Universitat Oberta de Catalunya

uoc.edu

Long non-coding RNAs annotation in *Strongylocentrotus purpuratus*. Studying the need of a reference genome.

Estudiante Sonia Doblado Martín
Máster universitario en Bioinformática y bioestadística UOC-UB
Área 1

Consultor/a Cinta Pegueroles Queralt
Profesor/a responsable de la asignatura: Marc Maceira Duch

Entrega: 01/2021

FICHA DEL TRABAJO FINAL

Título del trabajo:	Long non-coding RNAs annotation in *Strongylocentrotus purpuratus.* Studying the need of a reference genome.
Nombre del autor:	Sonia Doblado Martín
Nombre del consultor/a:	Cinta Pegueroles Queralt
Nombre del PRA:	Marc Maceira Duch
Fecha de entrega:	01/2021
Titulación:	Máster universitario en Bioinformática y bioestadística UOC-UB
Área del Trabajo Final:	Área 1
Idioma del trabajo:	Castellano
Palabras clave	lncRNA, transcriptoma, *de novo* assembly

Resumen del Trabajo (máximo 250 palabras):

El objetivo del trabajo es averiguar si es posible anotar lncRNAs de manera efectiva sin disponer del genoma publicado de la especie objetivo como referencia. Para ello, en este estudio se anotan los lncRNAs de una especie cuyo genoma sí está disponible, con lncRNAs ya anotados y de la cual existen datos de RNA-Seq, como es *Strongylocentrotus purpuratus*. Estos datos de experimentos RNA-Seq previos se utilizaron para ensamblar un transcriptoma *de novo* mediante el software Trinity. A partir de dicho transcriptoma, se anotaron lncRNAs de dos modos distintos vía FEELnc: i) se obtuvo una lista de posibles lncRNAs filtrando los transcritos *de novo* y ii) usamos un algoritmo de *machine learning* entrenado con lncRNAs ya anotados para la especie. Por último, comparamos estos resultados entre ellos y contra el set de lncRNAs obtenido usando el genoma de referencia, siendo éste el protocolo estándar.

Según nuestros resultados, sería posible anotar los lncRNAs sin genoma de referencia. Hay una diferencia en la cantidad de lncRNAs anotados entre nuestros dos sets de sólo un 1,03%, y hemos anotado un 41,6% de los lncRNAs ya anotados con genoma de referencia para *S. purpuratus* en el estudio más reciente disponible.

Abstract (in English, 250 words or less):

The aim of our project is to investigate the feasibility of annotating lncRNAs without a reference genome. In order to estimate the accuracy of the annotations, we selected a species (*Strongylocentrotus purpuratus)* having both a reference genome and RNA-Seq data publicly available. First, we obtained a transcriptome *de novo* assembly, using the software Trinity. We then used two approaches to annotate lncRNAs via FEELnc: i) we obtained a list of putative lncRNAs by filtering the *de novo* transcripts using several criteria and ii) we used a machine learning method fed with already annotated lncRNAs. Finally, we compared the results from both methods between them and also to the lncRNAs set obtained mapping directly to a reference genome, which is the standard protocol.

According to our results, it would be possible to annotate the lncRNAs without a reference genome. There's a difference in the amount of annotated lncRNAs between our two sets of only 1,03%, and we have found 41,6% of the lncRNAs annotated using a reference genome in the most recent paper available for *S. purpuratus*.

Índice

Lista de figuras

Lista de tablas

1. Introducción

1.1 Contexto y justificación del Trabajo

1.1.1. Descripción general

El presente Trabajo Final de Máster (TFM) lleva por título “Long non-coding RNAs annotation in *Strongylocentrotus purpuratus*. Studying the need of a reference genome”. La pregunta que se intentará responder es si es posible anotar lncRNAs sin un genoma de referencia. Los lncRNAs (*long non-coding RNAs*) son RNAs celulares endógenos que se asume que no codifican para proteínas, y con una longitud mayor de 200pb. Los lncRNA presentan una menor expresión que los genes codificantes y están mucho menos conservados[(1)] [(2)]. El protocolo estándar para anotar lncRNAs consiste en mapear los transcritos candidatos a ser lncRNAs a un genoma de referencia. Aunque la cantidad de especies con lncRNAs anotados va en aumento, dicha cantidad está limitada, ya que muchas, sobre todo las no-modelo, no disponen aún de un genoma publicado que pueda emplearse como referencia.

Para la anotación de lncRNAs en la que no se dispone de un genoma de referencia, se parte de la realización de un transcriptoma *de novo* mediante ensambladores como Trinity. Como los lncRNAs presentan una expresión más baja comparados con los genes codificantes de proteínas, las alineaciones *de novo* pueden eliminar los transcritos con baja expresión de los conjuntos de datos[(3)] [(4)], por lo que puede ser que no se anoten correctamente. Ésto es lo que vamos a evaluar y este factor es la razón por la que la anotación con un genoma de referencia debe ser la opción de preferencia.

El tema del trabajo es comparar anotaciones de lncRNAs en el erizo morado *Strongylocentrotus purpuratus* (Stimpson, 1857), siguiendo dos metodologías diferentes. El primer paso fue la obtención de un transcriptoma *de novo* a razón de experimentos previos de RNA-Seq[(5)], a partir del cual se anotarán los lncRNAs. En la primera metodología, se identificaron los lncRNAs sin usar ningún tipo de referencia. En la segunda metodología, se obtuvieron los lncRNAs ya anotados[(6)] en experimentos previos para los cuales se utilizó el genoma publicado de *S. purpuratus*; a continuación se empleó este set de

lncRNAs para entrenar un algoritmo y anotar los lncRNAs presentes en nuestro transcriptoma *de novo*. Por último, se compararon nuestros resultados con los obtenidos en estudios previos en los que se usó el genoma como referencia[6].

La importancia del trabajo radica en la comparación de los resultados de ambas metodologías con los resultados obtenidos por otros investigadores empleando un genoma de referencia, como herramienta para dilucidar si es en realidad necesario disponer de dicho genoma o no a la hora de anotar lncRNAs. La comparación entre nuestras metodologías nos permitirá también saber si el hecho de tener lncRNAs anotados para la especie objetivo u otra especie cercana mejora nuestras anotaciones.

1.1.2.- Justificación del trabajo

El estudio de los lncRNAs es un campo relativamente reciente. Incluso la descripción de estos es tan poco detallada que, aunque con los conocimientos actuales nos es suficiente para poder realizar una anotación, dicha identificación como lncRNA puede estar englobando una gran variedad de tipos de secuencias. La literatura científica no es por tanto demasiado abundante, y lo encontrado en la mayoría de los estudios son anotaciones de lncRNAs basadas en un genoma de referencia[7][8]. ¿Existen estudios en los que no se utilice un genoma de referencia?[9][10]. Sí, pero en cantidad reducida, por lo que este trabajo estaría realizando una aportación importante a este campo de estudio. Pero lo que no se ha encontrado en la literatura es un estudio que realice una comparación entre lo obtenido por un mismo equipo de investigadores, con genoma y sin genoma de referencia, sobre una misma especie. Es aquí donde el presente trabajo realizará la mayor aportación al campo del estudio de los lncRNAs, ya que esta comparación nos servirá para determinar si es necesario o no el genoma de referencia. Eliminar la concepción de que es necesario tener un genoma de referencia podría las puertas a una gran cantidad de estudios con especies no modelo, o especies cuyo genoma no ha sido publicado ni lo será en los próximos años.

Para añadirle valor, se ha elegido como objetivo *Strongylocentrotus purpuratus* por ser una especie de un grupo con poca representación en los estudios de lncRNAs como son los equinodermos, pero que reúne las condiciones mencionadas de tener el genoma publicado y haber sido protagonista de estudios RNA-Seq.

1.2 Objetivos del Trabajo

En este apartado se enumeran los objetivos generales y específicos del TFM. Se estimaron tres objetivos generales, que cubren la mayor parte del trabajo. Éstos se consideraron esenciales para el correcto desarrollo del trabajo, ya que si no se hubiese conseguido alguno de los tres el TFM no estaría completo. Los tres objetivos generales engloban a su vez cuatro objetivos específicos, más concretos; los tres primeros son igualmente necesarios para poder decir que este trabajo se ha llevado a cabo con éxito. El objetivo específico número 4 no es esencial para el desarrollo del trabajo, pero se llevó a cabo una vez se hubieron conseguido el resto de objetivos, al disponer del tiempo suficiente para al menos realizar una caracterización a grandes rasgos de los lncRNAs anotados.

1.2.1.- Objetivos generales

1. Obtención de un transcriptoma *de novo* utilizando los experimentos RNA-Seq sobre la especie disponibles en bases de datos públicas.
2. Anotación lncRNAs a partir del *de novo assembly* usando dos metodologias: predicción de *coding potential* y predicción de lncRNAs vía *machine learning*.
3. Comparación de resultados de la anotación de lncRNAs obtenidos en este trabajo con los obtenidos mediante *pipelines* utilizadas en experimentos previos.

1.2.2. Objetivos específicos.

1) Predecir el *coding potential* de las regiones anotadas *de novo*.
2) Utilizar los lncRNAs ya anotados en *S. purpuratus* como *training set*, estudiar la posible predicción de secuencias que constituyan lncRNAs en el genoma completo de *S. purpuratus*.
3) Comparativa de los lncRNAs anotados con las dos metodologías y con los lncRNAs en la literatura mapeando contra un genoma de referencia.
4) Caracterizar y validar la funcionalidad de los lncRNAs.

1.3 Enfoque y método seguido

Para la realización de este trabajo se han empleado, como ya se ha mencionado, dos metodologías diferentes: por un lado, se obtuvo un transcriptoma *de novo* y se anotaron los lncRNAs, mientras que por otro lado se utilizaron lncRNAs de experimentos previos para entrenar un algoritmo de predicción.

1.3.1. Obtención de transcriptoma *de novo*

Se comenzó el proyecto con la obtención del transcriptoma *de novo* siguiendo el flujo de trabajo o *pipeline* publicado en el trabajo de Ceschin *et al.* (2020)[(11)]. Esta *pipeline* incluye desde el control de calidad de los datos hasta la caracterización de los lncRNAs. Los pasos descritos en el artículo se realizaron sobre datos de RNA-Seq de la especie localizados en la literatura científica disponible. Algunos pasos difieren debido a ajustes en la planificación. Por ejemplo, no se ha empleado el transcriptoma clusterizado por CD-HIT, ya que el proceso ha sido largo y ha dado diversos errores, por lo que no ha dado tiempo a terminarlo. De haber esperado a completarlo con éxito, no habríamos dispuesto de un transcriptoma *de novo* para realizar los siguientes pasos de la *pipeline*. Se intentó sustituir Annocript por Trinotate, pero también hubo problemas con el tiempo de procesado de los datos. Al final, la anotación se realizó mediante BLAST. De todos modos, se siguió el proceso de Trinotate,

por lo que se dispone de una base de datos SQLite con toda la información obtenida sobre el transcriptoma obtenido vía Trinity. Esto se explica con más detalle en el Capítulo 2.

1.3.2. Búsqueda de lncRNAs ya anotados con genoma de referencia

Mientras se completaba el ensamblaje del transcriptoma *de novo*, se obtuvieron los lncRNAs de *S. purpuratus* ya anotados. Estos lncRNAs se utilizaron como *training set* con el objetivo de predecir los lncRNAs en el transcriptoma *de novo* de *S. purpuratus*.

1.3.3. Comparación de los resultados obtenidos en ambas metodologías

Por último, se compararon los lncRNAs obtenidos según las diferentes metodologías. El método a seguir será la búsqueda intensiva de bibliografía para conocer las últimas actualizaciones en el campo de los lncRNAs, y se discutirá sobre si es necesario o no disponer de un genoma de referencia para la anotación de los mismos. Se intentará dar respuesta a la pregunta que ha originado este TFM.

1.4 Planificación del Trabajo

En este apartado se incluye la última versión de la planificación del trabajo. En total, ha sufrido dos modificaciones, siendo la aquí presentada la tercera versión. Las diferentes modificaciones han sido debidas a fallos humanos, pero sobre todo a fallos técnicos.

El retraso en el día de entrega de la PEC2 era de alrededor de una semana en relación a lo establecido en el cronograma original. Esto fue debido en su mayor parte a fallos informáticos, y a una subestimación del tiempo de procesado de los datos. En cuanto a los fallos informáticos, fue necesario instalar en mi ordenador personal el Sistema Operativo (SO) Ubuntu, ya que

previamente había estado empleando una máquina virtual con ese SO en Windows, pero daba demasiados errores. En relación al procesado de los datos, la descarga de los archivos SRA de experimentos RNA-Seq de *S. purpuratus* y su posterior control de calidad mediante FastQC tardó varios días más de lo esperado, ya que me fue imposible encontrar la forma de realizarlo en el clúster, con la diferencia de capacidad de procesado entre mi máquina y el clúster que ello implica. Este error me hizo valorar la importancia de saber lanzar un *job* correctamente.

En la fecha de entrega de la PEC3, se estimó que el retraso en relación al plan original era de unos 10-12 días en relación a la planificación original. No pudo calcularse de forma exacta debido a la cantidad de actividades no programadas que se llevaron a cabo, lo cual adelantó trabajo para las tareas que aún quedaban por completar. Como ocurrió en la PEC anterior, el retraso fue debido en su mayor parte a fallos informáticos, y a una subestimación del tiempo de procesado de los datos. En esta ocasión, el error fue confiar en exceso en que el procesado por parte de Trinity iba a terminar a tiempo. Este programa fue el empleado desde un inicio para realizar el ensamblaje *de novo* del transcriptoma, lo cual supone la base del TFM. Al entrar en el mes de diciembre sin haber terminado ese ensamblaje, se optó por correr de forma paralela otro programa más rápido pero con el mismo fin, para aumentar las posibilidades de obtención del transcriptoma a tiempo. El programa elegido, fue SOAPdenovo-Trans. El día 6 de diciembre terminó con éxito el proceso de Trinity de forma inesperada, por lo que esta acción de mitigación no habría sido necesaria, pero da más opciones a la hora de analizar datos.

Los cambios más relevantes de planificación en relación al cronograma original son una mayor duración de las tareas de anotación (tareas 5, 8 y 9) y un retraso en la fecha de consecución del hito de finalización de ensamblaje *de novo*.

1.4.1. Tareas:

Las tareas se pueden dividir en 3 grupos fundamentales, según las diferentes metodologías a comparar descritas en apartados anteriores. Esto es, el primer grupo reúne las tareas a realizar en la anotación *de novo* (tareas de 1 a 6), el segundo grupo de tareas engloba las predicciones mediante los lncRNAS ya anotados en estudios previos, (tareas de 7 a 10) y el último grupo de tareas se reserva para la comparación de resultados y la elaboración de la memoria (tareas 11 y 12):

1. Búsqueda de todos los datos disponibles de RNA-Seq para *S. purpuratus*
2. Control de calidad y *trimming* (FastQC + Trimmomatic)
3. Ensamblaje *de novo* (Trinity)
4. Clustering jerárquico (CD-HIT)
5. Anotación de transcritos (Annocript + FEELnc)
6. Caracterización: niveles de expresión, análisis comparativo con especie cercana.
7. Obtención de los lncRNAs ya anotados en estudios anteriores.
8. Predicción mediante algoritmo entrenado con dichos lncRNAs ya anotados.
9. Anotación del segundo set de transcritos (Annocript +FEELnc)
10. Caracterización del segundo set de transcritos: niveles de expresión, análisis comparativo con especie cercana.
11. Comparación de los resultados obtenidos en los dos bloques de tareas.
12. Elaboración de la memoria del TFM, y preparación de la defensa.

1.4.2. Calendario

En la Figura 1 puede observarse el cronograma con el periodo de realización para cada una de las tareas. Como ya se ha comentado, han sido necesarias dos modificaciones de la planificación original, una en cada una de las dos últimas PECs; sólo se incluye el cronograma para la última versión de la planificación. En tonos rojizos aparecen las tareas relacionadas con el primer bloque (*de novo assembly* y anotación sin referencias), mientras que en tonos verdosos aparecen las tareas relacionadas con el segundo bloque (uso de algoritmo entrenado). En cuanto al tercer bloque, la tarea relativa a la comparación de metodologías aparece en azul, mientras que la redacción de la memoria aparece en amarillo. Esta última tarea se realizó de manera simultánea a las otras, para ganar tiempo. Aun así, se reservaron unos días al final del proyecto exclusivamente para la elaboración de la memoria, que en caso de que surja algún problema durante el desarrollo del trabajo servirán de comodín para las tareas esenciales a completar.

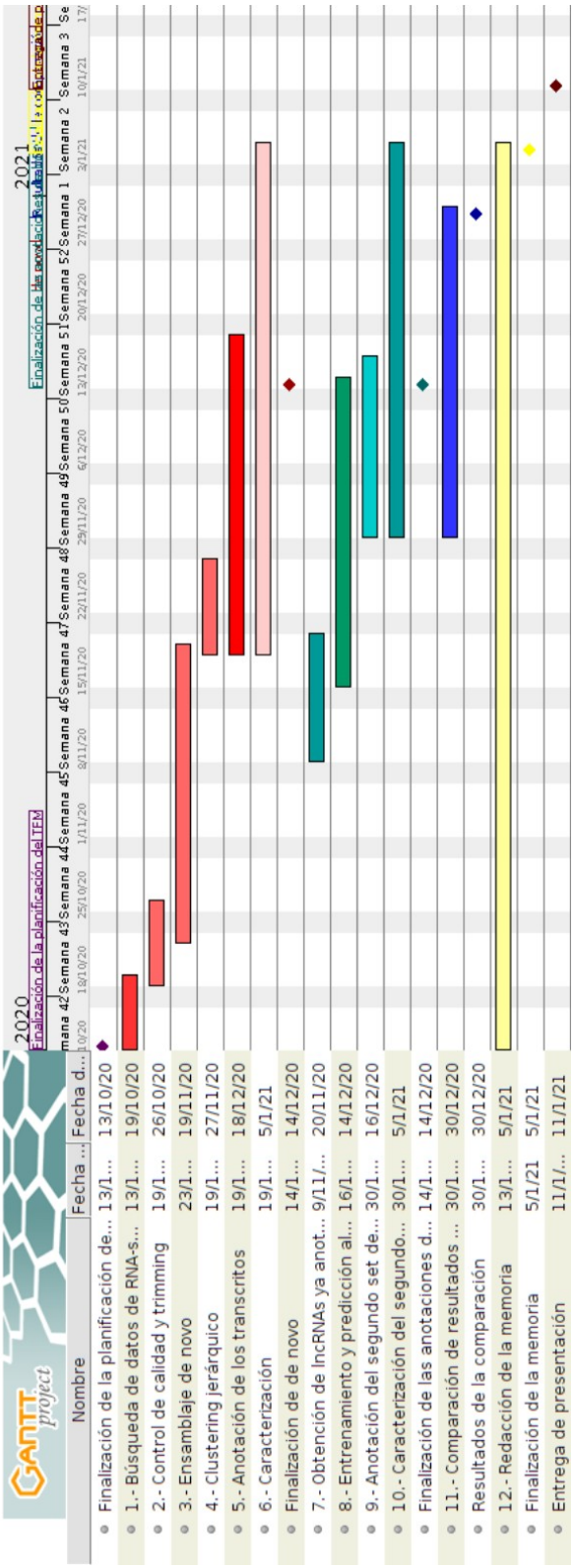

Figura 1: Cronograma representando la última versión de la planificación del trabajo. En tonos rojizos aparecen las tareas relacionadas con el primer bloque (*de novo* y anotación sin referencias), mientras que en tonos verdosos aparecen las tareas relacionadas con el segundo bloque (uso de algoritmo entrenado). En cuanto al tercer bloque, la tarea relativa a la comparación de metodologías aparece en azul, mientras que la redacción de la memoria aparece en amarillo. Los hitos aparecen señalados con un rombo.

Los hitos descritos en el siguiente apartado aparecen señalados con un rombo.

Se decidió que las tareas de caracterización sólo se llevarían a cabo en caso de que se disponga de tiempo una vez se han obtenido los dos sets de transcritos, ya que no son esenciales para el proyecto, aunque sí que añadirían información de valor. Es por ello que aparecen en el calendario "desconectadas" del resto de tareas de su bloque.

1.4.3. Hitos

Para mantener la coherencia, se intentó que los hitos identificados a continuación se correspondiesen dentro de lo posible con las fechas de entrega de las diferentes PECs. Esto se pensó así para tener nueva información que presentar entre las PEC 2 y 3, y forzar la redacción de los resultados que se fuesen obteniendo en las diferentes metodologías. A continuación aparecen los hitos con las fechas en que se planeaba obtenerlos. Entre paréntesis se indica la fecha en que el tercer hito se consideró alcanzado, único hito que se vio afectado por los retrasos del proyecto:

- **13/10/2020 –** Finalización de la planificación del trabajo
- **27/11/2020 – (14/12/2020)** Finalización de las anotaciones de lncRNAs *de novo*
- **14/12/2020 –** Finalización de las anotaciones de lnRNAs con genoma de referencia
- **30/12/2020 –** Resultados de la comparación de metodologías de anotación
- **5/01/2020 –** Finalización de la memoria del TFM

1.5 Breve sumario de productos obtenidos

- Archivos SRA obtenidos (BioProject PRJNA81157) en formato .fastq
- Obtención de un informe sobre la calidad de cada uno de esos archivos .fastq después de eliminar los adaptadores mediante *FastQC y MultiQC.*
- Resultado de ensamblaje mediante SOAPdenovo-Trans en contigs y *scaffolds*, en formato .fa. Cálculo del tamaño medio de los *inserts*.
- Resultado de ensamblaje mediante Trinity, en formato .fasta.
- Resultados de % de *alignments* vía Bowtie para los transcriptomas de Trinity y SOAPdenovo-Trans.
- Resultados de Transrate para el transcriptoma *de novo* ensamblado con SOAPdenovo-Trans.
- Base de datos SQLite con resultado de anotación vía Trinotate del transcriptoma de Trinity sin clusterizar. Incluye, entre otros: resultados de regiones homólogas obtenidos a partir de BLAST (blastx para transcriptomas completos y blastp para archivos .pep resultados de Transdecoder), tanto para transcriptomas *de novo* como para transcriptoma de Spur_5.0.
- Resultados de FEELnc para el transcriptoma *de novo* sin genoma de referencia y mediante algoritmo entrenado con lncRNAs conocidos.

1.6 Breve descripción de los otros capítulos de la memoria

- 2. Materiales y métodos: se describen brevemente los programas utilizados, los recursos necesitados y las cuestiones por las que se han variado los programas contemplados en la planificación original. Se detalla el flujo de trabajo seguido para la obtención de los resultados en las dos metodologías empleadas.
- 3. Resultados: incluye gráficos y tablas que muestran los principales resultados del proyecto, así como información adicional que puede ser utilizada en otros estudios, obtenida en *pipelines* parciales realizadas durante el proyecto.

- 4. Discusión: en este capítulo se interpretan los resultados mostrados en el capítulo anterior, poniéndolos en contexto y razonando si coinciden o no con los esperado.
- 5. Conclusiones: principales ideas sacadas del experimento. Se mencionan posibles pasos en proyectos futuros, y lecciones aprendidas de los errores cometidos.
- 6. Glosario: contiene una breve descripción de los términos y acrónimos más empleados en este documento.
- 7. Bibliografía: lista numerada de las referencias empleadas.
- 8: Anexos. Incluye el Anexo I.

2. Materiales y métodos

Debido a los recursos informáticos que necesita el procesado de los archivos empleados en este proyecto, los análisis se han hecho desde el clúster del grupo de investigación al que pertenece la directora del TFM. Esto agiliza los tiempos de procesado pero también supone un inconveniente, que es el de la dependencia del equipo de soporte informático a la hora de instalar nuevo software. Por ello, a la hora de analizar los datos, se ha priorizado el uso de programas ya instalados en el clúster, siendo ésta la razón por la cual muchas de las aplicaciones empleadas no coinciden con la última versión disponible.

En la Figura 2 puede observarse el flujo de trabajo seguido en el proyecto, incluyendo las dos metodologías mencionadas en el apartado anterior.

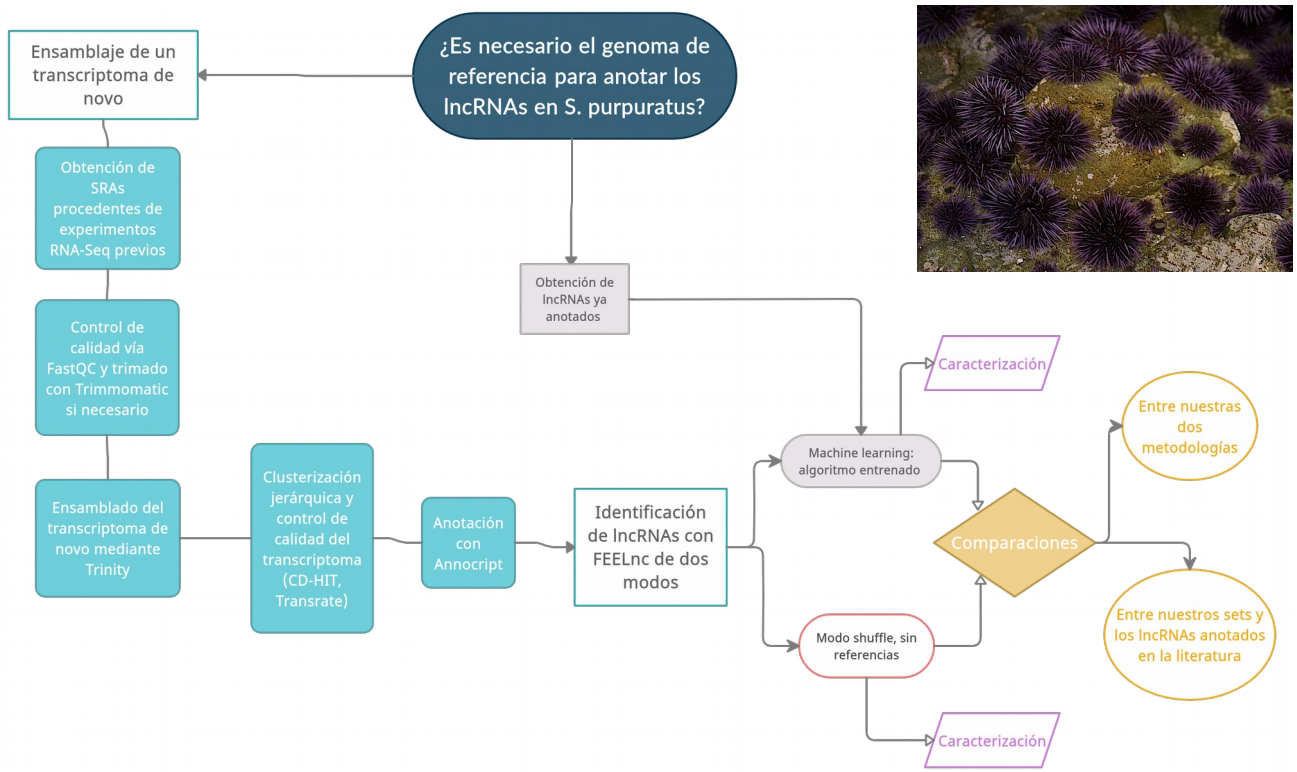

Figura 2: Diagrama de flujo de la *pipeline* empleada. Se parte de la pregunta de si es necesario un genoma para anotar los lncRNAs en *S. purpuratus*, Para contestar, se realiza el ensamblaje de un transcriptoma *de novo* (pasos en color verde agua, rectángulos redondeados). Una vez anotado, se identifican los lncRNAs con dos metodologías. En oval rojo con fondo blanco, aparece la metodología sin referencias, en la cual se identifican los lncRNAs mediante el modo shuffle del programa FEELnc. En gris aparece la metodología en la cual se emplean lncRNAs ya anotados para la especie para entrenar un algoritmo que identifique lncRNAs en nuestro transcriptoma. Cada una de esas metodologías será después comparada entre ellas y con el set de lncRNAS disponible en la literatura (rombo y círculos dorados). Las tareas de caracterización de cada uno de nuestros sets, marcadas con paralelogramo rosa. En la esquina superior derecha se incluye una imagen de varios individuos de la especie *S. purpuratus.*

(Fuente de la imagen :commons.wikimedia.org//wiki/File:Purple_Sea_Urchin_-_Strongylocentrotus_purpuratus_(16455860102).jpg).

2.1.- Obtención de datos RNASeq y control de calidad

El primer paso del proyecto fue la descarga de los datos correspondientes al BioProject PRJNA81157 (SRA:SRP014690)[12] . Se eligió este BioProject mediante investigación de la literatura científica preexistente[6] [13][14]. En todos los artículos científicos consultados sobre transcriptoma de *S. purpuratus* se menciona este BioProject como referencia. Elegir las mismas muestras permitirá la comparación de resultados.

A continuación, se realizó el control de calidad para cada uno de los archivos mediante FastQC v.011.9[15]. Al observarse aparentes problemas de calidad, se realizó un *trimming* con Trimmomatic v0.39[16] sólo en un par de archivos de los de peor calidad, a modo de comprobación. Después se realizó un nuevo control de calidad a estos archivos "*post-trimming*", y se comparó con el resultado obtenido antes del *trimming*. Como el resultado fue que aproximadamente un ~5% de los *reads* contenían adaptadores, se estimó que en este caso el *trimming* era necesario. Por ello, se utilizó Trimmomatic en todos los archivos, y se volvió a realizar un control de calidad mediante FastQC. Para una mejor visualización de los resultados de este control de calidad, se empleó MultiQC[17]. MultiQC *t*ambién permite visualizar las variables calculadas por FastQC, como puede ser el contenido en adaptadores de las muestras, el número de *reads* de cada secuencia o la puntuación de calidad de las secuencias. En el Anexo I puede verse una tabla resumen para todos los archivos empleados elaborada por MultiQC.

2.2.- Ensamblaje de transcriptoma *de novo*

Una vez se comprobó que los datos eran de la calidad adecuada, se procedió al siguiente paso: se emplearon los archivos resultantes del *trimming* para el ensamblaje de un transcriptoma *de novo* mediante Trinity v2.11.0[18]. Trinity es un programa que ensambla un transcriptoma *de novo* a partir de

datos de experimentos RNA-Seq a lo largo de tres fases: Inchworm, Chrysalis y Butterfly. En la Figura 3 se ilustra el funcionamiento del programa.

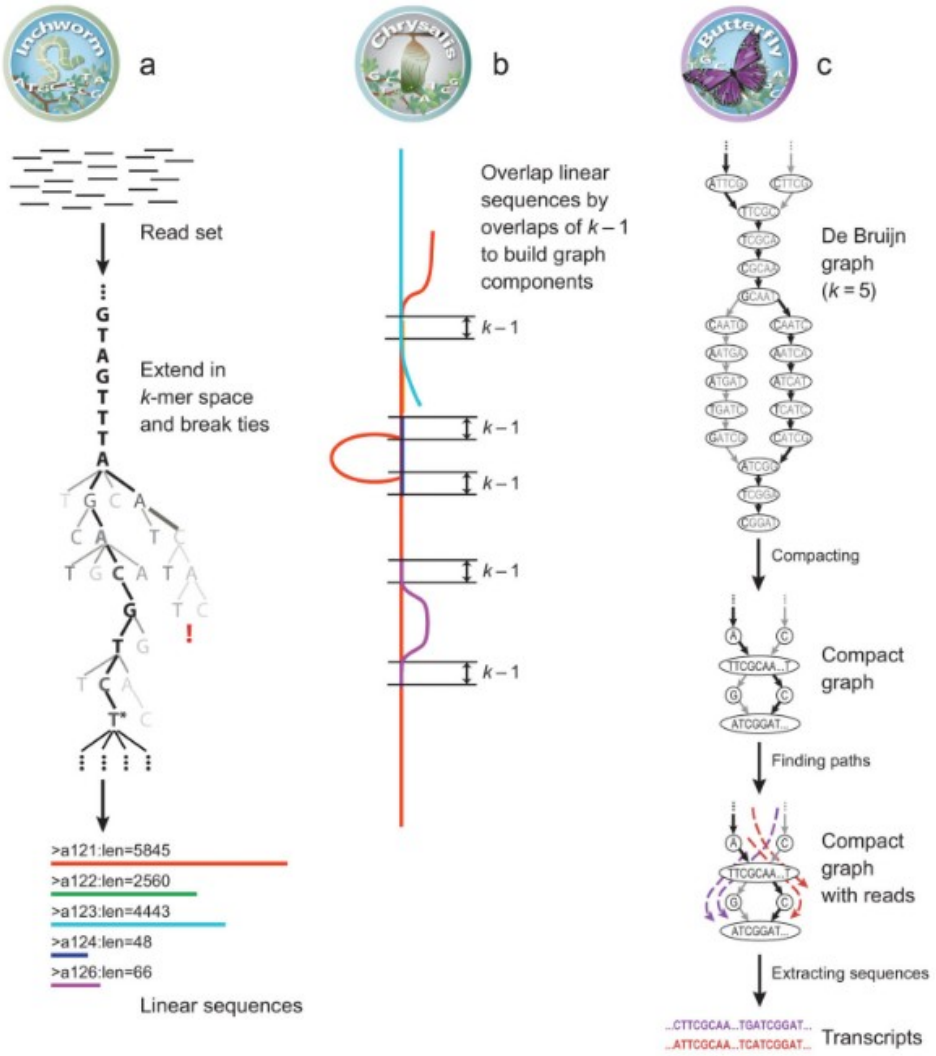

Figura 3: Resumen del funcionamiento de Trinity. a) Inchworm ensambla los *reads* del conjunto de datos (línea negra, arriba) buscando rutas en un gráfico de k-mers (medio), lo que resulta en una colección de contigs lineares (líneas de colores, abajo). b) Chrysalis une los contigs en pools si comparten por lo menos un k-1-mer y los reads abarcan toda la unión; después construye gráficos de Bruijn individuales para cada uno de esos *pools* (líneas de colores). (c) Butterfly coge cada uno de esos gráficos de Bruijn de Chrysalis (arriba), y corta los bordes "falsos" y compacta rutas lineales (medio). Después, concilia el gráfico con los reads (flechas de colores, abajo) y pares (no se muestra), y muestra como *output* una secuencia lineal por cada forma de *splice* o transcrito parálogo reflejado en el gráfico (abajo, secuencias de colores)[(18)].

El procesado de Trinity comenzó a finales de octubre, pero el procesado de los datos fue mucho más complejo y largo de lo esperado. Hubo que comenzarlo de nuevo varias veces, tanto por fallos del programa como por errores propios, como ocurrió con la normalización de los datos. Se asumió que el programa lo haría de forma automática, pero a la versión instalada en el clúster había que indicarle expresamente que debía realizar la normalización. Este error provocó un nuevo inicio del proceso de ensamblaje en noviembre. En la fecha de entrega de la PEC2, el proyecto se encontraba estancado en esta tarea de ensamblaje *de novo*. Desde entonces, se tomaron medidas de mitigación como el cambio de *software* para el ensamblaje *de novo*, ya que Trinity no mostraba signos de avanzar lo suficientemente rápido como para tener resultados antes de la fecha de entrega de la memoria del TFM. Por este riesgo de no tener datos a presentar, se valoraron otros programas que cuentan con la misma validación en la literatura científica existente pero se consideran mucho más rápidos[19][20][21]. Con el objetivo de ilustrar acción de mitigación y las diferencias entre los ensambladores disponibles, en la Figura 4 se muestran los resultados de una comparación de un estudio previo del tiempo de procesado y la RAM necesaria para los ensambladores más utilizados.

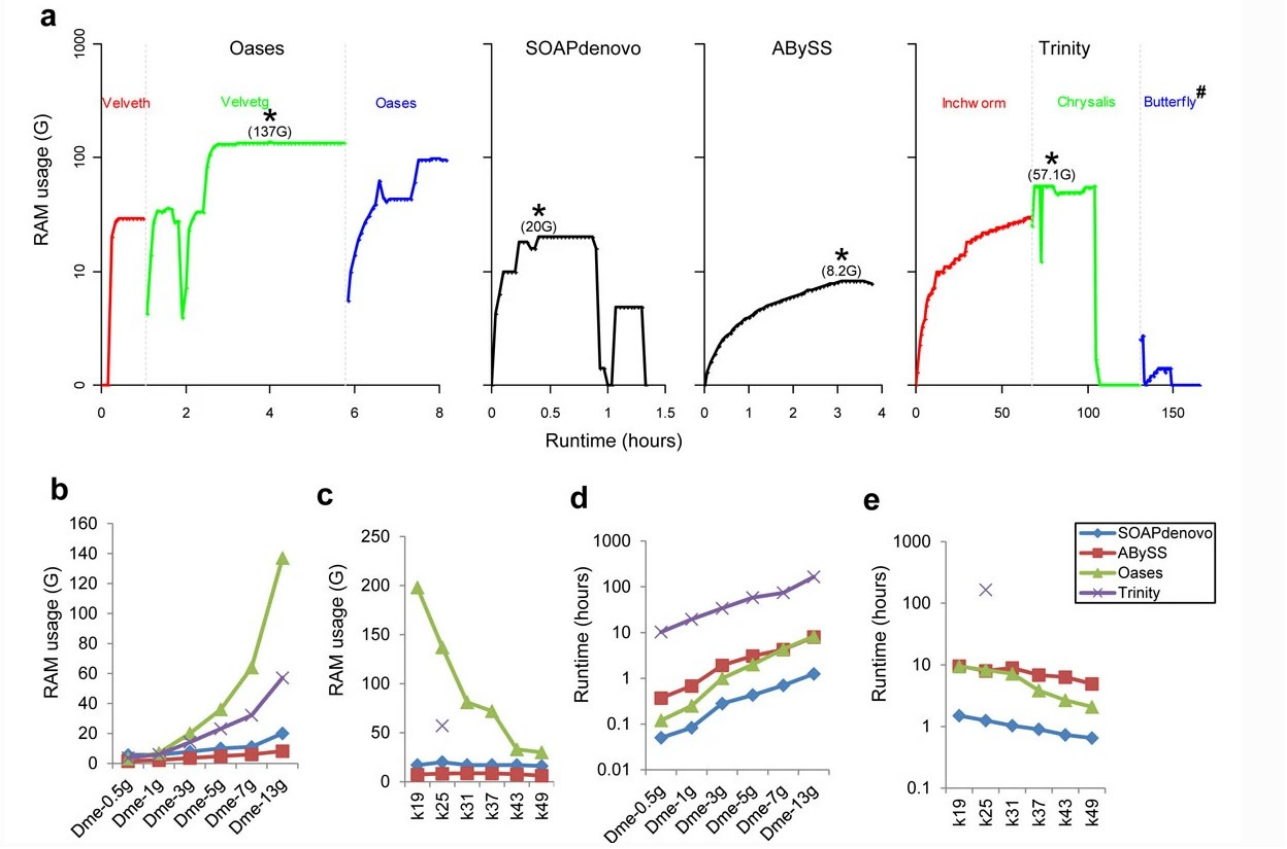

Figura 4: Tiempo de procesado y uso de RAM para los ensambladores Trinity, Oases, AbySS y SOAPdenovo. a) Tiempo real monitorizado y uso de RAM para cada método usando el conjunto de datos de *Drosophila melanogaster* Dme-13g. El uso máximo de RAM para cada ensamblador está marcado con un asterisco, y las tres etapas de Oases y Trinity están marcadas con colores diferentes (rojo: Velvet e Inchworm; verde: Velvetg y Chrysalis; azul: Oases y Butterfly). Se muestra uso de RAM (b) y el tiempo de procesado (d) de cada método usando diferentes cantidades de *inputs* con un valor de k-mer de 25. Se muestra el uso de RAM (c) y el tiempo de procesado (e) de cada método usando el dataset de Dme-13g con diferentes valores de k-mer[(21)].

Utilizando los *reads* ya normalizados mediante Trinity, se seleccionó SOAPdenovo-Trans v 1.0.4[(22)] como nuevo *software* de ensamblaje. Para este programa, hacen falta algunos pasos previos que no fueron necesarios con Trinity. SOAPdenovo-Trans necesita que el usuario personalice un archivo de configuración en el que figuran variables como el tamaño medio de los *inserts*, o la longitud máxima de los *reads*. Los parámetros de los *reads* fueron fáciles de calcular, pero no ocurrió lo mismo con los tamaños medios de los *inserts*. No se encontró información en los artículos en los que se mencionaban los datos del BioProject empleado, y tampoco se encontraron los parámetros exactos empleados en los *arrays* de Illumina IIx en los experimentos RNA-Seq de los que se obtuvieron esos *reads*, aunque deberían ser incluidos en los metadatos de los archivos SRA producto de dichos experimentos. En otros

experimentos con *S. purpuratus*, se menciona que el tamaño medio de los *inserts* fue de 300bp[(23)] , por lo que se realizó un transcriptoma con este parámetro. Por otro lado, se calculó el tamaño de los *inserts* mediante Bowtie2 v2.1.0[(24)] , y se realizó la media con R v3.4.4[(25)] . Este análisis arrojó un resultado de 139bp, por lo que se volvió a repetir SOAPdenovo-Trans pero con este parámetro. Al analizar el porcentaje de alineamiento mediante Bowtie2 (descrito más adelante) y comparar ambas opciones, los resultados fueron idénticos, con un 60.04%. Se siguió la *pipeline* con los resultados obtenidos teniendo en cuenta la opción del *insert size* de 139bp.

En cuanto a los recursos empleados se lanzó el primer *job* con SOAPdenovo-Trans el día 2 de diciembre. Se utilizaron 10GB de memoria virtual para ambos programas, con 18 *cores* en el caso de Trinity y 10 en el caso de SOAPdenovo-Trans. Este último programa tardó en realizar el ensamblaje entre 70 y 90 minutos, frente a las semanas que ha tardado Trinity. El día 6 de diciembre, de forma inesperada, Trinity acabó su proceso. Por ello, y para minimizar los riesgos de no disponer de datos a analizar, se siguió la *pipeline* original con los *contigs* resultado de SOAPdenovo-Trans (es decir, se sometió al paso de clusterización jerárquica), pero se realizó una *pipeline* "alternativa" con el archivo .fasta resultante del procesado de Trinity sin clusterizar. Esta *pipeline* alternativa se inició para aprovechar los tiempos muertos típicos entre procesos bioinformáticos y los programas que ya están instalados en el clúster, con la idea no sólo de tener más información, sino de ir probando las herramientas para tenerlas listas a la hora de realizar la *pipeline* principal. Este hecho ha resultado crucial, ya que como se explica en los siguientes párrafos, la calidad del transcriptoma obtenido via SOAPdenovo-Trans era de una calidad bastante menor que la del obtenido vía Trinity, por lo que finalmente se empleó éste último transcriptoma para la anotación de los lncRNAs. El haber comprobado ya que los programas de los siguientes pasos funcionaban correctamente, y haber solucionado los errores en caso de haberlos, permitió la finalización del proyecto, aún habiendo obtenido el transcriptoma apenas un mes antes de la fecha de entrega de esta memoria.

Para evaluar la calidad de los transcriptomas, se consideró que la mejor opción era el programa Transrate v1.0.1[(26)]. Tras no poder utilizarlo en el

clúster, se instaló en local, pero el ordenador carecía de la potencia suficiente como para correrlo sin que el SO lo finalizase ("*killed*"). Se buscaron alternativas para poder estudiar la calidad del transcriptoma *de novo* y se optó por Bowtie2. Bowtie2 permite calcular el porcentaje de *alignments* del transcriptoma y los *reads* empleados. Para ello, hay que realizar primero un *index* con el transcriptoma *de novo*, que después podrá emplearse para compararlo con los reads empleados para realizar el mismo, en este caso, los archivos normalizados por Trinity. Los resultados obtenidos son de un 60,04% de alineamientos para SOAPdenovo-Trans, y de un 83,66% para Trinity. La calidad de Trinity entraba dentro de lo que por consenso se tiene por buena calidad (70-90%)[(27)].

Aunque con el empleo de Bowtie2 ya podría considerarse suficiente como para deducir la calidad del transcriptoma, se siguieron estudiando otras alternativas. Otra forma de las formas de evaluar la calidad del transcriptoma fue mediante BUSCO v2/3[(28)], que evalúa la presencia de BUSCOs, que es un set de *core genes* muy conservados y que por lo tanto deberían estar correctamente ensamblados. Para ello se empleó el servidor en línea gvolante[(29)]. No se desistió en el intento de conseguir resultados vía Transrate, y finalmente hubo éxito. Por lo tanto, se obtuvieron datos sobre la calidad del transcriptoma con tres programas diferentes: Bowtie2, BUSCO y Transrate.

Tanto con los datos de SOAPdenovo-Trans como con los de Trinity se inició el proceso de *clustering* mediante CD-HIT v4.8.1[(30)]. Éste es un programa utilizado para clusterización y comparación de secuencias de nucleótidos y proteínas. Lo que se persigue con este programa es el de identificar quimeras, transcritos redundantes y posible fragmentación en el ensamblaje, con el fin de mejorar la calidad del transcriptoma *de novo*. En particular, se ha empleado CD-HIT-EST, que agrupa en clusters proteínas (DNAs) similares que cumplen con unos umbrales determinados fijados por el usuario (en este caso, valores por defecto)[(31)]. El *job* para cada uno de los transcriptomas fue lanzado el día 6 de diciembre, con 10GB de memoria virtual en 10 *cores*, acabando por primera vez para Trinity el día 17 de diciembre. Los días de diferencia con el fin de la tarea anterior fueron empleados en intentar analizar la calidad del transcriptoma *de novo* con Transrate, y en la búsqueda de alternativas al no poder realizarlo con éste. No se ha conseguido finalizar el

proceso con éxito para ninguno de los transcriptomas, dando error al guardar el archivo de *output*.

Por falta de tiempo, y debido al alto porcentaje de alineamiento arrojado por Bowtie2, se ha seguido el trabajo con los datos de Trinity sin clusterizar.

2.3.- Anotación

Annocript[32] dio diversos problemas de instalación, de modo que se optó por anotar el transcriptoma utilizando BLAST, con el objetivo de eliminar de nuestro transcriptoma todo aquello que pudiera ser una proteína. Para agilizar el proceso, se seleccionaron los mRNAs del transcriptoma de *S. purpuratus* ya publicado (anotación Spur_5.0)[33], se convirtieron en una base de datos de blast y se compararon con el transcriptoma obtenido vía Trinity. De este modo, se eliminó todo lo que puede codificar para proteína, por lo que el archivo de *input* para el siguiente programa, FEELnc, será de un tamaño mucho menor, con la idea de que el tiempo de procesado fuese mucho menor. Se empleó un valor de e de $10e^{-3}$, se solicitaron 10 *cores* y 15GB. Tras realizar esta operación, quedaron 14844 transcritos candidatos a lncRNAs en nuestro transcriptoma *de novo*.

De modo paralelo se empleó la *pipeline* de Trinotate, aprovechando que ya se encontraba instalado en el clúster y para hacer un mejor uso de los ya mencionados “tiempos muertos”. Trinotate es un programa de anotación de transcritos, particularmente para transcriptomas ensamblados *de novo*, que utiliza los resultados de otros programas ampliamente utilizados como BLAST[34] para búsqueda de homologías, Transdecoder, para predecir las regiones codificantes de los transcriptomas, o HMMER/PFAM para la identificacion de dominios de proteinas, además de hacer uso de otras bases de datos de anotaciones como eggNOG, GO y Kegg, entre otras funciones[48]. Todos esos datos se integran en una base de datos SQLite que permitirá anotar o buscar cualidades específicas del transcriptoma, según lo que busque el investigador. Este programa podría encargarse de la anotación de los transcriptomas como alternativa a Annocript. Para ello se utilizó el

transcriptoma de Trinity sin clusterizar. Esta *pipeline* incluyó la búsqueda de homólogos mediante BLAST para Trinotate: la base de datos de proteínas requerida por Trinotate es Uniprot. Esta base de datos se han tenido que descargar y formatear mediante el comando makeblastdb, ya que Uniprot no se encontraba en el clúster. Se ha empleado blastx para los transcriptomas completos y blastp para los archivos .pep resultantes de Transdecoder. También se ha realizado blastn en el transcriptoma *de novo* obtenido mediante Trinity, utilizando la base de datos de blast nt. No se esperó a terminar el *pipeline* de Trinotate para seguir con el proyecto por lo largo del mismo y la falta de tiempo, pero se consideró que podría añadir información de valor.

El archivo resultado de eliminar las regiones codificantes del transcriptoma de Trinity sin clusterizar vía BLAST, con esos 14844 transcritos, fue entonces utilizado con el programa FEELnc[35], empleado para identificar lncRNAs. De este programa, se empleó su módulo de cálculo de *coding potential*, $\text{FEELnc}_{\text{codpot}}$. Este módulo calcula un potencial de codificación para cada uno de los transcritos del *input* mediante un algoritmo *Random Forest* basado en el paquete randomForest[36] de R. El módulo se ha empleado en sus dos variantes:

-**modo shuffle**: en base a su *coding potential* sin tener genoma de referencia mediante el uso de Ushuffle[37]. Los mRNA conocidos necesarios para emplear el método de *shuffle* han sido obtenidos por dos métodos diferentes, para disponer de más opciones: mediante BLAST (descrito en el apartado siguiente) y a través del archivo existente de regiones codificantes (cds) para Spur_5.0 en NCBI[38] . Para la comparación de datos, se usaran los resultados obtenidos utilizando este segundo archivo, al ser el que menos problemas ha dado con Ushuffle.

-**algoritmo entrenado**: se obtuvieron los lncRNAs ya anotados en otros experimentos y se utilizaron como *training set* del algoritmo del programa.

Por último, se realizó un BLAST de los lncRNAs obtenidos en ambos métodos contra los lncRNAs obtenidos en el estudio más reciente disponible [6] ,

ya mencionado anteriormente, con un valor de e de $10e^{-3}$. El objetivo es comparar los lncRNAs obtenidos a raíz de un transcriptoma *de novo* (nuestros sets) contra los lncRNAs anotados usando un genoma de referencia (set del estudio previo).

2.4.- Caracterización

Siguiendo trabajos previos[39][40][41], se está estudiando la estructura secundaria de los sets de lncRNAs obtenidos por FEElnc mediante CROSSalign[42], que permite evaluar la conservación de la estructura secundaria de los RNA mediante el algoritmo CROSS (*Computational Recognition of Secondary Structure*) y el algoritmo DTW (*Dinamic Time Warping*). Se ha empleado la versión en línea del *software*, aplicando el modo *Standard DTW* tanto al set de lnRNAs obtenidos por el modo *shuffle* como a aquellos obtenidos utilizando el *training set*. En la fecha de entrega de la presente memoria, los datos siguen en proceso, por lo que no se han podido añadir los resultados a este documento. Pero se ha dejado que continúe porque estos resultados pueden ser de valor para proyectos futuros.

Se ha realizado también un estudio mediante catRAPID[43] de la interacción de los lncRNAs anotados en ambos métodos con las proteínas de *S. purpuratus*. Por tiempo, se han excluido las proteínas basadas en predicciones (categoría XP), utilizando únicamente aquellas categorizadas como mRNA (categoría NM). Idealmente, para la caracterización se emplearían ambas categorías de proteínas. Como resultado se obtiene una lista de las proteínas con las que interaccionan, la cualificación de la interacción (entre 0 y 3, siendo 3 el máximo), el *z-score*, el *ranking* y un enlace a su página de Uniprot. De esta lista, se han seleccionado las proteínas con un porcentaje de fortaleza de interacción (*Interaction Score*%) mayor del 10% y se ha elaborado una tabla con el código de identificación GO *idGO*, el tipo de proceso en que interviene, y una pequeña descripción de su función.

3. Resultados

Tras realizar el *trimming* a las secuencias para eliminar los adaptadores, se realizó un análisis de la calidad de las mismas mediante FastQC, y se empleó MultiQC para visualizar de una forma más sencilla los resultados. Como puede verse en la Figura 5, los *reads* que serán empleados en el ensamblaje *de novo* son considerados de buena calidad.

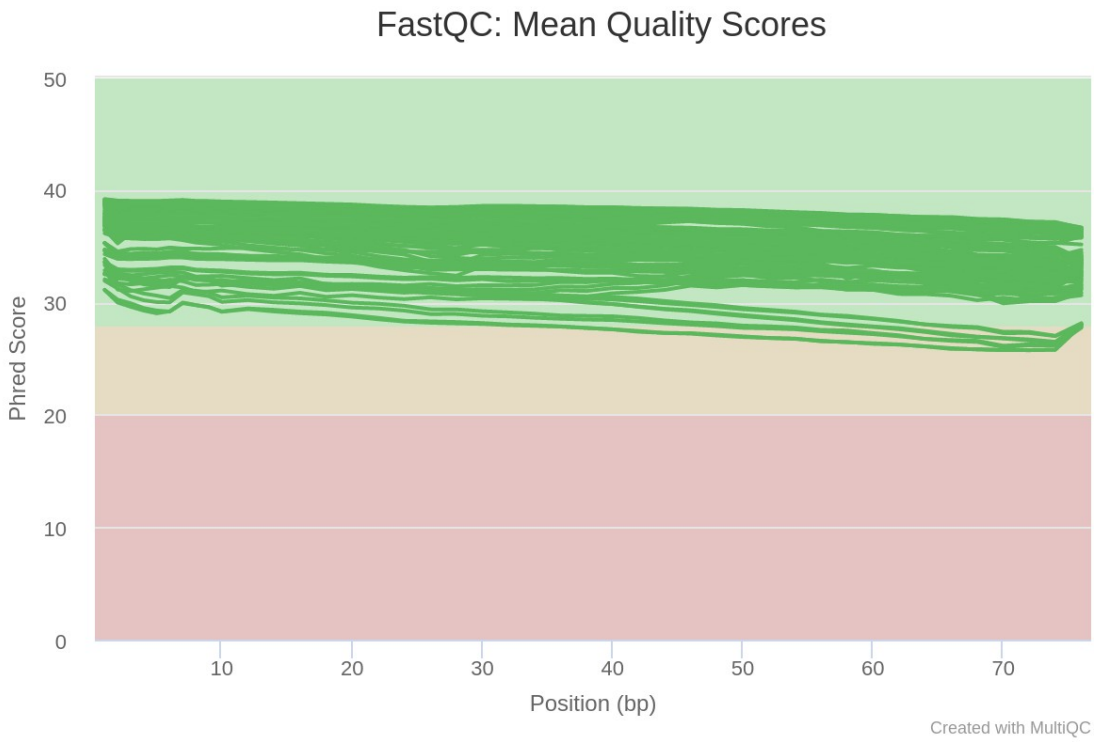

Figura 5: puntuaciones de calidad media por base y por secuencia, resultado de eliminar los adaptadores con Trimmomatic y evaluado por FastQC

En la Figura 6 se observa el porcentaje de citosina y guanina según la longitud de las secuencias empleadas. Se observa que para la mayoría de secuencias este porcentaje se encuentra entre el 30 y el 50%.

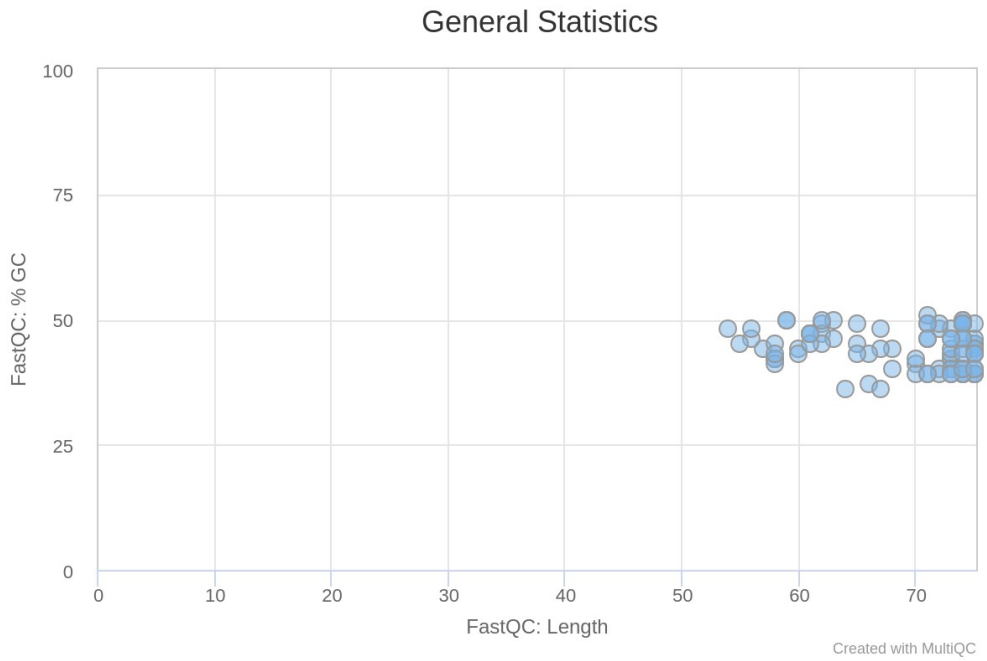

Figura 6: Gráfico del porcentaje de citosina y guanina (%GC) contra la longitud de los *reads* empleados en el proyecto después del proceso de *trimming*.

A continuación, en la Tabla 1 se muestran los diferentes resultados obtenidos al ensamblar el transcriptoma *de novo* mediante Trinity y mediante SOAPdenovo-Trans. Cabe destacar que lo único en lo que se asemejan los resultados de ambos programas es en el porcentaje de guanina y citosina (GC). SOAPdenovo-Trans ha identificado más del triple de contigs, y sin embargo Trinity considera que el número de contigs con un ORF es de casi el doble de aquellos considerados por SOAPdenovo-Trans. Por la longitud media puede deducirse que SOAPdenovo-Trans ha considerado contigs de un tamaño mucho menor que Trinity.

	Trinity	**SOAPdenovo-Trans**
Number of contigs	907236	2904191
Contigs with an ORF	74315	28928
Mean contigs length (bp)	503,61	47.93
N50	593	11267
GC content (%)	39,00 %	39,00 %

Tabla 1: Estadísticas básicas obtenidas tras la realización de los ensamblajes de transcriptomas *de novo* para los programas Trinity y SOAPdenovo-Trans.

La Tabla 2 muestra los porcentajes de alineamiento de los transcriptomas obtenidos *de novo* según Bowtie2. El porcentaje de alineamientos es mayor en el transcriptoma de Trinity que en el obtenido vía

SOAPdenovo-Trans, es decir, es de mejor calidad. Destaca también que el porcentaje de *reads* alineados de forma concordante más de una vez es de 38,57% en Trinity, mientras que con SOAPdenovo-Trans es de únicamente el 0,1%.

Bowtie2	**Trinity**	**SOAPdenovo-Trans**
Reads aligned concordantly 0 times	33.14%	69.76%
Reads aligned concordantly exactly 1 time	28.28%	30.13%
Reads aligned concordantly >1 times	38.57%	0.1%
Overall alignment rate	**83.66%**	**60.04%**

Tabla 2: Porcentajes de alineamientos obtenidos vía Bowtie2 para los transcriptomas de Trinity y SOAPdenovo-Trans.

Las puntuaciones de los dos transcriptomas según Transrate se muestran en la Tabla 3. La puntuación de Transrate (*Assembly Score*) es muy baja para ambos transcriptomas, ya que la puntuación máxima es de 1 y la mínima de 0. La puntuación tan baja de SOAPdenovo-Trans puede deberse a que, al haber considerado muchos más contigs pero de menor tamaño, hayan aparecido un mayor número de duplicados o fragmentaciones.

TRANSRATE v1.0.1	**Trinity**	**SOAPdenovo-Trans**
Transrate Assembly Score	0.0769	0.0087
Transrate Optimal Score	0.2667	0.1265
Transrate Optimal Cutof f	0.4762	0.3787
good contigs	272630	289412
p good contigs	0.3	0.1265

Tabla 3: Puntuaciones referentes a la calidad de los transcriptomas de Trinity y SOAPdenovo-Trans. , obtenidas vía Transrate. Se han seleccionado aquellos estadísticos más relevantes para el proyecto.

Se aplicó BUSCO en ambos transcriptomas, siendo considerado como más completo el transcriptoma ensamblado vía Trinity, como puede observarse en la Tabla 4. En este transcriptoma se han encontrado un 94,72% de BUSCOs completos, y si se incluyen los fragmentados, se han encontrado los 303 grupos de BUSCOs de eucariotas. En el transcriptoma de SOAPdenovo-Trans faltan el 3% de los BUSCOs.

BUSCO V2/V3 – Euk	Trinity	SOAPdenovo-Trans
Complete BUSCOs – C	94.72%	50.8%
Complete and single-copy BUSCOs – S	45.5%	45.5%
Complete and duplicated BUSCOs – D	49.2%	5.3%
Fragmented BUSCOs – F	5.3%	46.2%
Missing BUSCOs – M	0,00 %	3,00 %
Total BUSCO groups searched	303	303

Tabla 4: Puntuaciones referentes al porcentaje de BUSCOs encontrados en los transcriptomas de Trinity y SOAPdenovo-Trans. Se han seleccionado el grupo de BUSCOs perteneciente al grupo de eucariotas (Euk).

Por todos estos resultados, se considera que fue una buena idea emplear el transcriptoma de Trinity sin clusterizar en el resto de tareas del trabajo. En la Tabla 5 se muestran las estadísticas básicas de este transcriptoma, calculadas por Trinity. En la Tabla 6, se muestran los estadísticos N50, media y mediana de la longitud de los contigs, y el total de bases ensambladas, también calculados por Trinity, para todos los contigs y sólo teniendo el cuenta la isoforma más larga de cada "gen". Trinity llama"gen" a un grupo de transcritos relacionados.

Total transcritos	907236
Total "genes" de Trinity	534770
%GC	38.8

Tabla 5: Estadísticas básicas del transcriptoma ensamblado por Trinity. Estas estadísticas han sido proporcionadas por el propio programa, e incluye el número de transcritos identificados, el número total de "genes".

	Todos los contigs	Sólo isoforma más larga
N50	593	500
Mediana de longitud	320	300
Media de longitud	503.61	460.42
Bases totales	456897182	2462176

Tabla 6: Estadísticas básicas del transcriptoma ensamblado por Trinity. Estas estadísticas han sido proporcionadas por el propio programa, e incluye el estadístico N50, la media y mediana de longitud de los contigs y el total de bases ensambladas.

Los resultados de la anotación realizada por Trinotate se encuentran recogidos en una base de datos SQLite, alojada en el servidor del grupo de investigación.

FEELnc se aplicó al transcriptoma sin clusterizar de Trinity, al cual se le eliminaron ya las zonas identificadas como codificantes vía BLAST, como ya

se ha comentado, para hacer el archivo de *input* más pequeño y que diera tiempo a obtener resultados. Se eligió este transcriptoma por obtener una mejor puntuación que el elaborado por SOAPdenovo-Trans, pero en un futuro sería interesante aplicarlo también a éste último y observar las posibles diferencias. En la Tabla 7 se observan la cantidad de lncRNAs anotados por el programa en cada uno de los modos. Se observa que el *Cutoff* calculado por el programa es de 0.609 en el modo *shuffle*, y de 0.283 cuando se utiliza el *training set* de lncRNAs previamente identificados. En la Figura 7, se observan las curvas ROC resultado del método *shuffle* y y en la Figura 8 las curvas correspondientes al método usando el algoritmo *Random Forest* entrenado con lncRNAs ya anotados. Estas curvas nos indican una exactitud diagnóstica muy buena, puede comprobarse visualmente ya que se alejan de la diagonal (que indicaría identificación de lncRNAs al azar) y se acercan al punto 0,1 (que indicaría una exactitud perfecta). Además, el AUC es de casi 1 (valor máximo) para los dos métodos.

	Shuffle	**Algoritmo RF**
Cutoff	0.6086	0.2826
lncRNAs	14789	14637
mRNAs	27	179

Tabla 7: Cantidad de lncRNAs y mRNAs identificado por cada uno de los métodos del módulo codpot del programa FEELnc. Se indica el *cutoff* para cada uno de los modos.

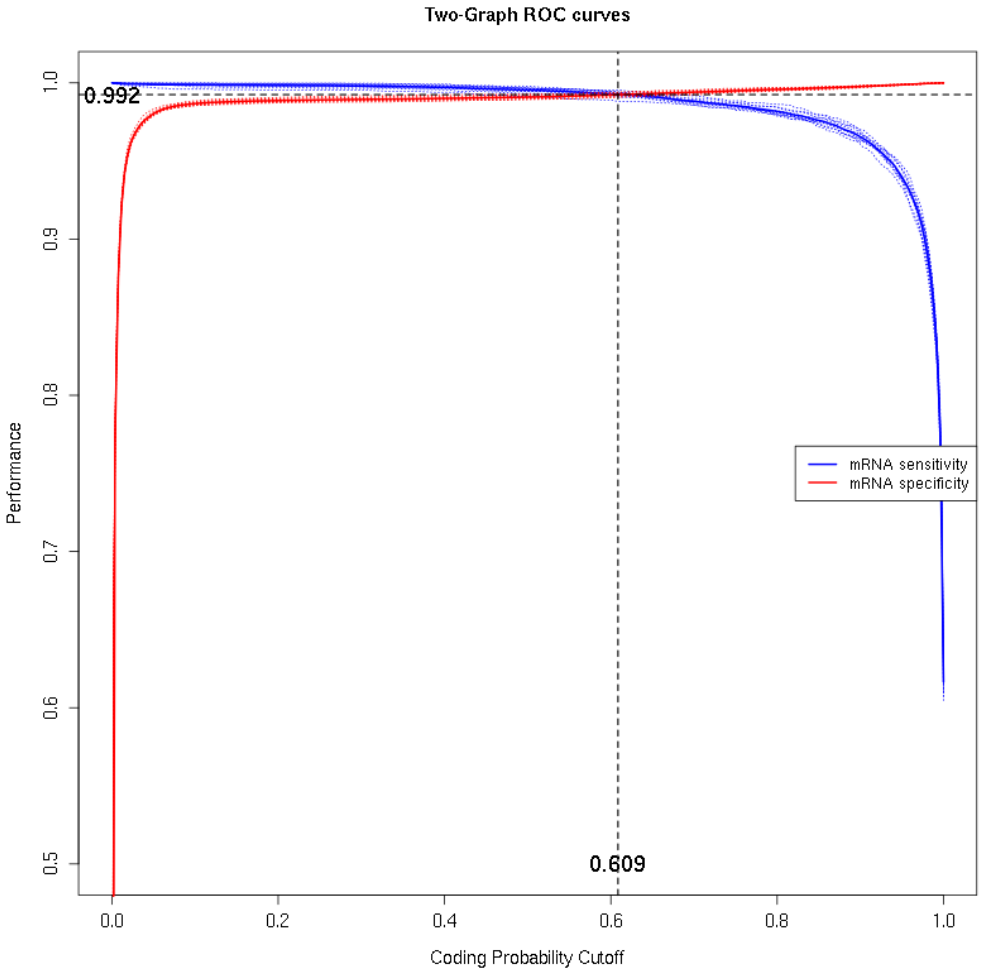

Figura 7: Curvas ROC resultado del modo *shuffle*, dentro del módulo codpot del programa FEELnc

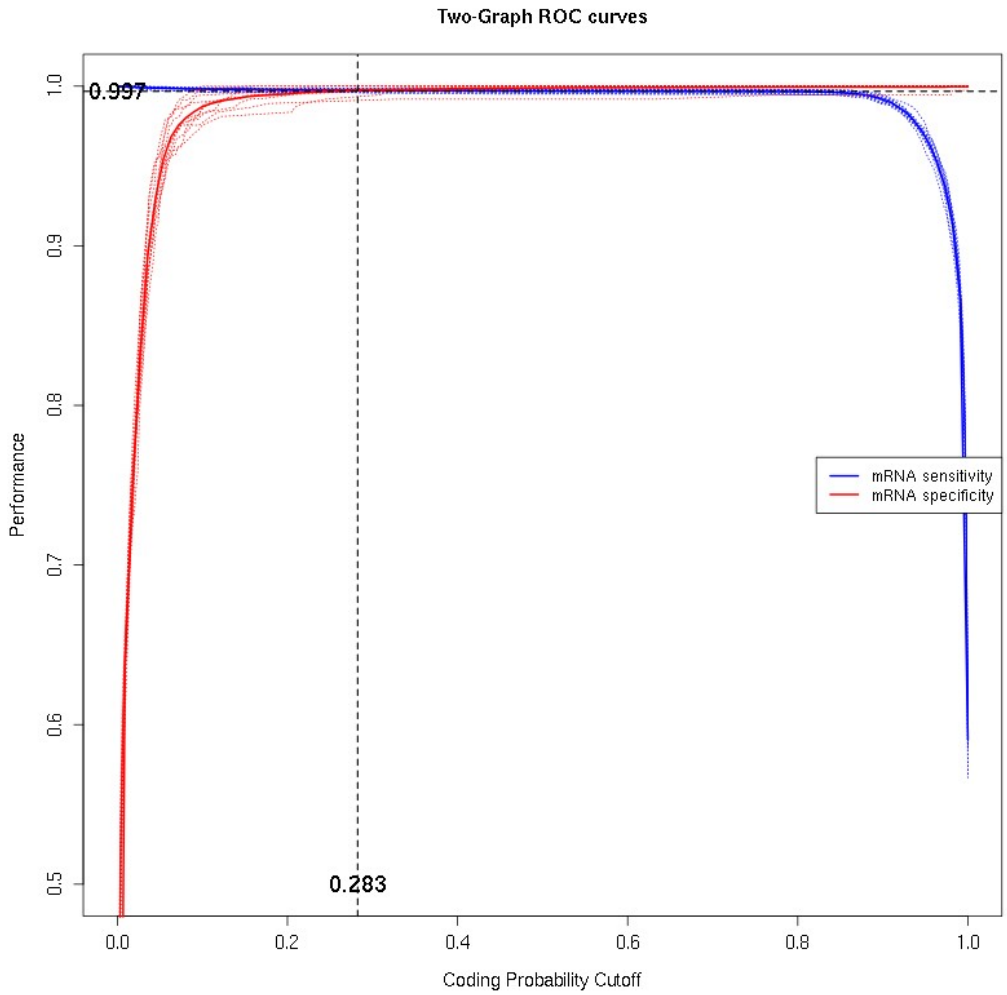

Figura 8: Curvas ROC resultado del modo referenciado, dentro del módulo codpot del programa FEELnc.

Para comparar los resultados obtenidos en ambos métodos, se realizó un blastn con un valor de e de $10e^{-3}$ de los lncRNAs obtenidos en el modo *shuffle* y en el modo referenciado contra las 9005 secuencias de lncRNAs identificadas el último estudio disponible para la especie[(6)]. Con los sets de ambos modos, se han encontrado resultados idénticos. Se han encontrado 3750 de los lncRNAs del estudio previo, siendo esto un 41,6% del total.

Un 42,8% de los lncRNAs en común aparecen fragmentados en nuestro set, es decir, coinciden con más de uno de nuestros transcritos. El porcentaje mínimo de identidad (*pident*) es de un 74,01%, siendo el promedio de un 93,2%. El *bit score* promedio es de 316,88, mientras que el *score* promedio es de 97,5.

En relación a la anotación funcional, la Figura 9 muestra las 10 proteínas identificadas por catRAPID como aquellas con mayor porcentaje de interacción con los lncRNAs anotados, para ambos sets producidos por FEELnc. Puede observarse que son las mismas proteínas en ambos casos. En la Tabla 8 se observa un resumen de las funciones de dichas proteínas.

Protein ID	RNA ID	Z-score	Discriminative Power (%)	Interaction Strength (%)	Domain	Motif	Ranking
NP_999787.1	TRINITY_DN41917_1	-0.18	45	84	no	no	
NP_999744.1	TRINITY_DN41917_1	-0.45	22	50	yes	no	
NP_001123287.1	TRINITY_DN41917_1	-0.53	20	32	yes	no	
NP_001075435.1	TRINITY_DN41917_1	-0.57	17	30	yes	no	
NP_999704.1	TRINITY_DN41917_1	-0.66	14	16	yes	no	
NP_001123279.1	TRINITY_DN41917_1	-0.76	14	8	yes	no	
NP_999824.2	TRINITY_DN41917_1	-0.83	14	8	yes	no	
NP_001229606.1	TRINITY_DN41917_1	-0.79	14	5	yes	no	
NP_001005725.1	TRINITY_DN41917_1	-0.83	14	5	yes	no	
NP_001124190.1_826-877	TRINITY_DN41917_1	-0.84	14	3	yes	no	

Protein ID	RNA ID	Z-score	Discriminative Power (%)	Interaction Strength (%)	Domain	Motif	Ranking
NP_999787.1	TRINITY_DN41917_1	-0.16	47	85	no	no	
NP_999744.1	TRINITY_DN41917_1	-0.43	22	53	yes	no	
NP_001123287.1	TRINITY_DN41917_1	-0.49	20	36	yes	no	
NP_001075435.1	TRINITY_DN41917_1	-0.54	17	34	yes	no	
NP_999704.1	TRINITY_DN41917_1	-0.64	14	18	yes	no	
NP_999824.2	TRINITY_DN41917_1	-0.81	14	9	yes	no	
NP_001123279.1	TRINITY_DN41917_1	-0.75	14	8	yes	no	
NP_001005725.1	TRINITY_DN41917_1	-0.80	14	6	yes	no	
NP_001229606.1	TRINITY_DN41917_1	-0.78	14	5	yes	no	
NP_001124190.1_826-877	TRINITY_DN41917_1	-0.83	14	4	yes	no	

Figura 9: Captura de pantalla de los resultados obtenidos por catRAPID. La tabla de arriba corresponde a los lncRNAs identificados con referencia, y la tabla inferior refleja los resultados de los lncRNAs identificados con el modo *shuffle*.

UnipProt entry	Función molecular	Proceso biológico
A2PZA6_STRPU	DNA-binding	Transcripción, regulación de la transcripción
Q26649_STRPU	DNA-binding	Regulación de la transcripción
B3FNS0_STRPU	DNA-binding, ecdysone binding	Transcripción, regulación de la transcripción, receptor en ruta de la ecdisona
A0A0B4J2U9_STRPU	Quinasa, Transferasa	Procesos metabólicos de ADP, AMP, GTP, ITP, purina
B3FNR8_STRPU	DNA-binding	Transcripción, regulación de la transcripción
B3VCG6_STRPU	Binding de ácidos nucleicos	
Q64HK6_STRPU	DNA-binding, RNA polimerasa II	Regulación positiva de la transcripción
Q9U0E3_STRPU		Movilidad celular por medio de cilios
O77156_STRPU	DNA-binding, RNA polimerasa II	Regulación positiva de la transcripción
Q6S5K1_STRPU	DNA-binding	Diferenciación celular, Regulación positiva de la transcripción

Tabla 8: Resumen de las proteínas con mayor porcentaje de interacción con los lncRNAS identificados y sus funciones.

4. Discusión

Los resultados obtenidos al analizar la calidad del transcriptoma ensamblado *de novo* avalan la idea de haber continuado la *pipeline* con el transcriptoma de Trinity, y no con el de SOAPdenovo-Trans, al ser la calidad del primero mucho mayor que la del segundo. En experimentos posteriores, se recomienda realizar la clusterización de los transcriptomas y repetir los análisis de control de calidad, para compararlos y ver si la clusterización mejora dicha calidad o no. En este caso, no se espera que la clusterización mejore en demasía el transcriptoma de Trinity porque ya presenta unas puntuaciones que lo identifican como de alta calidad, aunque sí podría mejorar su puntuación en Transrate. Las puntuaciones obtenidas por Trinity tanto en BUSCOs como en porcentaje de alineamiento sirven para dar por bueno el transcriptoma, y permiten continuar el flujo de trabajo sin dudas de que el transcriptoma de que se parte es válido. En estudios posteriores, podría valorarse la posibillidad de unir ambos transcriptomas, ya que se ha visto que el uso de un transcriptoma fruto de la unión («*merge*») de varios transcriptomas ensamblados *de novo* puede aumentar su calidad[(44)(45)].

En cuanto a los resultados de la anotación de los lncRNAs mediante FEELnc, se observa un número similar de lncRNAs anotados en ambos modos. Lo que sí varía de forma muy apreciable es el valor del *cutoff* calculado por el programa, siendo casi el doble en el modo *shuffle* que en el modo referenciado. Tiene sentido que el programa haya estimado un *cutoff* mucho mayor en el modo sin referencias, ya que lo único con lo que cuenta de entrada es con un conjunto de mRNAs, es decir, puede estimar a partir de dónde se considera codificante un transcrito en el modelo, y al no ajustar demasiado ese *cutoff* no estaría dejando fuera a posibles candidatos a lncRNAs, que tienen por definición un potencial de codificación bajo. Siguiendo esta lógica, calcula, como se observa en la Tabla 7, un número mayor de lncRNAs que en el modo referenciado. Ésta diferencia de lncRNAs entre el modo *shuffle* y el modo referenciado podrían considerarse entonces como falsos positivos, contando como positivo que un transcrito se identifique como lncRNA. Si todos los

lncRNAs “de más” que calcula el modo *shuffle* son falsos positivos, estaríamos hablando de un 1,03% de falsos positivos. Parece un error más que aceptable, si se tiene en cuenta que, de asumirlo, esta *pipeline* podría ayudar a identificar lncRNAs en multitud de organismos no modelo cuyo genoma no esté publicado. El porcentaje de error tan bajo también nos indica que el hecho de entrenar el algoritmo con lncRNAs conocidos no mejora mucho la predicción. Y esta mejora será aún menor cuando se empleen lncRNAs anotados en una especie cercana y no en la especie objetivo, que es el método común en este tipo de estudios.

El número de lncRNAs identificados en este trabajo, 14789 para el modo *shuffle* y 14367 para el modo referenciado, difiere de la cantidad anotada en estudios anteriores[(6)], que es de 9005. Esto parece indicar que la posible eliminación de transcritos con bajos niveles de expresión en el ensamblaje del transcriptoma *de novo* no sería un problema a la hora de anotar lncRNAs. La diferencia en la cantidad de lncRNAs podría deberse al uso en dichos estudios de la versión 4.2 del genoma de *S. purpuratus*. Para comprobar esta teoría, se realizó el proceso en modo *shuffle* con el mismo *input*, pero utilizando como cebador en *shuffle* los mRNAs anotados en dicha versión del genoma. Se obtuvieron 4209 lncRNAs, cifra algo más cercana a los 9005 transcritos de lncRNAs que se anotan en el último artículo disponible sobre esta especie[(46)]. Oficialmente, en la anotación de Spur_4.2 se idenficaron 3,487 lncRNAs (*release* 101)[(49)]. Es decir, aún utilizando la misma versión del genoma que en el artículo con el que se comparan nuestros resultados, se obtiene una cantidad diferente de lncRNAs, si bien esa cantidad es más similar. Por tanto, no puede decirse que la diferencia se deba, al menos únicamente, al uso de una versión diferente del genoma. El mayor número de lncRNAs obtenidos mediante el *de novo* transciptome podría deberse a la fragmentación de algunas anotaciones, como sugieren los resultados del blast.

Los resultados de la comparación entre nuestros sets de lncRNAs y los lncRNAs anotados con genoma de referencia (41,6% de los lncRNAs detectados), arrojan un porcentaje bajo de transcritos en común pero dentro de lo esperable, ya que puede ser que nuestras anotaciones estén fragmentadas.

Se han detectado por tanto menos de la mitad de los lncRNAs ya anotados para *S. purpuratus*, siendo una posible línea de investigación futura averiguar la razón por la cual se han obtenido estos resultados.

El hecho de que un 42,8% de los lncRNAs de la base de datos de blast aparezcan fragmentados tiene sentido, ya que en nuestro caso se ha realizado un transcriptoma *de novo*, y como se ha visto (en nuestro caso entre SOAPdenovo-Trans y Trinity), la longitud de los transcritos ensamblados varía según el ensamblador. Es decir, puede ser que esta fragmentación se deba en parte a que los transcritos de nuestro transcriptoma *de novo* son más cortos que los obtenidos a partir del genoma de referencia. En estudios futuros debe por tanto valorarse el uso de programas que reduzcan esta fragmentación, como CD-HIT.

Los resultados de la caracterización nos dan una idea del papel de los lncRNA anotados. Estarían involucrados en la regulación de la transcripción, coincidiendo con lo encontrado en estudios previos[(2)(3)]. Llama la atención su possible relación con la movilidad celular, de lo cual apenas se han encontrado referencias, pero existen pruebas del rol epigenético que pueden tener los lncRNAs en la reproducción, por ejemplo, influyendo en la movilidad de los espermatozoides[(50)].

5.Conclusiones

Los resultados de este proyecto darían una respuesta afirmativa a la pregunta que originó el TFM: **sí sería posible anotar lncRNAs de una especie sin disponer de genoma de referencia a partir del ensamblaje *de novo* de su transcriptoma**. La clave estaría en conseguir un transcriptoma *de novo* de alta calidad. Si con un transcriptoma *de novo* de calidad mejorable como el nuestro se han anotado aproximadamente el 42% de los lncRNAs anotados en un estudio en el cual se emplea un genoma de referencia, todo parece indicar que ese porcentaje aumentaría al incrementar la calidad del transcriptoma *de novo*. Por tanto, estos resultados deberían tenerse en cuenta como preliminares; antes de poder dar una respuesta definitiva a la pregunta que ha originado este trabajo sería recomendable repetir los análisis sin el límite de tiempo que supone la entrega de la presente memoria. Esta limitación, más los diversos problemas técnicos surgidos durante el desarrollo del proyecto, han hecho que en ocasiones se haya tenido que emplear una solución que, si bien no es errónea, no sería la ideal. Esto puede aplicarse por ejemplo a la eliminación previa de todos los mRNAs del transcriptoma de Trinity gracias a que ya están identificados en el transcriptoma de *S. purpuratus* bajo la publicación del genoma Spur_5.0, o al uso de CD-HIT para mejorar la calidad del transcriptoma *de novo*.

Todos los objetivos planteados en la planificación del proyecto han sido alcanzados, habiendo sido capaces de adaptar la planificación cuando ha sido necesario. Como ya se ha comentado, para conseguir acabar a tiempo se han tenido que desarrollar acciones de mitigación y desviaciones de la *pipeline* original descritas en el capítulo 2. Estas desviaciones han sido posibles gracias a que la metodología prevista para el TFM fue adecuada: se marcaron los pasos a seguir y objetivos a lograr de forma clara y concisa, lo que permitió buscar opciones alternativas en los momentos en que el procesado de datos parecía encallado con un determinado *software*.

En un futuro, se recomienda repetir los análisis realizados en este proyecto sin límite de tiempo, sin tener en cuenta que los genes codificantes para proteínas están ya identificados, siguiendo la siguiente ruta :

- en transcriptoma: transdecoder en transcriptoma (longestORF y predict) ->blastx (Uniprot)
- en longestORF.pep: blastp (uniprot) -> hmmscan -> signalp

La caracterización de los lncRNAs también podría mejorarse en proyectos futuros mediante el uso del módulo *classifier* de FEELnc, que permite identificar clases de lncRNAs.

Por último, se recomienda también emplear esta misma *pipeline* en una especie similar sin genoma publicado, como puede ser *Paracentrotus lividus*. Esto nos permitiría no sólo comprobar que efectivamente se pueden anotar lncRNAs sin disponer de genoma de referencia, sino comparar las caracterizaciones y clases de los lncRNAs identificados. Convendría también repetir la *pipeline* con el transcriptoma de Trinity aquí obtenido tras ser procesado por CD-HIT, y comparar los resultados.

6. Glosario

- **Adaptador**: secuencias utilizadas para que la librería de Illumina se una a las secuencias de la muestra.
- **Alignment**: alineamiento, forma de organizar las secuencias para identificar regiones similares, ya sean similitudes estructurales, funcionales o evolutivas.
- **Clusterización**: unión en grupos de transcritos que se supone codifican para peptidos o proteínas similares.
- **Contig**: conjunto de secuencias de ADN superpuestas usadas para crear un mapa que reconstruya la secuencia original de un cromosoma o una región cromosómica.
- **Core**: anglisicismo utilizacon en informática que se refiere al núcleo del procesador o CPU.
- **Insert**: secuencia de ADN o ARN que está "insertada" entre los adaptadores.
- **Job**: Proceso informático que se envía al servidor y no se termina ("*detach*") cuando se apaga la máquina local
- **mRNA**: ARN mensajero.
- **ORF**: Opening Reading Frame. Secuencia comprendida entre un codón de inicio y otro de terminación de la traducción, excluyendo los intrones.
- **Pipeline**: conjunto de pasos y procesos bioinformáticos ordenados necesarios para el tratamiento de los datos iniciales y la obtención del resultado deseado.
- **Read**: secuencia de ADN identificada por los secuenciadores.
- **RNA-Seq**: RNA-Sequencing. Técnica que identifica la cantidad y secuencias de RNA en una muestra empleando lo que se conoce como Next Generation Sequencing (NGS). Se emplea tanto para estudiar transcriptomas como para analizar patrones de expresión génica codificados en el ARN[47].
- **Training set**: conjunto de datos de muestra correctamente etiquetados e identificados empleados para entrenar un algoritmo de *machine learning*.
- **Trimming**: eliminación de los finales de las secuencias con la idea de eliminar las secuencias correspondientes a los adaptadores.

7. Bibliografía

(1) Tripathi, R., Chakraborty, P., & Varadwaj, P. K. (2017). Unraveling long non-coding RNAs through analysis of high-throughput RNA-sequencing data. *Non-coding RNA research*, 2(2), 111-118.

(2) Long, Y., Wang, X., Youmans, D. T., & Cech, T. R. (2017). How do lncRNAs regulate transcription?. *Science advances*, *3*(9), eaao2110.

(3) Perry, R. B. T., & Ulitsky, I. (2016). The functions of long noncoding RNAs in development and stem cells. *Development*, *143*(21), 3882-3894.

(4) Housman, G., & Ulitsky, I. (2015). Methods for distinguishing between protein-coding and long noncoding RNAs and the elusive biological purpose of translation of long noncoding RNAs. *BioRxiv*, 017889.

(5) Goff, L. A., & Rinn, J. L. (2015). Linking RNA biology to lncRNAs. *Genome research*, 25(10), 1456-1465.

(6) Hezroni, H., Koppstein, D., Schwartz, M. G., Avrutin, A., Bartel, D. P., & Ulitsky, I. (2015). Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell reports*, *11*(7), 1110-1122.

(7) Jia, H., Osak, M., Bogu, G. K., Stanton, L. W., Johnson, R., & Lipovich, L. (2010). Genome-wide computational identification and manual annotation of human long noncoding RNA genes. *Rna*, *16*(8), 1478-1487.

(8) Chen, B., Zhang, Y., Zhang, X., Jia, S., Chen, S., & Kang, L. (2016). Genome-wide identification and developmental expression profiling of long noncoding RNAs during Drosophila metamorphosis. *Scientific reports*, *6*(1), 1-8.

(9) Gupta, M. K., Donde, R., Gouda, G., Vadde, R., & Behera, L. (2019). De novo assembly and characterization of transcriptome towards understanding molecular mechanism associated with MYMIV-resistance in Vigna mungo-A computational study. *BioRxiv*, 844639.

(10) Harris, Z. N., Kovacs, L. G., & Londo, J. P. (2017). RNA-seq-based genome annotation and identification of long-noncoding RNAs in the grapevine cultivar ‘Riesling’. *BMC genomics*, 18(1), 937.

(11) Ceschin, D. G., Pires, N. S., Mardirosian, M. N., Lascano, C. I., & Venturino, A. (2020). The Rhinella arenarum transcriptome: de novo assembly, annotation and gene prediction. *Scientific reports*, *10*(1), 1-8.

(12) https://www.ncbi.nlm.nih.gov/bioproject/PRJNA81157Última visita: 27/12/2020

(13) Tu, Q., Cameron, R. A., & Davidson, E. H. (2014). Quantitative developmental transcriptomes of the sea urchin Strongylocentrotus purpuratus. *Developmental biology*, *385*(2), 160-167.

(14) Tu, Q., Cameron, R. A., Worley, K. C., Gibbs, R. A., & Davidson, E. H. (2012). Gene structure in the sea urchin Strongylocentrotus purpuratus based on transcriptome analysis. *Genome research*, *22*(10), 2079-2087.

(15) https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ Última visita: 27/12/2020

(16) Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. "Trimmomatic: a flexible trimmer for Illumina sequence data." *Bioinformatics* 30.15 (2014): 2114-2120.

(17) Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047-3048.

(18) Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., ... & Chen, Z. (2011). Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature biotechnology*, *29*(7), 644.

(19) https://www.rna.uni-jena.de/supplements/assembly/ Última visita: 27/12/2020

(20) Hölzer, M., & Marz, M. (2019). De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. *Gigascience*, *8*(5), giz039.

(21) Zhao, Q. Y., Wang, Y., Kong, Y. M., Luo, D., Li, X., & Hao, P. (2011, December). Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. I*BMC bioinformatics* 12(S2). BioMed Central.

(22) Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., ... & Zhou, X. (2014). SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics*, *30*(12), 1660-1666.

(23) Lebedev, E. E., Ostromyshenskii, D. I., Solovyeva, A. I., Turenko, A. S., Drozdov, A. L., Podgornaya, O. I., & Adonin, L. S. (2019). The Transposons of the Sea Urchin Strongylocentrotus intermedius Agassiz, 1863: In Silico Versus In Vitro. *Russian Journal of Marine Biology*, *45*(6), 418-424.

(24) http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml Última visita: 27/12/2020

(25) https://cran.r-project.org/ Última visita: 27/12/2020

(26) Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J. M., & Kelly, S. (2016). TransRate: reference-free quality assessment of de novo transcriptome assemblies. *Genome research*, *26*(8), 1134-1144.

(27) https://github.com/trinityrnaseq/trinityrnaseq/wiki/RNA-Seq-Read-Representation-by-Trinity-Assembly Última visita: 27/12/2020

(28) Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210-3212.

(29) https://gvolante.riken.jp/analysis.html Última visita: 27/12/2020

(30) Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, *28*(23), 3150-3152.

(31) http://weizhongli-lab.org/cd-hit/ Última visita: 27/12/2020

(32) Musacchia, F., Basu, S., Petrosino, G., Salvemini, M., & Sanges, R. (2015). Annocript: a flexible pipeline for the annotation of transcriptomes able to identify putative long noncoding RNAs. *Bioinformatics*, *31*(13), 2199-220

(33)https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Strongylocentrotus_purpuratus/102/ Última visita: 27/12/2020

(34) Madden, T. (2013). The BLAST sequence analysis tool. In *The NCBI Handbook [Internet]. 2nd edition*. National Center for Biotechnology Information (US).

(35) Wucher, V., Legeai, F., Hedan, B., Rizk, G., Lagoutte, L., Leeb, T., ... & Cirera, S. (2017). FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic acids research*, *45*(8), e57-e57.

(36) RColorBrewer, S., & Liaw, M. A. (2018). Package 'randomForest'. *University of California, Berkeley: Berkeley, CA, USA*.

(37) Jiang, M., Anderson, J., Gillespie, J., & Mayne, M. (2008). uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts. *BMC bioinformatics*, *9*(1), 192.

(38) https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/002/235/GCF_000002235.5_Spur_5.0/ Última visita: 27/12/2020

(39) http://openaccess.uoc.edu/webapps/o2/bitstream/10609/97967/6/dmiquelbTM0619mem%C3%B2ria.pdf Última visita: 27/12/2020

(40) Kirk, J. M. (2019). Functional Classification of Long Non-coding RNAs. *UNC*. Chapell Hill, North Carolina.

(41) Kirk, J. M., Kim, S. O., Inoue, K., Smola, M. J., Lee, D. M., Schertzer, M. D., ... & Horning, C. R. (2018). Functional classification of long non-coding RNAs by k-mer content. *Nature genetics*, 50(10), 1474-1482.

(42)http://service.tartaglialab.com/static_files/algorithms/crossalign/documentation.html Última visita: 27/12/2020

(43) http://s.tartaglialab.com/static_files/shared/documentation.html Última visita: 27/12/2020

(44) Sadat-Hosseini, M., Bakhtiarizadeh, M. R., Boroomand, N., Tohidfar, M., & Vahdati, K. (2020). Combining independent de novo assemblies to optimize leaf transcriptome of Persian walnut. *PloS one*, *15*(4), e0232005.

(45) Nakasugi, K., Crowhurst, R., Bally, J., & Waterhouse, P. (2014). Combining transcriptome assemblies from multiple de novo assemblers in the allo-tetraploid plant Nicotiana benthamiana. *PloS one*, *9*(3), e91776.

(46)https://www.cell.com/cms/10.1016/j.celrep.2015.04.023/attachment/b304e66b-2467-43ef-bc01-84866af8ee11/mmc3.xlsx Última visita: 27/12/2020

(47) https://www.technologynetworks.com/genomics/articles/rna-seq-basics-applications-and-protocol-299461 Última visita: 27/12/2020

(48) https://github.com/Trinotate/Trinotate.github.io/wiki Última visita: 27/12/2020

(49) https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Strongylocentrotus_purpuratus/101/ Última visita: 27/12/2020

(50) Liu, Y., Sun, Y., Li, Y., Bai, H., Xue, F., Xu, S., ... & Chen, J. (2017). Analyses of long non-coding RNA and mRNA profiling using RNA sequencing in chicken testis with extreme sperm motility. *Scientific reports*, *7*(1), 1-8.

8. Anexos

ANEXO I: Tabla resumen de secuencias empleadas para la elaboración del transcriptoma.

Sample Name	% Dups	% GC	Length	M Seqs
SRR531843_1P	41.2%	40%	75 bp	35.0
SRR531843_1U	11.2%	43%	58 bp	2.8
SRR531843_2P	40.6%	40%	74 bp	35.0
SRR531843_2U	4.8%	36%	67 bp	0.2
SRR531853_1P	37.8%	46%	73 bp	19.2
SRR531853_1U	6.9%	50%	59 bp	0.5
SRR531853_2P	35.0%	46%	71 bp	19.2
SRR531853_2U	3.9%	47%	61 bp	0.1
SRR531860_1P	47.8%	43%	75 bp	44.3
SRR531860_1U	13.3%	45%	61 bp	2.0
SRR531860_2P	46.8%	43%	74 bp	44.3
SRR531860_2U	3.4%	43%	65 bp	0.3
SRR531948_1P	35.7%	46%	73 bp	19.9
SRR531948_1U	6.7%	50%	59 bp	0.6
SRR531948_2P	34.2%	46%	71 bp	19.9
SRR531948_2U	5.2%	47%	61 bp	0.1
SRR531949_1P	44.7%	43%	75 bp	42.6
SRR531949_1U	13.0%	45%	62 bp	1.8
SRR531949_2P	44.5%	43%	75 bp	42.6
SRR531949_2U	4.3%	43%	66 bp	0.2
SRR531950_1P	42.0%	46%	74 bp	27.6
SRR531950_1U	10.8%	48%	56 bp	1.0
SRR531950_2P	40.9%	46%	74 bp	27.6
SRR531950_2U	8.8%	46%	63 bp	0.1
SRR531951_1P	44.4%	39%	73 bp	34.1
SRR531951_1U	15.4%	41%	58 bp	4.0
SRR531951_2P	44.0%	39%	73 bp	34.1
SRR531951_2U	11.7%	39%	71 bp	1.6
SRR531952_1P	41.3%	49%	74 bp	22.4
SRR531952_1U	24.9%	50%	62 bp	1.2
SRR531952_2P	44.7%	49%	74 bp	22.4
SRR531952_2U	5.9%	49%	71 bp	0.0
SRR531953_1P	39.8%	39%	74 bp	32.1
SRR531953_1U	9.7%	40%	68 bp	4.2
SRR531953_2P	38.0%	39%	72 bp	32.1
SRR531953_2U	10.6%	39%	74 bp	2.8

Sample Name	% Dups	% GC	Length	M Seqs
SRR531954_1P	48.6%	50%	74 bp	22.4
SRR531954_1U	28.5%	50%	63 bp	1.1
SRR531954_2P	54.0%	50%	74 bp	22.4
SRR531954_2U	10.5%	51%	71 bp	0.0
SRR531955_1P	37.7%	40%	74 bp	36.6
SRR531955_1U	9.6%	43%	60 bp	2.8
SRR531955_2P	36.2%	40%	73 bp	36.6
SRR531955_2U	5.7%	39%	70 bp	1.0
SRR531956_1P	43.7%	49%	74 bp	27.0
SRR531956_1U	16.1%	49%	65 bp	1.1
SRR531956_2P	49.4%	49%	75 bp	27.0
SRR531956_2U	8.7%	49%	71 bp	0.0
SRR531957_1P	44.0%	39%	75 bp	35.8
SRR531957_1U	10.2%	42%	58 bp	1.9
SRR531957_2P	44.1%	39%	74 bp	35.8
SRR531957_2U	7.2%	36%	64 bp	0.2
SRR531958_1P	56.3%	44%	75 bp	40.4
SRR531958_1U	15.7%	47%	62 bp	0.9
SRR531958_2P	55.4%	44%	75 bp	40.4
SRR531958_2U	7.8%	44%	67 bp	0.2
SRR531964_1P	42.3%	49%	74 bp	22.5
SRR531964_1U	15.0%	49%	62 bp	1.0
SRR531964_2P	48.2%	49%	74 bp	22.5
SRR531964_2U	9.4%	49%	72 bp	0.0
SRR531996_1P	44.3%	44%	74 bp	30.0
SRR531996_1U	10.9%	45%	55 bp	3.1
SRR531996_2P	43.3%	44%	73 bp	30.0
SRR531996_2U	6.6%	44%	68 bp	0.1
SRR532046_1P	38.7%	39%	75 bp	34.2
SRR532046_1U	8.3%	42%	58 bp	1.9
SRR532046_2P	39.2%	39%	75 bp	34.2
SRR532046_2U	6.3%	37%	66 bp	0.2
SRR532055_1P	48.9%	46%	75 bp	44.6
SRR532055_1U	14.3%	47%	61 bp	1.4
SRR532055_2P	48.1%	46%	74 bp	44.6
SRR532055_2U	4.9%	45%	65 bp	0.2
SRR532074_1P	36.8%	45%	75 bp	14.3
SRR532074_1U	5.4%	46%	56 bp	1.0
SRR532074_2P	36.7%	45%	75 bp	14.3
SRR532074_2U	4.5%	43%	73 bp	0.1
SRR532121_1P	49.5%	40%	75 bp	36.7

Sample Name	% Dups	% GC	Length	M Seqs
SRR532121_1U	15.2%	44%	60 bp	1.9
SRR532121_2P	48.5%	40%	74 bp	36.7
SRR532121_2U	12.2%	39%	71 bp	0.5
SRR532143_1P	42.3%	42%	73 bp	36.6
SRR532143_1U	17.8%	45%	58 bp	3.0
SRR532143_2P	41.5%	42%	73 bp	36.6
SRR532143_2U	11.6%	42%	70 bp	1.2
SRR532151_1P	54.3%	48%	73 bp	30.8
SRR532151_1U	15.3%	48%	54 bp	3.4
SRR532151_2P	52.0%	48%	72 bp	30.8
SRR532151_2U	15.2%	48%	67 bp	0.1
SRR533746_1P	40.0%	40%	72 bp	36.0
SRR533746_1U	13.0%	44%	57 bp	2.5
SRR533746_2P	39.8%	40%	73 bp	36.0
SRR533746_2U	10.8%	41%	70 bp	1.3