Comparación de métodos de cálculo de expresión génica basados en RNA-seq.

María Elena Campoy García
Máster en Bioinformática y Bioestadística
Análisis bioinformático de exomas en cáncer

Daniel Aguilar Villalba
Carles Ventura Royo
02/01/2019

Esta obra está sujeta a una licencia de Reconocimiento-NoComercial-SinObraDerivada 3.0 España de Creative Commons

FICHA DEL TRABAJO FINAL

Título del trabajo:	Comparación de métodos de cálculo de expresión génica basados en RNA-seq.
Nombre del autor:	*María Elena Campoy García*
Nombre del consultor/a:	*Daniel Aguilar Villalba*
Nombre del PRA:	*Carles Ventura Royo*
Fecha de entrega (mm/aaaa):	01/2019
Titulación:	*Máster en Bioinformática y Bioestadística*
Área del Trabajo Final:	*M0.180-TFM-Bioinformática y Bioestadística. Area 4, aula 1.*
Idioma del trabajo:	*Español*
Palabras clave	*RNA-seq, expresión génica, software.*

Resumen del Trabajo.

Debido al auge de la tecnología RNA-seq, múltiples programas de análisis de expresióngenética se encuentran disponibles. En el grupo de investigación donde se ha realizado este proyecto, el software usado para analizar datos de RNA-seq es STAR y RSEM. Dado que RSEM necesita un paso previo ejecutado por STAR, es interesante comparar su rendimiento con el de programas que no lo necesitan como Kallisto, Sailfish y Salmon. Por tanto, se han analizado 17 muestras paired-end procedentes de controles sanos de un estudio anterior. Los resultados obtenidos muestran que RSEM es el programa que más tiempo emplea en generar los resultados debido a su dependencia del mapeado de STAR y que, a pesar de tener un porcentaje superior de observaciones mapeadas, sus estimaciones de expresión son muy parecidas a las generadas por Kallisto, Sailfish y Salmon, sobre todo a nivel de gen y no de transcritos. Estos 3 últimos programas, debido a la similitud en su funcionamiento, muestran

unas estimaciones muy parecidas. En un futuro sería recomendable profundizar en esta comparación realizando un análisis paralelo con datos single-end y observando las diferencias en el reconocimiento de ARNs pequeños por cada método.

Abstract:

Due to the rise of the RNA-seq technology, multiple software related to expression analysis is available. In the group of investigation where this project has been elaborated, RSEM and STAR are the programs used to quantify the expression of isoforms and genes. RSEM needs a prior step by STAR so it would be of interest to make a comparison between the performance of this program and Kallisto, Sailfish and Salmon which do not need a prior step. Therefore, 17 paired-end samples from a prior project have been analysed. The results show that RSEM is the program that spends more time in performing the analysis owing to its dependency of the STAR step and in spite of a bigger percentage of mapped reads, the expression estimates are very similar to those generated by Kallisto, Sailfish and Salmon. Besides, the results are more similar in terms of gen instead of transcripts. These last 3 programs, which are very similar in their running, show very similar estimates. In the future, it would be recommendable to go in depth in this comparison performing the same analysis with single-end samples and looking at the differences in the recognition of small RNAs by each method.

Índice

1. Introducción 1
1.1 Contexto y justificación del Trabajo 1
1.2 Objetivos del Trabajo 1
1.3 Enfoque y método seguido 2
1.4 Planificación del Trabajo 2
1.5 Breve sumario de productos obtenidos 5
1.6 Breve descripción de los otros capítulos de la memoria 5
2. Resto de capítulos 5
2.1 Introducción 5
2.2 Materiales y métodos 14
2.3 Resultados 18
2.4 Discusión 31
3. Conclusiones 38
3.1 Cumplimiento de los objetivos 38
3.2 Seguimiento de la planificación 38
3.3 Vistas al futuro 39
4. Glosario 40
5. Bibliografía 41
6. Anexos 45

Lista de figuras

1. Workflow del procedimiento de análisis en RNA-seq.
2. Workflow del análisis bioinformático llevado a cabo en el proyecto.
3. Boxplot de la distribución de los tiempos de ejecución del análisis.
4. Boxplot del porcentaje de *reads* mapeados por RSEM, Sailfish y Salmon.
5. Diagrama de barras del porcentaje de genes y transcritos detectados por cada programa.
6. Distribución de las medidas de expresión por transcritos entre RSEM y Kallisto.
7. Distribución de la desviación estándar de cada transcrito entre RSEM y Kallisto.
8. Boxplot de la desviación estándar por transcritos de cada programa.
9. Relación entre la expresión y la desviación estándar de cada transcrito en RSEM.
10. Relación entre el coeficiente de variación de cada transcrito entre RSEM y Kallisto.
11. Distribución de las medidas de expresión por genes entre RSEM y Kallisto.
12. Distribución de la desviación estándar de cada gen entre RSEM y Kallisto.
13. Boxplot de la desviación estándar por genes de cada programa.
14. Relación entre la expresión y la desviación estándar de cada gen en RSEM.
15. Relación entre el coeficiente de variación de cada gen entre RSEM y Kallisto.
16. Boxplot de los tiempos de ejecución de cada programa al analizar archivos con menor número de *reads*.
17. Diagrama de barras del porcentaje de genes y transcritos detectados por cada programa al analizar archivos con menor número de *reads.*

1. Introducción

1.1 Contexto y justificación del Trabajo

Este TFM surge por la necesidad de mejorar el análisis de datos de RNA-seq (la aplicación de cualquier técnica de secuenciación de nueva generación al estudio del ARN[1]). El avance y desarrollo de algoritmos bioinformáticos ha hecho que hoy en día exista una gran disponibilidad de herramientas para ejecutar cualquier tipo de análisis. Tener un conocimiento adecuado de los recursos que podemos utilizar para llevar a cabo nuestro trabajo nos permite obtener un mejor rendimiento y unos resultados más ajustados a la realidad. Es decir, realizar una comparación entre los programas disponibles para analizar datos RNA-seq con el pipeline (encadenamiento de acciones necesarias para realizar un análisis informático) de referencia del laboratorio, permitirá aumentar el conocimiento sobre el uso de toda esta tecnología, mejorando y aumentando la eficacia. Actualmente, en el grupo de investigación Genómica Traslacional del Cáncer Colorrectal (IDIBAPS/Hospital Clínic de Barcelona) se utilizan STAR y RSEM para llevar a cabo el mapeado (asociación de cada fragmento de ADN complementario a la zona del genoma a la que corresponde) y cuantificación de los datos (cálculo de cuántos fragmentos se unen a cada zona del genoma). A partir de la comparación, el grupo podrá conocer las fortalezas y debilidades de cada software y escoger el que mejor se ajuste a los datos.

1.2 Objetivos del Trabajo

1. Objetivos generales

a) Ejecutar correctamente el software de análisis de datos derivados de RNA-seq.

a.1) Búsqueda e instalación del software disponible para el análisis de datos derivados de RNA-seq: STAR, RSEM, Kallisto, Salmon, Sailfish.

a.2) Establecimiento de un pipeline adecuado con cada uno de los programas.

b) Comparar el rendimiento de diferentes programas de análisis de expresión de genes basado en RNA-seq.

b.1) Análisis de la base de datos con los pipelines establecidos.

b.2) Comparación en conjunto de los resultados de cada análisis.

b.3) Extracción de la explicación biológica y técnica de los resultados.

1.3 Enfoque y método seguido

El enfoque escogido para llevar a cabo el proyecto es secuencial:

- Primero se llevará a cabo la instalación del software a utilizar. Una vez instalados, se crearán los pipelines a utilizar posteriormente a partir de muestras más pequeñas que las definitivas. Esto permitirá detectar posibles errores de ejecución o de instalación en los programas más rápidamente que si se ejecutara con los datos finales (17 muestras pertenecientes a controles sanos de un estudio anterior), ya que al ser estas más grandes, el análisis tarda más en ejecutarse y se emplearía más tiempo en la detección de errores derivados de un mal uso del software.
- En segundo lugar, se realizará el análisis de las muestras definitivas cuando lo anterior quede establecido. En este punto solamente se detectarían errores derivados de la propia base de datos, no del software (habrían sido detectados previamente).
- Por último, al tener los resultados de este análisis, se compararán con el software de referencia STAR + RSEM para conseguir el objetivo principal del trabajo: observar las ventajas e inconvenientes de cada programa para poder escoger el más adecuado para cada situación.

1.4 Planificación del Trabajo

4.1 Tareas

Para buscar e instalar el software necesario se emplearán 3 semanas.

Conseguir el pipeline adecuado para cada programa se llevará a cabo junto con el primer paso.

En la segunda parte del trabajo, se analizará una misma base de datos con lo obtenido previamente. Es estima una duración de 2 semanas.

Finalmente, la comparación de los resultados y extracción del significado biológico se llevará a cabo en una semana.

4.2 Calendario

En el calendario se puede observar, en negro, las desviaciones respecto al calendario original (en azul oscuro).

	PEC 1		PEC 2					PEC 3				PEC 4		PEC 5 a	PEC 5 b
			16/10 – 19/11					20/11 – 17/12				18/12-02/01		Elaboración presentación	Defensa publica
	02/10 – 15/10		16-22/10	23-29/10	30/10-5/11	6-12/11	13-19/11	20-26/11	27/11-3/12	4-10/12	11-17/12	18-24/12	25/12-02/01	03-10/01	14-23/01
Plan de trabajo															
a.1) Búsqueda e instalación del software disponible para el análisis de datos derivados de RNA-seq: STAR, Rsem, Kallisto, Salmon, Sailfish.															
a.2) Establecimiento de un pipeline adecuado con cada uno de los programas.															
b.1) Análisis de la base de datos con los pipelines establecidos.															
b.2) Comparación en conjunto de los resultados de cada análisis.															
b.3) Extracción de la explicación biológica y técnica de los resultados.															
Redacción de la memoria y presentación.															
Elaboración de la presentación.															
Defensa pública.															

1.5 Breve sumario de productos obtenidos

Los resultados finales del proyecto se presentarán de la manera especificada en el plan docente: se entregará el plan de trabajo, la memoria, la presentación virtual y la autoevaluación del proyecto.

1.6 Breve descripción de los otros capítulos de la memoria.

La estructura de la memoria del proyecto estará compuesta de varios apartados. El primero de ellos será la introducción donde se tratará sobre los aspectos generales del proyecto (contexto, justificación, objetivos, planificación...). A continuación, se tratarán los distintos apartados que componen normalmente un trabajo científico: una introducción más extensa sobre el estado actual del tema a tratar, materiales y métodos donde se explicará con detalle los datos y programas usados y cómo se han usado, y, por último, un tercer apartado de resultados donde se expondrá lo que se ha obtenido, además de la discusión de los mismos. Para finalizar, habrá un apartado de conclusión, glosarios y términos, bibliografía y anexos

2. Resto de capítulos

2.1 Introducción

Una de las áreas de interés de la biología molecular siempre ha sido la transcriptómica, término que no fue acuñado hasta 2009, por Wang y su equipo[2] como el conjunto completo de transcritos que hay en una célula y su cantidad en cierto momento del desarrollo o bajo condición fisiológica, siendo sus objetivos principales catalogar todos los transcritos (incluyendo ARN mensajero (ARNm), ARN no codificante (non-coding RNA) y ARN pequeño (small RNA)), determinar la estructura transcripcional de los genes y cuantificar los cambios en los niveles de expresión de cada transcrito durante el desarrollo y bajo distintas circunstancias[2,3].

En la década de los 90[4] surgieron los microarrays como técnica para analizar la expresión diferencial de genes a nivel de ARN entre grupos en distintos estados de desarrollo o entre grupos sanos y enfermos[3]. Su capacidad de poder analizar

simultáneamente miles de transcritos ha hecho que se produzcan importantes avances en los puntos antes mencionados[4]. Sin embargo, la propia naturaleza de los microarrays (están basados en la hibridación y puede haber hibridación de una sonda independientemente de su nivel de expresión[4]) limita su capacidad de catalogar y cuantificar moléculas de ARN expresadas en distintas condiciones[3].

Actualmente, aunque estas técnicas siguen usándose, la aparición de las técnicas de secuenciación de última generación (NGS por sus siglas en inglés) ha producido una revolución en el mundo científico debido a la gran cantidad de información disponible que está generando y a la aparición de multitud de herramientas informáticas nuevas para analizar todos estos datos. Una tecnología que ha surgido basada en NGS dentro de la transcriptómica, es el RNA-seq. Algunas de estas plataformas de secuenciación han sido desarrolladas por 454 Life Sciences (Roche) e Illumina (Solexa sequencing)[4]. Esta tecnología nos permite producir cientos de millones de *reads* o lecturas de secuencias que a su vez es información sobre billones de bases nucleotídicas[1] a diferencia de los métodos anteriores que no proporcionan la secuencia de bases[2]. Además, nos permite conocer todas las moléculas de ARN existentes en una célula o tejido en un momento determinado y bajo unas circunstancias específicas, aportando una cobertura mucho mayor y una mejor resolución que las técnicas anteriores[5]. Algunas de las aplicaciones de esta tecnología son: descubrimiento de transcritos, anotación del genoma, estudio de los mecanismos de la regulación génica, análisis de expresión diferencial genética, análisis de expresión alternativa, análisis de expresión específica de alelos, detección de ARN editado, detección viral, detección de fusión de genes y detección de otros tipos de variantes[3,6]. Rápidamente ha reemplazado a los microarrays por su mejor resolución y alta reproducibilidad[3], además de no depender de la anotación del genoma anterior a la selección de sondas y evitar los sesgos introducidos durante la hibridación de los microarrays[7].

Aun así, esta tecnología también tiene algunos inconvenientes: los experimentos con RNA-seq solamente representan una instantánea de la expresión genética de una muestra de células estática y la expresión de ARN no siempre se correlaciona con la expresión proteica[6]. Además, también existen desafíos (muchos relacionados con la purificación, calidad y cantidad de la muestra): uno

de ellos tiene que ver con los tipos de ARN y la cantidad en que se encuentra cada uno de ellos: los más abundantes son los ARN ribosomales, mientras que aquellos que codifican proteínas se encuentran en baja proporción. Por tanto, es importante considerar el enriquecimiento de las librerías para poder extraer la máxima información posible del ARN que nos interesa. En comparación con el ADN, el ARN también resulta extremadamente frágil.

A continuación, se van a exponer los procedimientos técnicos y bioinformáticos de los que se compone este tipo de análisis (figura 1):

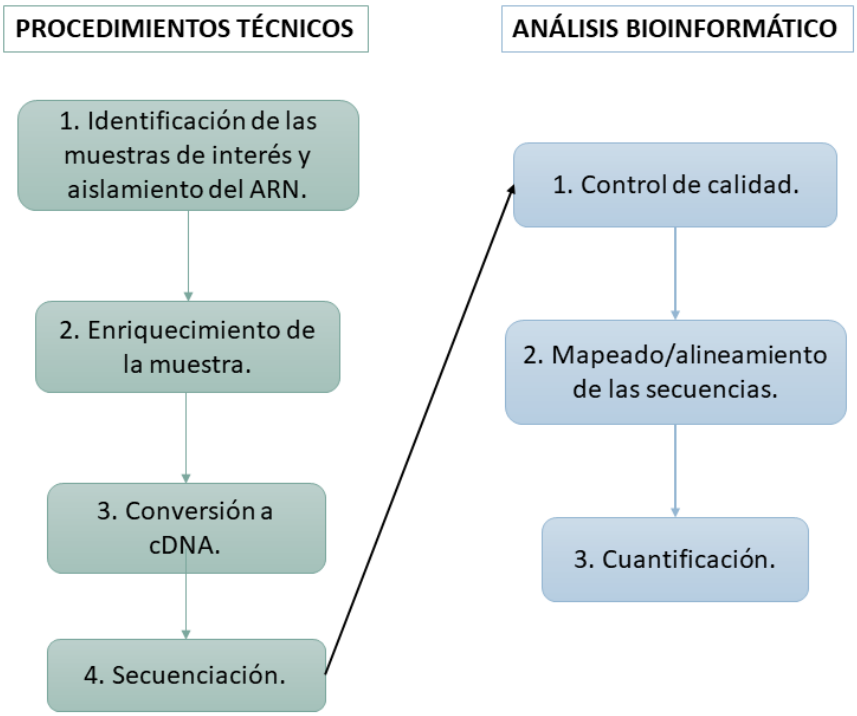


Figura 1: esquema de la metodología de análisis en RNA-seq.

PROCEDIMIENTOS TÉCNICOS

El procedimiento técnico para elaborar las librerías de RNA-seq normalmente utiliza transcripción inversa (conversión de una molécula de ARN a ADN) y es el siguiente[2,5,6]:

Identificación de las muestras biológicas de interés y aislamiento del ARN: Tener claro el experimento a realizar es esencial ya que las muestras deben escogerse entre los cientos de tipos celulares existentes en casi 200 tejidos

distintos (solamente considerando la especie humana). También se debe tener en cuenta el momento de la extracción de ARN y el objetivo del experimento[5]. El ARN de dichas muestras es aislado y debe tener una calidad suficiente para producir una buena librería. La calidad del ARN extraído se mide mediante el número de integridad del ARN (RIN) con valores del 1 (mínima calidad) al 10 (máxima calidad)[5].

Enriquecimiento de la muestra: Es posible seleccionar las moléculas de ARN de interés para mejorar el experimento mediante distintos métodos. En las células, la gran mayoría del ARN que encontramos es ARN ribosómico (ARNr) (95% del total). Si este no es eliminado antes de construir la librería, hará que se reduzca la profundidad de la secuenciación limitando la detección de otros ARNs menos abundantes. Esto se puede llevar a cabo incrementando la detección de ciertos transcritos[5] mediante el enriquecimiento de las colas poly-A (secuencia de varias adeninas que se añaden al final del ARNm maduro para aumentar su estabilidad), reduciendo la cantidad de ARNr, realizando una selección a partir del tamaño o mediante el uso de técnicas de normalización de ADN complementario (ADNc). Es conveniente conocer cómo funciona cada una de estas técnicas para poder aplicarlas al experimento en función de nuestro objetivo[6].

Conversión a ADNc: Posteriormente, el ARN es fragmentado y convertido en ADNc mediante una reverso transcriptasa (enzima que realiza la conversión de ARN a ADNc). Se pueden usar adaptadores y se puede decidir si conservar o no la información sobre a qué cadena pertenece cada secuencia, ya que conocer esto puede ser importante en el caso de querer detectar nuevos transcritos[5].

Secuenciación: Cada molécula (amplificada o no) es secuenciada mediante un secuenciador de última generación. Existen distintos secuenciadores, siendo Illumina el más usado[5]. La elección de una plataforma de secuenciación u otra, la instrumentación, la longitud de los *reads* o el uso de *single-end* (secuenciación de un solo extremo de un fragmento) o *paired-end* (secuenciación de los 2 extremos de un fragmento) también pueden influenciar en cierta medida nuestros resultados, por lo que conviene igualar estos parámetros dentro del mismo experimento[6]. Finalmente, se obtienen pequeñas secuencias, llamadas *reads*. Estas se incluyen en un fichero de tipo FASTQ. Los archivos FASTQ están estructurados de la siguiente manera: cada secuencia consta de 4 líneas. La

primera de ellas es su ID (con información adicional o no) y comienza con el símbolo '@', la segunda es la secuencia de bases nucleotídicas de dicha secuencia, la tercera comienza con el símbolo '+' y a continuación podemos encontrar de nuevo el ID o información adicional y, por último, la cuarta línea es un índice de calidad de casa base. A partir de dichos ficheros se realizan el resto de análisis: calidad, alineado, cuantificación etc.

ANÁLISIS BIOINFORMÁTICO

A partir del paso de secuenciación, los siguientes pasos a realizar siguen métodos bioinformáticos.

Control de calidad: antes de seguir avanzando en el análisis es conveniente realizar un análisis de calidad del archivo FASTQ que contiene las secuencias para asegurarnos de que tienen la calidad suficiente para realizar el resto del análisis con garantía. Este paso puede consistir tanto en la retirada de los adaptadores usados en los pasos previos, como en la comprobación la calidad de las lecturas.

Los adaptadores son pequeños oligonucleótidos unidos a los extremos de las secuencias de interés durante la secuenciación con herramientas NGS. Su retirada de las secuencias (proceso denominado *adapter trimming*) es esencial para poder leer los datos correctamente. Son fáciles de identificar ya que se especifican en el kit con el que se realiza la secuenciación. Para realizar este procedimiento se usará el programa Skewer[8]: nos permite retirar los finales de los *reads* donde la calidad de las bases desciende hasta un nivel donde podemos obtener errores en la secuencia. Este proceso se denomina *quality trimming*. Este último procesado puede ser esencial según el tipo de análisis que se busque[3].

También se puede analizar la calidad de las propias secuencias para identificar cualquier tipo de problema o sesgo que afecte a nuestros datos y, por tanto, eliminarlo o excluir la muestra del análisis. En este trabajo se usará el software FastQC[9] que, entre otras cosas, indica la calidad de la secuencia base a base, el contenido de cada base, el contenido de G y C por base y por secuencia, contenido de Ns o secuencias duplicadas o sobrerrepresentadas [10].

Mapeado o alineamiento de las secuencias: Uno de los mayores retos trabajando con RNA-seq es, precisamente, el mapeado de los *reads*

secuenciados de forma precisa[11]. Como previamente se han fragmentado y secuenciado todas las moléculas, el resultado son fragmentos de una longitud determinada que es necesario mapear para saber a qué zona del genoma (y, por tanto, a qué gen) pertenecen. Al tratar con el transcriptoma celular tenemos que tener en cuenta que los transcritos sufren procesos de *splicing* en los que se eliminan los intrones para dar lugar al ARNm, quedando regiones no contiguas juntas en el genoma. Por tanto, cuando leemos el ADNc (formado a partir del ARNm) no estamos leyendo la secuencia real tal y como se encuentra en el genoma. Es necesario mapear estas secuencias no-adyacentes al genoma de referencia o al transcriptoma. Además, existen otros 2 motivos que hacen que esta tarea necesite un gran requerimiento computacional[11]: el primero de ellos tiene que ver con intentar alinear de forma precisa secuencias que contienen errores, inserciones y deleciones causadas por variaciones genómicas o errores de secuenciación. El segundo está relacionado con mapear *reads* que se expanden por dos o más exones. Además, hay que añadir la presencia de múltiples copias de secuencias genómicas idénticas o parecidas que se transcriben a ellas mismas[12].

El mapeado puede hacerse respecto al genoma o al transcriptoma. Hacerlo a este último impide que se descubran nuevas isoformas de un gen (o no anotadas) y hace que la investigación se enfoque solamente en la cuantificación[13].

Cuantificación: es la aplicación más común de RNA-seq[13]. En este punto se cuenta el número de *reads* alineados a cada transcrito. Lógicamente, aquellos fragmentos pertenecientes a genes que están más expresados, aparecerán más veces. En este paso nos encontramos con una dificultad: hay *reads* que no se mapean únicamente a un solo transcrito o gen[14]. Existen distintos tipos de unidades con las que medir la expresión de cada transcrito o gen y no todos se pueden usar para comparar muestras entre ellas, como ocurre con los *raw counts*, que son las medidas sin procesar. Estas unidades están afectadas por factores como la longitud del transcrito, el número total de *reads* y otros sesgos de los secuenciadores[13]. Para poder comparar medidas de expresión entre muestras se utilizan unidades como RPKM (Reads per kilobase of exon per million), FPKM (Fragments per kilobase of exon per million) o TPM (Transcripts per million). Esta última medida es la que se considera que es la más apropiada

para comparar muestras de distintos orígenes y composición, aunque también puede sufrir algunos sesgos[13]. Además, dado que la técnica RNA-seq permite contrastar las diferencias en expresión genética entre células con distintas condiciones, es necesario realizar una normalización de los datos para realizar la comparación, ya que cada muestra puede tener distintas coberturas de secuenciación (en NGS, el número de veces que cada base nucleotídica es leída) y composición de ARN. Por esta razón, realizar una buena normalización es esencial para comparar los niveles de expresión genética y estimar los parámetros del modelo[15].

Por tanto, debido a la gran acogida y expansión de esta técnica, en los últimos años se ha creado multitud de software destinado a la ejecución de distintos análisis. Precisamente, este es el punto de inicio de este trabajo. El grupo de investigación trabaja con datos RNA-seq procedentes de tumores y frecuentemente realiza análisis de cuantificación. El software de referencia que el grupo viene utilizando es STAR[12] y RSEM[14], pero debido a la existencia de programas que realizan también el paso de cuantificación, cabe realizar una comparación entre estos programas y otros existentes. Los elegidos para la comparación son Kallisto[16], Salmon[17] y Sailfish[18] (pertenecientes a los denominados *"ultra-fast aligners"*[19]). Las características de cada programa se comentan a continuación:

STAR (Spliced Transcripts Alignment to a Reference) alinea secuencias no contiguas al genoma de referencia directamente, trabajando en dos pasos (*seed searching step y clustering/stitching/scoring step*), y constituye más tarde la base para ejecutar otro software. La novedad que este algoritmo introdujo es la capacidad de alinear secuencias de ARN más largas que se originan a partir de la NGS, siendo capaz de alinear tanto secuencias cortas como largas[12].

RSEM, Kallisto, Salmon y Sailfish nos permiten estimar los niveles de expresión de genes a partir de datos de RNA-seq. La principal diferencia entre RSEM y Kallisto-Salmon-Sailfish es que el primero necesita un paso de mapeado previo realizado mediante STAR, mientras que los otros 3 programas no necesitan ejecutar este paso ya que están basados en la idea de que no se necesitan alineaciones precisas para poder generar las estimaciones de la expresión de cada transcrito o gen[19].

Los métodos que siguen cada uno de ellos son los siguientes:

- **RSEM** (RNASeq by Expectation Maximization): Está diseñado para trabajar con *reads* alineados directamente a los transcritos en vez de utilizar todo el genoma. Así se obtienen varias ventajas como evitar los saltos generados por el *splicing* (proceso post-transcripcional mediante el cual se eliminan ciertos fragmentos de del ARN durante su maduración), facilitar el alineamiento en especies sin un genoma secuenciado pero con un transcriptoma caracterizado y permitir un alineamiento más rápido ya que el tamaño del transcriptoma es más pequeño que el del genoma completo[14]. Este programa opera en 2 pasos[3]: el primero consiste en la generación y preprocesado de un grupo de transcritos y el segundo en el alineamiento de las lecturas a los transcritos de referencia seguido de una estimación de la abundancia de cada transcrito. Los resultados de la cuantificación se obtienen también a nivel génico. Genera las estimaciones de la expresión de genes y transcritos con el método Maximum-Likelihood usando un algoritmo EM (Expectation-Maximization)[3].
- **Kallisto**: este software realiza un pseudoalineamiento entre cada *read* y los transcritos a partir del índice previamente generado. Trata de identificar el transcrito a partir del cual se ha originado cada *read* pero no trata de señalar cómo se alinea cada uno con los transcritos (es decir, en qué parte de los transcritos está alineado cada *read*), ya que esta última parte no es necesaria para asegurar una buena precisión en las medidas de expresión obtenidas. Esto se puede realizar mediante un *hashing* (generar un valor numérico para cada elemento de una cadena de caracteres para acelerar el procesado de datos) rápido de k-mers (todas las sub-secuencias de longitud *k* dentro de una secuencia). Este software utiliza un algoritmo EM y las estimaciones de la expresión de cada transcrito se obtienen mediante un método de Maximum-Likelihood[16].
- **Salmon**: está diseñado para aprovechar las ventajas de trabajar con varios procesadores y su proceso de alineamiento cuenta con 3 componentes: un mapeado ligero, una fase *online* (donde se estima la expresión inicial y los parámetros auxiliares mediante una variante de una inferencia Bayesiana o un método Maximum-Likelihood) y una fase *offline* (donde se refinan las estimaciones de la expresión con un algoritmo EM o un EM bayesiano). En conjunto, las 2 últimas fases crean un modelo probabilístico óptimo de la

abundancia de cada transcrito ya que el programa incorpora información que no tienen en cuenta Kallisto o Sailfish. El tipo de mapeado que usa se denomina *quasi-mapping*. Durante el mapeado, se rastrea la posición y orientación de los fragmentos mapeados, y esta información se usa junto a las estimaciones de abundancia obtenidas en la fase online para computar las probabilidades condicionales. Estas probabilidades se usan para estimar parámetros del modelo y de sesgo que, a su vez, actualizan las estimaciones de abundancia. En este modo, el archivo de entrada son el transcriptoma y un grupo de reads, realizándose la cuantificación de estos *reads* directamente, sin generar ningún archivo intermediario, ahorrando tiempo y espacio[17].

- **Sailfish**: este software tampoco necesita un alineamiento previo para realizar la cuantificación. Al evitar el alineamiento, también se reduce el tiempo necesario para realizar cualquier análisis. Las primeras versiones del programa construían su índice desde un set de tránscritos de referencia y de un valor de k-mer específico, empleando en el paso de cuantificación un algoritmo EM, parecido al de RSEM [3] denominado SQUAREM [18]. Desde la versión 0.7.0 en adelante, se sustituyó el alineamiento mediante k-mer por un *quasi-mapping* igual que el que usa Salmon. Una vez más, está basado en la idea de que no es necesaria una gran precisión en el mapeado para obtener resultados de cuantificación precisos. Se utiliza un algoritmo VBEM (*variational Bayesian expectation maximization*) para realizar las estimaciones de expresión [20].

La principal diferencia entre RSEM y el resto de programas reside en la fase de mapeado o alineamiento. STAR realiza un mapeado en el que se genera un archivo BAM (Binary Alignment/Map: versión comprimida del archivo SAM con al menos 11 campos de información sobre el alineado), donde aparte de estar cada *read* alineado base por base hay mucha más información que puede ser usada en otro tipo de análisis. Kallisto, Sailfish y Salmon solamente indican a qué parte del genoma (transcrito o gen) pertenece cada *read* pero sin alinear base por base ni generar archivos intermedios. Es decir, mientras que con la información generada por STAR se podrían realizar otros tipos de análisis, con la información producida al construir los índices de Kallisto, Sailfish y Salmon solamente se puede realizar una cuantificación de la expresión genética. Además, al generar cada uno su índice a partir del transcriptoma, tampoco pueden usarse para la

identificación de nuevos transcritos. La ventaja de Kallisto, Salmon y Sailfish reside en la disminución del tiempo empleado en analizar las muestras al no realizar un mapeado tan preciso.

Programa	Mapeado	Algoritmo cuantificación
RSEM	Mediante STAR	ML/EM
Kallisto	Pseudoalineamiento	ML/EM
Sailfish	Quasi-mapping	ML/VBEM
Salmon	Quasi-mapping	ML/VM o EM bayesiano

Tabla 1: tabla resumen de las características de cada programa. ML: Maximum-Likelihood. EM: Expectation-Maximization.

2.2 Materiales y métodos

Los datos que se han usado para llevar a cabo este trabajo corresponden a muestras *paired-end* que provienen de un estudio en el que se analizaron los patrones de expresión en sangre periférica para estudiar los efectos sistémicos de la hemorragia aneurismal subaracnoidea. Se han escogido solamente las muestras que provienen de controles sanos para evitar cualquier desviación o sesgo producidos por una sobre o infraexpresión de algunos genes asociada a cualquier enfermedad. Los códigos del estudio son los siguientes: SRA: SRP150595 y BioProject: PRJNA473142[21]. Las muestras seleccionadas han sido: SRR7346345, SRR7346342, SRR7346340, SRR7346337, SRR7346336, SRR7346334, SRR7346331, SRR7346329, SRR7346323, SRR7346320, SRR7346316, SRR7346313, SRR7346312, SRR7346310, SRR7346306, SRR7346302 y SRR7346300. En la figura 2 se puede ver un esquema de la metodología del análisis.

Control de calidad: Como primer paso se ha realizado un análisis de calidad de las muestras. Aunque ya se preveía que esta era buena debido a que proceden de estudios previos, se ha comprobado que así era y también se han eliminado algunos de los adaptadores que aún quedaban. El procedimiento ha sido el siguiente: mediante el software Trim-galore (versión 0.4.5)[22] se ha analizado la

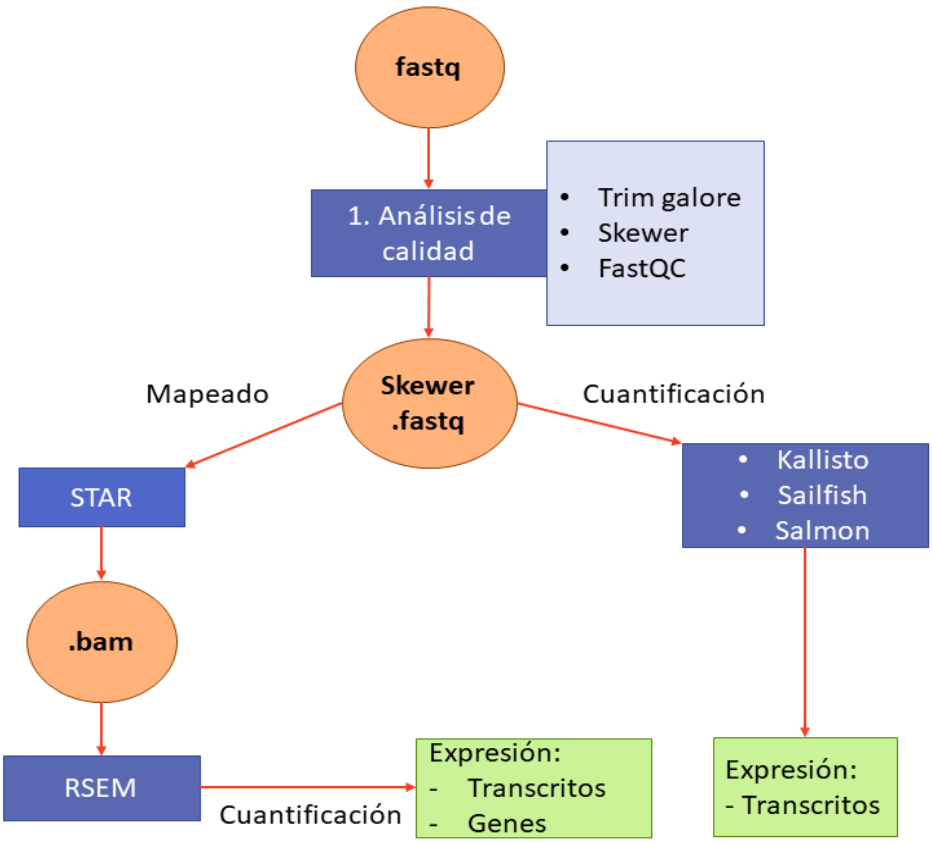


Figura 2: esquema de la metodología seguida en el proyecto. En azul se muestran los programas utilizados. En naranja aparecen los archivos que utiliza cada programa y en verde los resultados obtenidos.

muestra SRR7346345 para obtener el adaptador usado en el proceso de secuenciación. Solamente se ha realizado este proceso una vez, ya que, al pertenecer todas las muestras al mismo estudio, todas han sido tratadas de la misma manera y, por tanto, su adaptador era "AGATCGGAAGAGC". Una vez aparece en pantalla el adaptador, se detiene el análisis y se usa este adaptador en el programa Skewer (versión 0.2.2) que es capaz de detectarlo en las secuencias y eliminar aquellas que lo portan, generando un nuevo archivo con las secuencias procesadas para cada muestra. Este archivo es el que se usará en análisis posteriores. Finalmente, las muestras son analizadas mediante FastQC (versión 0.11.7) con el siguiente criterio: si superan el análisis de calidad de las secuencias *per base*, se considera que son de buena calidad. Este análisis no se supera normalmente por una degradación general de la calidad[10].

STAR y RSEM necesitan construir un índice del genoma de la especie de la que procede la muestra antes de realizar cualquier otro paso. En este caso se ha

utilizado la versión del genoma humano GRCh38 para STAR (versión 2.6.1) y RSEM (versión 1.3.1) junto con el fichero de anotaciones *primary assembly* de GENCODE v28. STAR se utiliza para mapear las muestras produciendo un archivo .bam que utiliza RSEM en la cuantificación. Kallisto (versión 0.43.1), Salmon (versión 0.9.2) y Sailfish (versión 0.11.3) construyen su índice a partir del transcriptoma de la especie. Para generar el archivo fasta con el que construir el índice se ha utilizado el propio genoma y el fichero de anotaciones con herramienta gtf_to_fasta de TopHat (versión 2.1.1)[23].

En la ejecución de todos los análisis se han usado 4 procesadores. De esta manera se puede realizar una comparación directa del tiempo empleado por cada programa sin necesidad de dividir entre el número de procesadores primero.

Los parámetros que se han usado en los programas STAR, RSEM, Kallisto, Sailfish y Salmon se encuentran en el anexo 1.

También se ha registrado el tiempo (en segundos) que emplea cada programa en ejecutar su análisis a partir de los registrados en los archivos de salida por los propios programas, excepto para RSEM que no ofrecía dicho registro y hubo que registrar el tiempo de creación de los archivos en el ordenador. A la hora de analizar estos datos, se ha utilizado para RSEM el tiempo empleado por dicho programa en realizar la cuantificación de cada muestra sumado al tiempo empleado por STAR en realizar el mapeado de cada muestra. Para Kallisto, Sailfish y Salmon se ha utilizado el tiempo empleado en el paso de cuantificación.

La comparación entre los resultados de los programas se realizará por transcritos y por gen. Aunque los resultados de Kallisto, Sailfish y Salmon solamente se obtienen por transcrito, se ha asociado cada transcrito a su gen de igual forma que se ha efectuado en RSEM: este software calcula la expresión de un gen sumando la expresión de todos sus tránscritos. Por tanto, para Kallisto, Sailfish y Salmon se han sumado las expresiones de todos los transcritos que pertenecen a un mismo gen. En la realización de los análisis posteriores se han utilizado los valores de expresión en TPM y se han normalizado aplicando logaritmo de la expresión en TPM + 1. Los análisis ejecutados (se ha usado R studio[24] versión 1.1.423 con R[25] versión 3.5.1) han sido los siguientes (se han repetido a nivel de transcrito y a nivel de gen):

- Tiempo empleado por cada programa en realizar la cuantificación. Esta comparación es importante ya que el uso de un programa más rápido daría lugar a una reducción de los tiempos computacionales y a una aceleración de la investigación.
- Porcentaje de *reads* mapeados en cada muestra por cada programa. Este porcentaje indica el nº de *reads* asociados al genoma/transcriptoma respecto a la muestra total. Un mayor porcentaje indica una mayor calidad del resultado del análisis y de las muestras.
- Número de genes y transcritos distintos reconocidos por cada programa. Para su detección se han contabilizado aquellos cuyo valor de expresión (TPM) es mayor de 0.1 [26].
- Correlación de Spearman entre los datos de expresión obtenidos para cada posible pareja de programas. Se ha elegido el coeficiente de Spearman ya que los datos no siguen una distribución normal por lo que hay que usar pruebas no paramétricas. También se han representado gráficamente las nubes de puntos de los valores de expresión registrados por cada programa. Este análisis permite observar cómo se relacionan las medidas de expresión tomadas por los distintos programas entre sí.
- Correlación de Spearman entre la desviación estándar de cada transcrito/gen de RSEM con Kallisto, Sailfish y Salmon. También se ha representado gráficamente la distribución de los valores de la desviación estándar de cada software, tanto la nube de puntos por parejas de comparación, como en forma de boxplot. Nos permitirá observar cómo varía la expresión medida entre cada programa por sí mismo y entre los demás.
- Tests estadísticos para comprobar la igualdad de varianzas: se ha realizado el test de la U de Mann-Whitney (tenemos datos que no siguen una distribución normal por lo que se realizan pruebas no paramétricas) para comprobar la igualdad de la media de la desviación estándar entre RSEM y Kallisto, Sailfish y Salmon, y también el test de Kolmogorov-Smirnov (prueba no paramétrica para comprobar la igualdad de distribuciones) para las mismas parejas.

- Gráficos de la relación entre la expresión y la desviación estándar de cada observación para cada programa. Observar esta distribución permite comprobar si es la esperada para los datos de expresión genética.
- Cálculo del coeficiente de variación (CV) de cada observación en cada programa. Esta medida se calcula dividiendo la desviación estándar de cada observación entre su media. Se ha representado gráficamente la relación entre los distintos programas.

También se ha reducido el tamaño de los archivos de las muestras en un 40%, 60% y 80% mediante seqtk (versión 1.3-r106)[27], utilizando siempre la semilla 123, y se han realizado los mismos análisis de cuantificación con RSEM, Kallisto, Sailfish y Salmon. El objetivo de este análisis es comprobar el rendimiento de estos programas cuando el tamaño de los archivos fastq va disminuyendo. Aunque lo ideal habría sido realizar este análisis reduciendo en intervalos del 10% cada archivo, no se disponía del tiempo necesario para generar 9 reducciones de los 34 archivos totales (2 por cada muestra (17)). Por tanto, se ha procedido a realizar este análisis con solamente 3 reducciones.

2.3 Resultados

- **Tiempo de ejecución de cada programa**

La distribución del tiempo de ejecución obtenido de cada programa se puede observar el boxplot de la figura 3.

La distribución de los tiempos de ejecución muestra que RSEM es el software que más tiempo emplea en proporcionar resultados de cuantificación. Queda muy por encima de Sailfish (el segundo más lento) que presenta unos tiempos más parecidos a Salmon y Kallisto (el más rápido). Es el único programa cuya distribución está por encima de la línea en y = 3600, es decir, su tiempo de ejecución está siempre por encima de una hora para cada muestra. RSEM presenta una mediana por encima de los 5000 segundos (83 minutos y 20 segundos) mientras que Sailfish no llega a 2000 segundos (33 minutos y 20 segundos). La mediana de Salmon y Kallisto ni siquiera llega a 1000 segundos (16 minutos y 40 segundos). Kallisto es también el programa que más estable resulta a la hora de realizar la cuantificación: la distribución de los tiempos medidos es la más pequeña de los 4 programas.

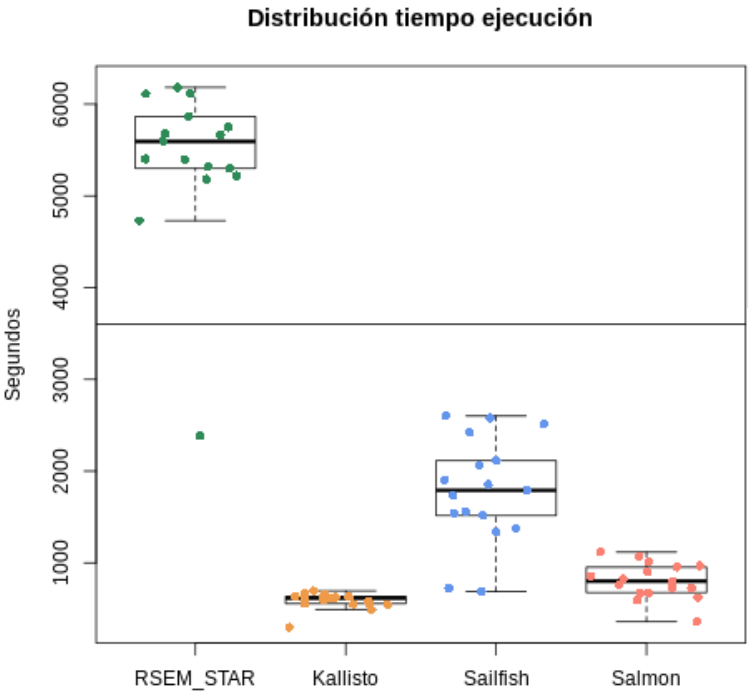


Figura 3: distribución del tiempo de ejecución de la fase de cuantificación empleado en cada muestra por cada programa. En el caso de RSEM se han incluido también el tiempo empleado por el programa STAR en el mapeado de las muestras. Se ha añadido una línea en y = 3600, para representar el tiempo correspondiente a una hora.

- **Porcentaje de *reads* mapeados.**

Un dato importante a tener en cuenta en los análisis de RNA-seq es el porcentaje de *reads* que se mapean en cada programa ya que puede ser un indicativo de la calidad de las propias muestras y del análisis en general. En el boxplot de la figura 4 se representa la distribución del porcentaje obtenido en cada muestra con los programas STAR, Sailfish y Salmon. La información proporcionada por STAR es la que usa RSEM posteriormente. Kallisto no proporciona dicha información.

En el boxplot podemos apreciar una clara diferencia entre los porcentajes obtenidos por STAR y los obtenidos por Sailfish y Salmon: en el primero todas las muestras están por encima del 90%, mientras que en los otros 2 programas solamente una muestra está por encima del 60% e incluso hay varias por debajo del 50%. Los porcentajes detectados por Sailfish y Salmon son prácticamente idénticos: utilizan ambos un *quasi-mappping.*

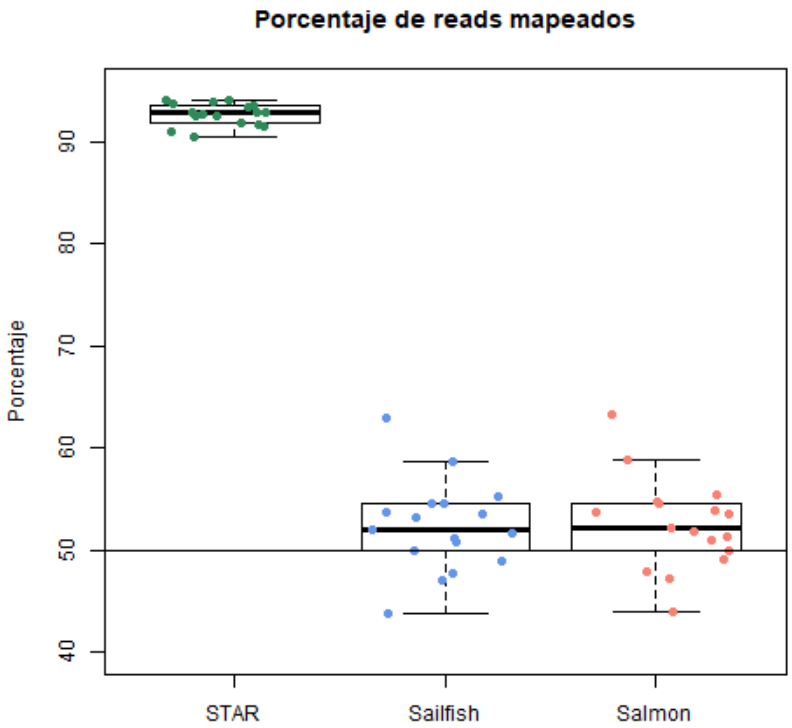


Figura 4: distribución del porcentaje de "reads" mapeados por cada programa.

Número de tránscritos y genes detectados por cada programa.

A partir de estos datos, podemos observar que RSEM es el programa que menos tránscritos y genes distintos detecta. Fijándonos en los porcentajes (tabla 2), se puede apreciar que todos los programas detectan una mayor proporción de genes que de transcritos distintos. Kallisto, Sailfish y Salmon están por encima de RSEM en todos los casos con porcentajes muy parecidos entre ellos. En la figura 5 se puede observar la representación gráfica de estos porcentajes.

	RSEM	Kallisto	Sailfish	Salmon
Tránscritos	16.328	35.051	34.054	33.838
Gen	24.961	44.062	44.242	43.936

Tabla 2: tabla con el porcentaje de transcritos y genes distintos reconocidos por cada programa.

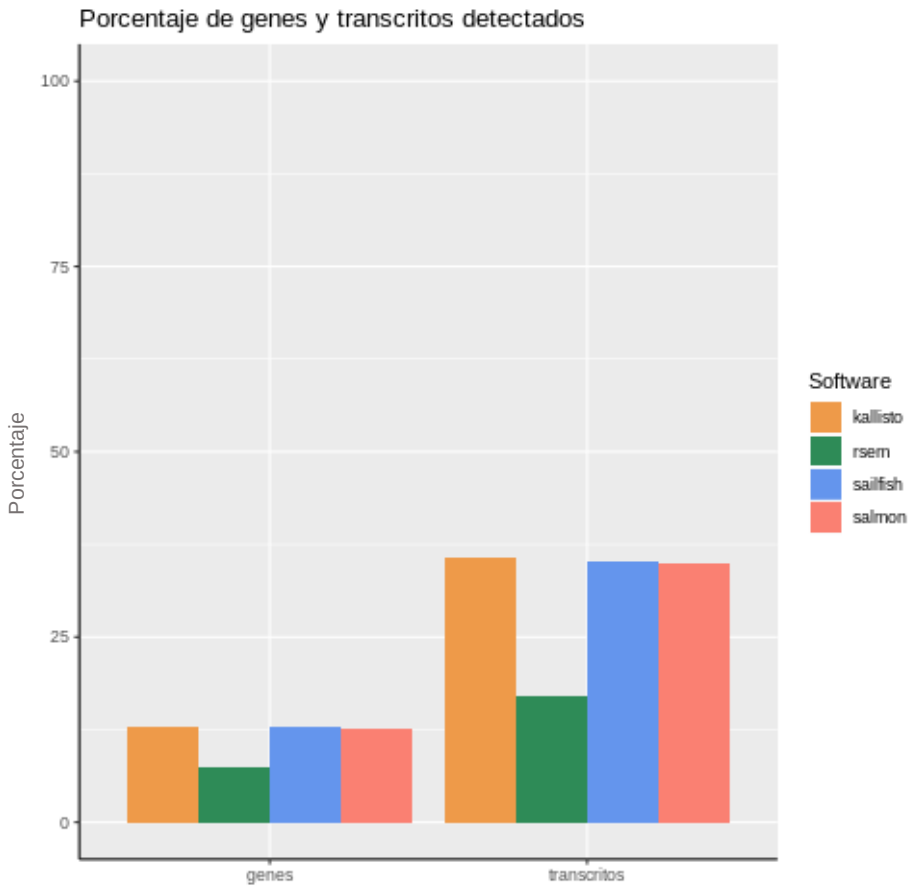


Figura 5: diagrama de barras que representa el porcentaje de genes y transcritos distintos reconocidos por cada programa.

- **Transcritos**

Correlación entre los datos de expresión obtenidos con RSEM, Kallisto, Sailfish y Salmon.

Para comprobar si los datos de expresión obtenidos por los 4 programas son parecidos entre ellos se ha calculado el coeficiente de Spearman entre todas las parejas de programas posibles y se han representado los valores de expresión obtenidos en cada pareja (figura 6 y anexo 1 (figuras)).

Los resultados del coeficiente de correlación de Spearman muestran una relación alta entre los valores de expresión obtenidos por RSEM y Kallisto, Sailfish y Salmon. Se ha obtenido un valor para RSEM y Kallisto, y RSEM y Salmon de 0,774. El mayor valor se observa entre RSEM y Sailfish con un coeficiente de 0,781. Todos los p-valores son menores de 2,2 e-16, por lo que las correlaciones son significativas.

Los coeficientes de correlación de las parejas de programas Kallisto, Sailfish y Salmon son superiores a los comentados anteriormente. Para Kallisto-Sailfish se ha obtenido un coeficiente de 0.994, 0.993 para Kallisto-Salmon y 0.994 para Sailfish-Salmon. Estos valores nos indican una relación casi perfecta y significativas, con p-valores < 2.2 e-16.

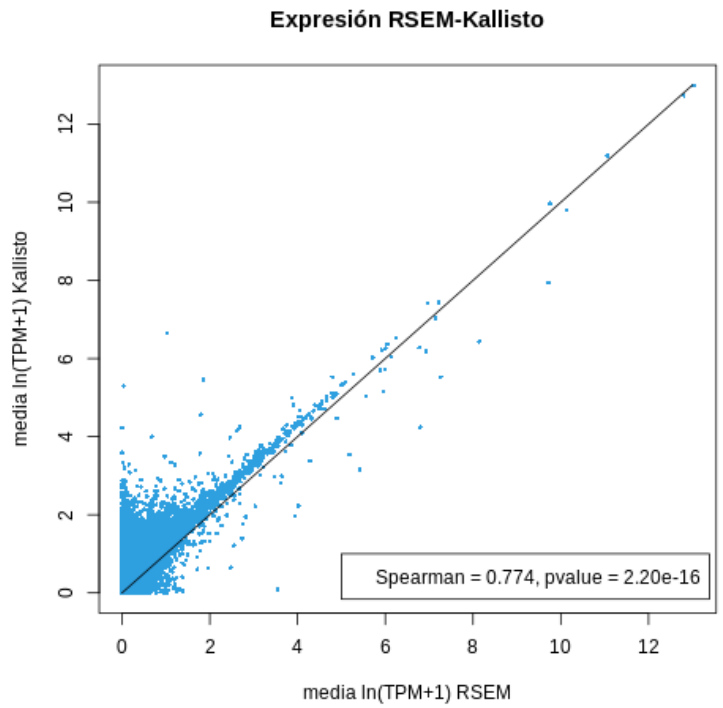


Figura 6: dispersión de puntos de la media del logaritmo del valor de la expresión de cada transcrito más uno entre RSEM y Kallisto. La diagonal observada se ha añadido para una mejor visualización y análisis de los datos y no se corresponde con la línea de regresión entre ambas variables.

Correlación entre la desviación estándar de los datos de expresión de RSEM, Kallisto, Sailfish y Salmon.

El análisis de la desviación estándar nos muestra lo siguiente: en la figura 7 se pueden observar la relación entre la desviación estándar de cada transcrito entre RSEM y Kallisto. El resto de comparaciones se encuentran en el anexo 2 (figuras). En la comparación entre la desviación estándar de las medidas de RSEM frente a las medidas de Kallisto, Sailfish y Salmon nos encontramos con una situación parecida a la anterior: tenemos un valor de correlación relativamente altos (los 3 son superiores a 0,650 con p-valores significativos). Además, también se observa que las observaciones tomadas por RSEM tienen

una menor desviación estándar que las del resto de programas en el boxplot de la figura 8. Podemos ver que la mediana del logaritmo de la desviación estándar de RSEM no llega a 0,01, mientras que para el resto de programas la mediana está por encima de 0,025.

Comparando entre el resto de programas, los coeficientes de correlación vuelven a ser más altos que al comparar con RSEM: para todas las parejas entre Kallisto, Salmon y Sailfish se encuentra por encima de 0.98. Es decir, las medidas tomadas por estos programas tienen una variación parecida.

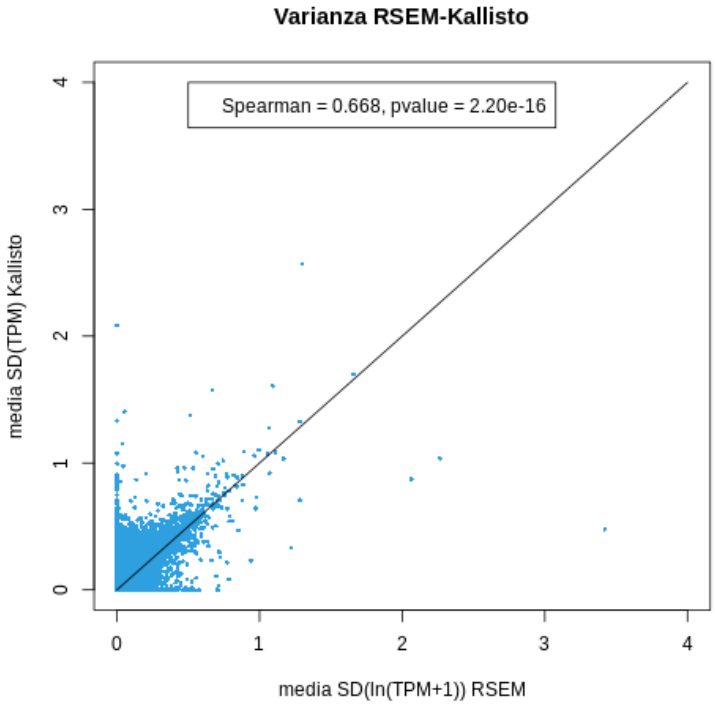


Figura 7: dispersión de puntos de la desviación estándar del logaritmo del valor de la expresión de cada transcrito más uno entre RSEM y Kallisto. La diagonal observada en cada gráfico se ha añadido para una mejor visualización y análisis de los datos y no se corresponde con la línea de regresión entre ambas variables.

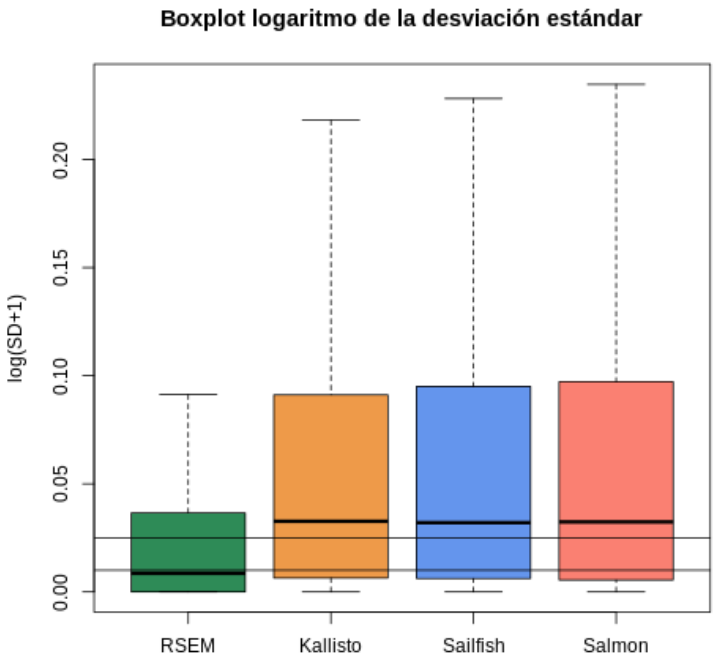


Figura 8: En este boxplot están representados el logaritmo de la desviación estándar + 1 de los 4 programas. Se han añadido 2 líneas de referencia en y = 0,01 e y = 0,025 para una mejor interpretación de los datos

Observando los p-valores obtenidos tanto en los test de la U de Mann-Whitney y Kolmogorov Smirnov (tabla 1 del anexo) observamos que las medias de la varianza son distintas para todas las comparaciones, al igual que la distribución de la varianza, con todos los p-valores por debajo de 0.05.

Relación entre la expresión y la desviación estándar en cada programa.

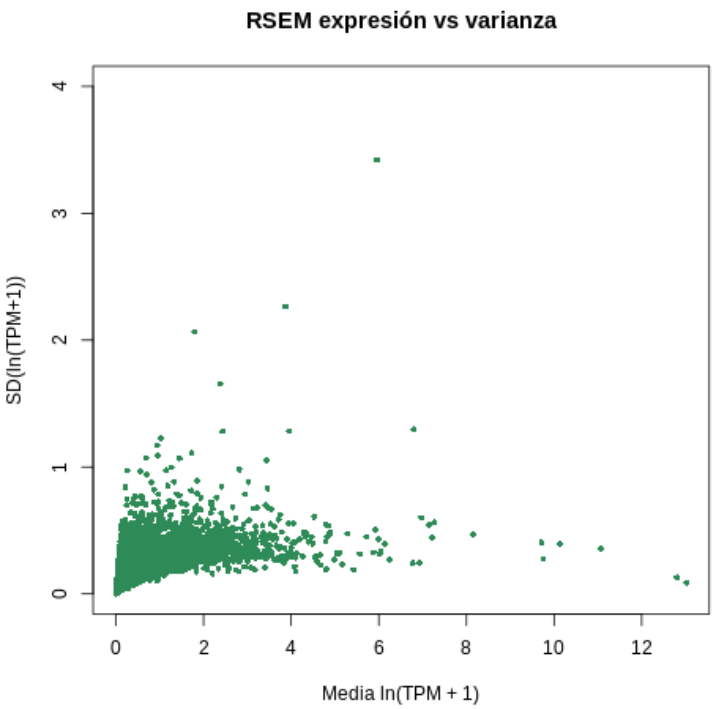


Figura 9: gráfico de la distribución de la varianza y la expresión de cada transcrito en RSEM.

En la gráfica de la figura 9 y en el anexo 3 (figuras) se puede observar la distribución de la desviación estándar y el logaritmo de la expresión de cada transcrito. En general se aprecia que conforme aumenta la expresión de los transcritos aumenta también su desviación estándar.

Coeficientes de variación.

Se ha calculado el coeficiente de variación de cada transcrito dividiendo su desviación estándar entre su media. Se ha comprobado si los transcritos obtienen un mismo coeficiente de variación con todos los programas calculando el coeficiente de correlación de Spearman entre todas las parejas de programas y se ha representado gráficamente.

Los coeficientes de correlación obtenidos son todos significativos (p-valor < 2.2 e-16) y se encuentran entre 0.747 y 0.769, indicando una buena correlación entre RSEM y el resto de programas. En las comparaciones entre Kallisto, Sailfish y Salmon, los coeficientes se encuentran entre 0.984 y 0.990, siendo muy cercanos a uno. Una vez más, se puede observar que hay mayor similitud entre los programas que usan un procedimiento parecido. En la figura 10 se puede observar la nube de puntos entre RSEM y Kallisto. En el anexo 4 (figuras) se encuentran el resto de gráficos.

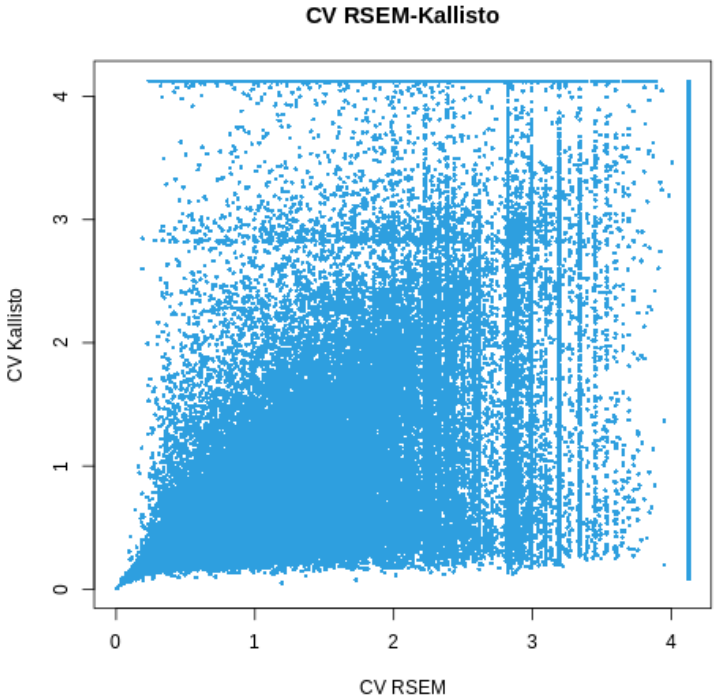


Figura 10: relación de los CV obtenidos para cada transcrito entre RSEM y Kallisto.

- **Genes**

Correlación entre los datos de expresión obtenidos con RSEM, Kallisto, Sailfish y Salmon.

Al igual que en el caso de los transcritos, también se ha querido observar la relación entre el valor de expresión medido para cada gen por todos los programas, utilizando todas las parejas posibles. En la figura 11 se puede observar el gráfico resultante entre RSEM y Kallisto. En el anexo 5 (figuras) se pueden observar el resto de comparaciones.

Volviendo a la división anterior entre RSEM y el resto de programas, vemos que para todas las parejas donde RSEM está presente, han aumentado los coeficientes de correlación hasta estar todos por encima de 0,870, con p-valores < 2,2E-16. Es decir, tenemos correlaciones muy altas y significativas trabajando con genes.

En el caso de las comparaciones donde RSEM no está presente se siguen obteniendo valores muy cercanos a 1: (0.993 para Kallisto-Sailfish, 0.994 para Kallisto-Salmon y 0.994 para Sailfish-Salmon).

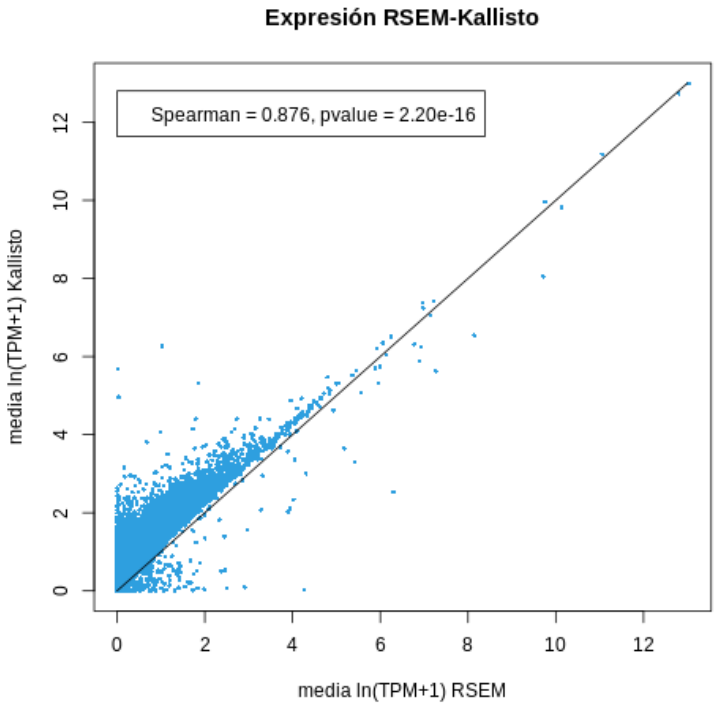


Figura 11: dispersión de puntos de la media del logaritmo del valor de la expresión de cada gen más uno entre RSEM y Kallisto. La diagonal observada en cada gráfico se ha añadido para una mejor visualización y análisis de los datos y no corresponde a la correlación entre ellos.

Correlación entre la desviación estándar de los datos de expresión de RSEM con Kallisto, Sailfish y Salmon.

En la comparación entre la desviación típica de la medida de expresión de cada gen por RSEM frente a las medidas de Kallisto, Sailfish y Salmon se observa lo siguiente: el coeficiente de correlación más alto se observa entre RSEM y Kallisto (figura 12) con un valor de 0,609 mientras que los demás coeficientes se quedan por debajo de 0,590. Los 3 son significativos. Como ocurre en el caso anterior, RSEM también presenta una menor varianza que Kallisto, Sailfish y Salmon. Se puede apreciar en la representación de la distribución de cada programa (figura 13) que la distribución de RSEM queda por debajo de la del resto de programas. La mediana del logaritmo de la desviación estándar está en torno a 0,05 en el

caso de Kallisto, Sailfish y Salmon, mientras que en el caso de RSEM apenas sobrepasa el 0.

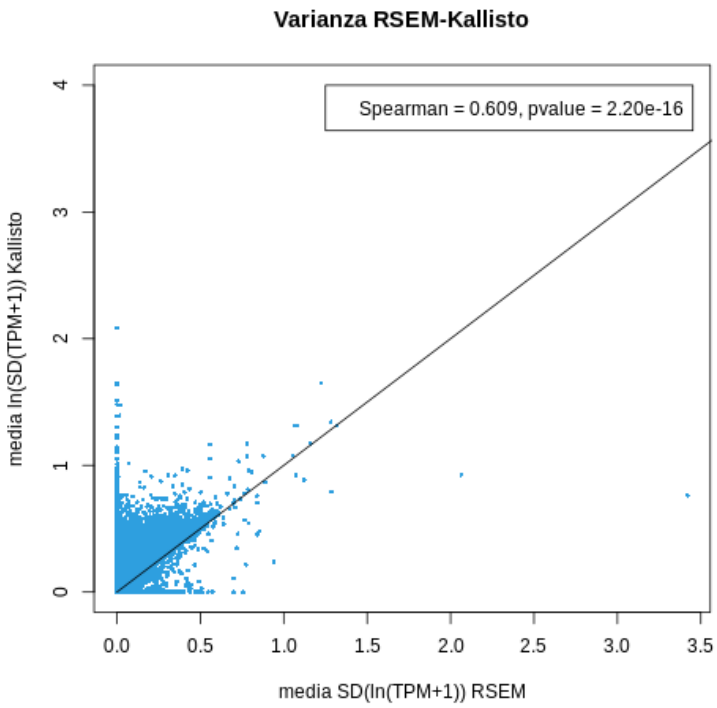


Figura 12: dispersión de puntos de la desviación estándar del logaritmo del valor de la expresión de cada gen más uno entre RSEM y Kallisto. La diagonal observada en cada gráfico se ha añadido para una mejor visualización y análisis de los datos y no corresponde a la correlación entre los datos.

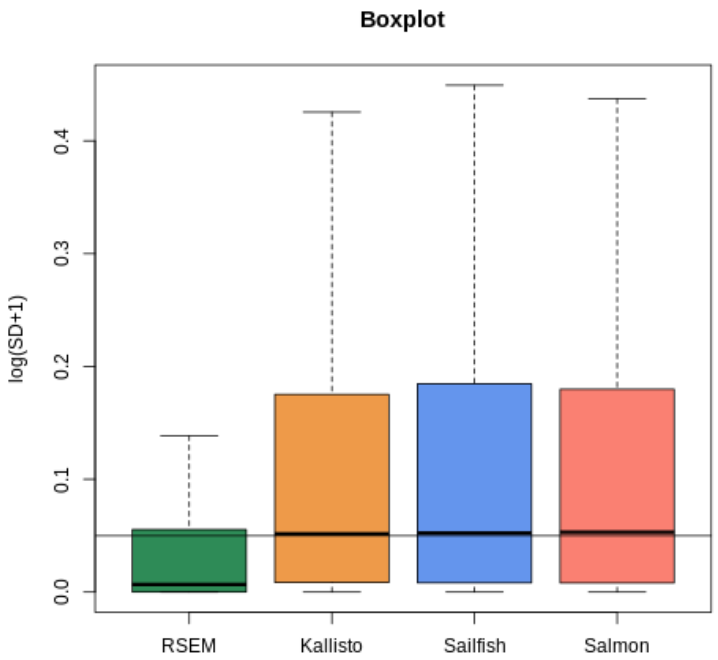


Figura 13: boxplot del logaritmo de la desviación estándar + 1 de los 4 programas. Se ha añadido una línea de referencia en y = 0.05 para la interpretación de los datos.

Los coeficientes de correlación obtenidos entre los programas que no necesitan un mapeado previo vuelven a ser más altos: los 3 se sitúan por encima de 0.99. Observando los p-valores obtenidos en los test estadísticos (tabla 2 del anexo) observamos que la media de la varianza es distinta en todas las comparaciones (test de la U de Mann – Whitney) al igual que su distribución (test de Kolmogorov-Smirnov). El resto de gráficas se encuentran en el anexo 6 (figuras).

Relación entre la expresión y la varianza de cada programa.

En la figura 14 se puede observar la distribución en RSEM de la desviación estándar y la expresión genética. La distribución en los demás programas se puede observar en el anexo 7 (figuras). Lo que nos indican estos gráficos es que los genes con una menor expresión presentan también una menor desviación

estándar, y que conforme aumenta la expresión aumenta el rango de desviación estándar presentado como ocurría en el caso anterior.

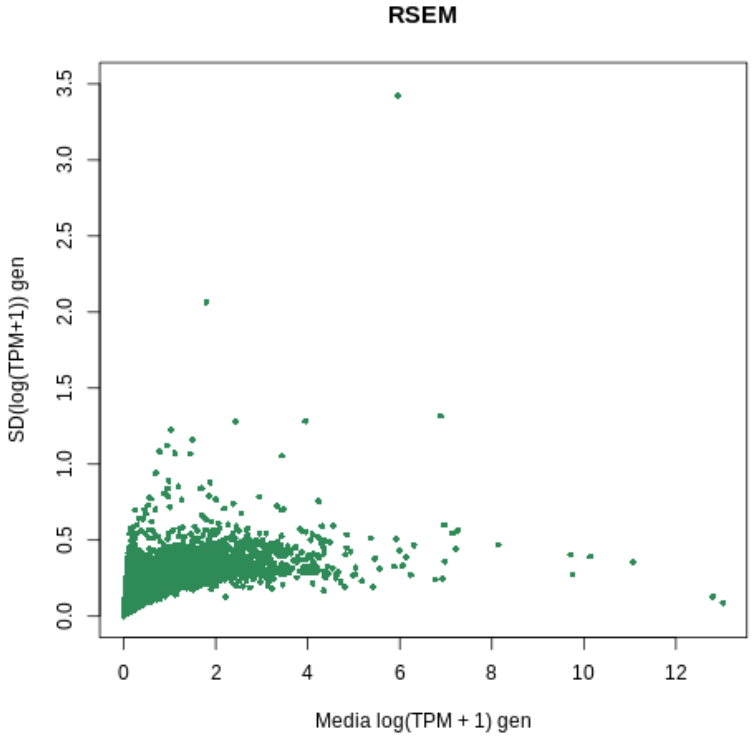


Figura 14: distribución de la desviación estándar y la expresión genética en RSEM.

Coeficiente de variación.

Se ha calculado el coeficiente de variación de cada gen de la misma manera que anteriormente y se ha calculado el coeficiente de correlación de Spearman entre todas las parejas de programas.

Los coeficientes de correlación obtenidos son todos significativos (p-valor < 2.2 e-16) y, en el caso de la comparación de RSEM con el resto de programas, aumenta hasta alrededor de un 0.887 con respecto a los transcritos (0.887 para RSEM-Kallisto, 0.888 para RSEM-Sailfish y 0.886 para RSEM-Salmon). En las comparaciones entre Kallisto, Sailfish y Salmon, los coeficientes son 0.992 para Kallisto-Sailfish, 0.994 para Kallisto-Salmon y 0.991 para Sailfish-Salmon, volviendo a ser muy cercanos a uno. Una vez más, se puede observar que hay mayor similitud entre los programas que usan un procedimiento parecido. En la figura 15 se puede observar la nube de puntos entre RSEM. En el anexo 8 (figuras) se encuentran el resto de gráficos.

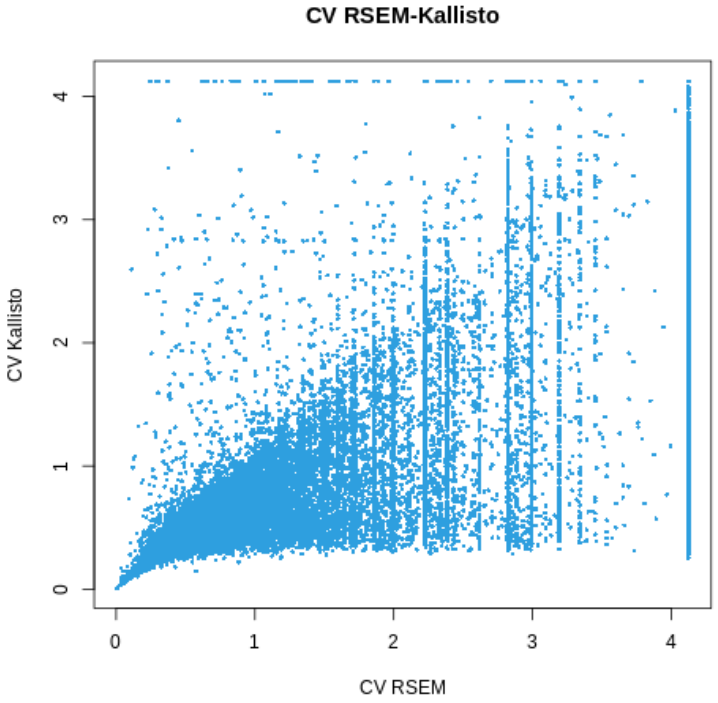


Figura 15: relación de los CV obtenidos para cada transcrito entre RSEM y Kallisto.

ANÁLISIS REDUCIENDO LA CANTIDAD DE *READS*

- **Tiempo de ejecución.**

En el boxplot (figura 16) se puede apreciar que cuando aumenta el tamaño de los archivos .fastq que estamos analizando, también aumenta el tiempo que emplea cada programa en ejecutar el paso de cuantificación. Dentro del conjunto global, la combinación de STAR + RSEM sigue consumiendo más tiempo que el resto de programas y las diferencias entre los programas se hacen más grandes conforme aumenta el tamaño de los archivos. Salmon es el programa que menos varía su tiempo de ejecución conforme aumenta el tamaño de los archivos, mientras que la combinación de STAR + RSEM es la más afectada por el tamaño. Examinando los datos por tamaños, vemos que Salmon es siempre el programa más rápido en todos los casos.

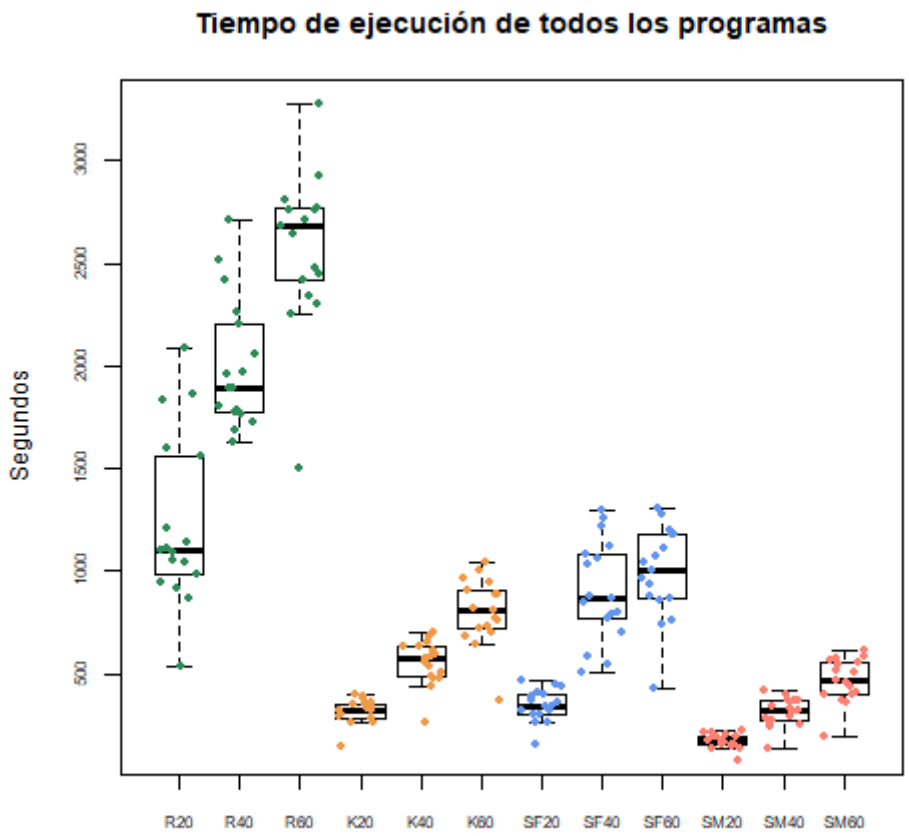


Figura 16: boxplot con la distribución de los tiempos de ejecución de cada programa con los tamaños distintos de los archivos. Los nombres son los siguientes: R – RSEM, K – Kallisto, SF – Sailfish, SM – Salmon. 20 corresponde con un archivo que solo tienen el 20% del tamaño original, 40 con un porcentaje del 40% y 60 del 60%.

- **Porcentaje de transcritos detectados.**

El porcentaje de transcritos detectados disminuye usando archivos más pequeños. Esta disminución es lógica, al igual que la disminución en el tiempo de ejecución, ya que al eliminar *reads* de los archivos originales estamos eliminando información. RSEM sigue siendo el que reconoce un menor porcentaje. El resto de programas arrojan porcentajes muy parecidos. En la figura 17 se puede observa un diagrama de barras que representa los porcentajes obtenidos.

El resto de análisis efectuados no presenta diferencias con respecto a los obtenidos anteriormente en el análisis de los archivos completos.

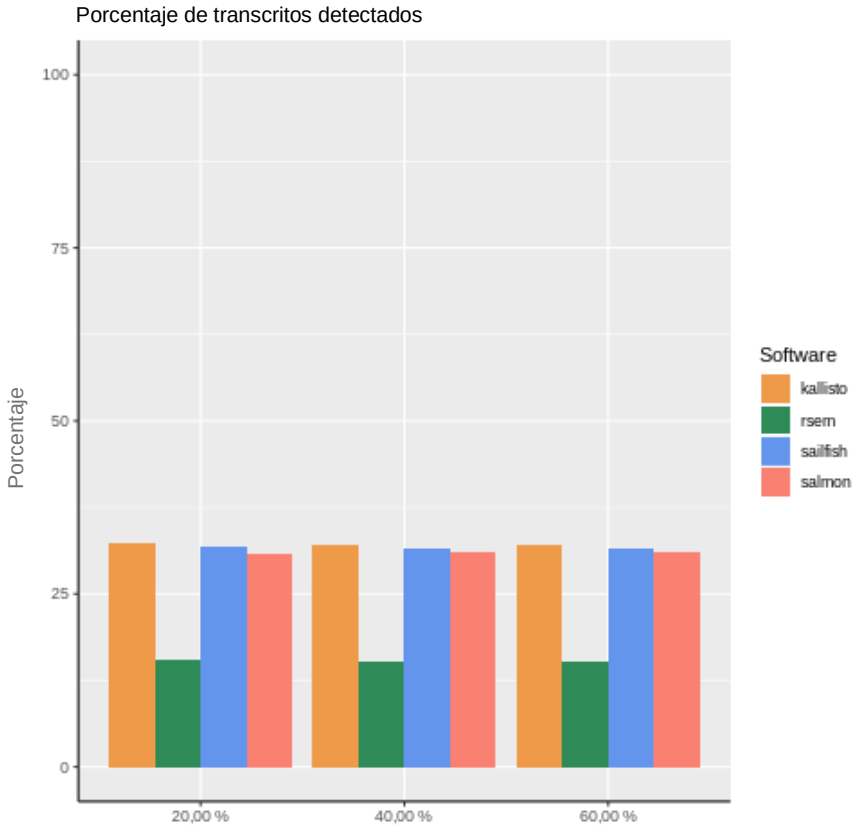


Figura 17: diagrama de barras que representa el porcentaje de transcritos distintos reconocidos por cada programa con tamaños de archivo que se corresponden con el 20%, 40%y 60% de los archivos originales.

2.4 Discusión

- **Tiempos de ejecución:**

El programa que más tiempo emplea en ejecutar el paso de cuantificación es RSEM. Al tener que realizar un paso previo con el programa de alineamiento STAR, el tiempo de ejecución sube, pero incluso sin contabilizar el tiempo empleado por STAR, sigue siendo mayor que los empleados por Kallisto, Sailfish o Salmon. Estos datos siguen la tendencia observada por Zhang et al., (2017)[19] que también reportan un mayor tiempo para RSEM que para Kallisto, Sailfish o Salmon. También se ha reportado que Sailfish es más eficiente con lo que respecta al tiempo que RSEM[28]. Para Kallisto, Sailfish y Salmon se han obtenido resultados muy parecidos entre ellos como ocurre en la comparación de Zhang et al., 2017[19]. Sailfish es el más lento de los 3, aunque aún muy por debajo de RSEM, mientras que Kallisto es el más rápido.

Respecto a cómo varía el tiempo de ejecución de la cuantificación cuando aumenta o disminuye el tamaño de los archivos, nos encontramos con la situación esperada: los archivos de menor tamaño tardan menos tiempo en ser procesados. Sin embargo, lo interesante de este análisis es observar el comportamiento de cada programa al variar el tamaño. La combinación STAR+RSEM es la que peor se adapta al aumento de tamaño mientras que Salmon es el programa más estable. En el primer caso, la dependencia de RSEM del alineamiento previo por parte de STAR es lo que hace que el tiempo total de cuantificación de cada muestra sea mucho mayor que los del resto de programas, ya que el tiempo empleado por RSEM para realizar la cuantificación no es mucho mayor que el de Kallisto, Sailfish y Salmon. La precisión que ofrece STAR en su mapeado (los porcentajes de las muestras están en torno al 90%) hace que la velocidad disminuya, aumentando considerablemente los tiempos de computación, al intentar producir alineamientos óptimos[20].

- **Porcentaje de *reads* mapeados y valores de expresión:**

Este porcentaje es muy informativo ya que de la cantidad de *reads* que el alineador sea capaz de detectar dependerá en gran medida la calidad del análisis. Un porcentaje muy bajo puede ser indicativo de una mala calidad de las muestras, de contaminación o incluso de un error en la ejecución del programa (por ejemplo, mapear una muestra a un genoma distinto al de su especie). Normalmente, dicho porcentaje debe situarse entre el 70 y el 90% en el caso de alinear *reads* de RNA-seq al genoma humano[13].

En este estudio se han obtenido unos porcentajes por encima del 90% para todas las muestras mediante el alineador STAR. Estos porcentajes reflejan una muy buena calidad de las muestras y del alineado. Por el contrario, los porcentajes obtenidos por Sailfish y Salmon son mucho: se sitúan entre el 50 y el 60%, e incluso algunas muestras se encuentran por debajo del 50%. Al tratarse de las mismas muestras, la diferencia en los porcentajes obtenidos radica en cómo construye su índice cada programa. Kallisto no ofrece esta información cuando se efectúa el paso de cuantificación de cada muestra por lo que no se puede conocer el porcentaje sobre el que se mueve.

Estos resultados muestran que STAR es el alineador que mejor desempeña su función, y, por consiguiente, ya que RSEM utiliza los archivos .bam generados por STAR, sus datos de expresión también tendrán una precisión mayor que los de Sailfish y Salmon.

Aun así, aunque el porcentaje de *reads* mapeados en Sailfish y Salmon es bastante más pequeño que el de STAR, los coeficientes de correlación de las medidas de expresión obtenidas por RSEM, Sailfish y Salmon son altos. La existencia de unos coeficientes de correlación altos indican que la influencia del mapeado en los resultados de cuantificación es pequeña[19]: aunque RSEM asegura unos resultados precisos, estos son muy parecidos a los del resto de programas. Aunque no se conozca el porcentaje que mapea Kallisto, también obtiene unos coeficientes de correlación altos con respecto al resto de programas. Una vez más, volvemos a la idea de que no se necesita un alineamiento óptimo entre los fragmentos y el transcriptoma para obtener unos buenos resultados en la cuantificación[20,29].

- **Relación entre los valores de expresión y desviación estándar.**

Para estudiar esta relación se ha utilizado el coeficiente de Spearman. En el caso de que todos los programas proporcionaran exactamente el mismo valor para todos los genes o transcritos, se obtendría una correlación de 1. Cuanto mayor sea la diferencia entre los resultados obtenidos, menor será el coeficiente.

La relación entre los valores de expresión se puede abordar desde el siguiente punto de vista: comparando entre RSEM, que necesita un mapeado previo, y el resto de programas (ninguno necesita este mapeado) y comparando entre los programas que no usan mapeado entre sí.

Respecto al primer caso, tanto para genes como para transcritos existe una alta correlación entre RSEM y el resto de programas. Para transcritos hemos obtenido correlaciones de alrededor de 0,75 en el caso de los tránscritos, y de 0,87 en el caso de los genes, que. además, son todas significativas. Es decir, todos los programas detectan casi por igual la expresión de los tránscritos o genes. Al unificar varios transcritos en un mismo gen sumando sus valores aumenta el valor de los coeficientes de correlación, ya que los distintos programas pueden asignar valores de expresión distintos a diferentes

transcritos, pero al realizar la suma estas diferencias desaparecen. Por tanto, aunque los distintos programas no reconozcan a los transcritos por igual, estas diferencias se hacen menores trabajando a nivel genético. Este aumento del coeficiente de Spearman cuando se trabaja con genes también ha sido reportado previamente en otra comparación de métodos de cuantificación de RNA-seq en el ámbito de la deconvolución[30].

Observando los resultados del segundo caso (comparando entre sí programas que no utilizan un mapeado previo), se puede apreciar que los coeficientes de correlación aumentan en el caso de transcritos y genes hasta 0.993 y 0.994. Esto es una relación casi perfecta. Estos resultados donde se obtiene un mayor coeficiente de correlación entre programas que utilizan el mismo tipo de aproximación también se ha reportado en otros estudios[26]. Sin embargo, sería necesario observar las diferencias en un futuro, ya que en otros estudios se han reportado diferencias entre la estimación de ciertos tipos de ARN, según se usara un mapeado previo o no[19], como por ejemplo, una peor detección por parte de los métodos sin mapeado de ARN poco abundantes o pequeños[26].

La desviación estándar también es más similar entre Kallisto, Sailfish y Salmon que entre RSEM y cualquiera de ellos. RSEM presenta una menor desviación estándar que el resto de programa en los resultados obtenidos.

En los gráficos de relación entre expresión y desviación estándar de cada transcrito o gen, encontramos que aquellos que tienen una mayor expresión también tienen una mayor desviación estándar. Esto está relacionado con la distribución de los datos de expresión: presentan una gran dependencia entre la media y la varianza de su nivel de expresión estimado de manera que los genes más expresados tienden a tener un mayor número de *reads* y por tanto, mayor variabilidad entre muestras, y viceversa[31]. Hay distintas opiniones respecto a la distribución que siguen los datos[32]: ciertos autores usan una distribución de Poisson[33], mientras que otras corrientes utilizan una distribución binomial negativa[33,34], incluso se han realizado aproximaciones con otras distribuciones[31,35]. Utilizar una distribución u otra tiene su importancia en el caso de realizar una búsqueda de los genes diferencialmente expresados[11] (se requieren modelos estadísticos que trabajen bien con las bases de datos

complejas que genera el RNA-seq[31]) o en el caso de contar con datos con dispersiones pequeñas[15]. Tanto si es una u otra, en ambas distribuciones la varianza está relacionada con la media, de manera que, si esta última aumenta, la varianza también lo hace y, por tanto, la desviación típica (recordemos que la desviación típica es la raíz cuadrada de la varianza). Por esta razón, en los datos de RNA-seq los genes o transcritos que están más expresados también tienen una mayor varianza y las distribuciones aquí obtenidas por todos los programas siguen esta lógica.

- **Coeficientes de variación.**

El coeficiente de variación se obtiene dividiendo la desviación estándar de un gen o transcrito a lo largo de varias muestras entre la media de expresión de dicho transcrito o gen. Es una medida que permite observar cómo es la varianza respecto a la media de una observación: el CV es menor si la desviación estándar es pequeña o la media es mayor. En este caso, si todos los transcritos o genes tuvieran la misma desviación estándar, aquellos más expresados (con una media mayor) tendrían un CV más bajo. El CV presenta muchas menos diferencias entre los programas que no necesitan un mapeado previo (Kallisto, Sailifish y Salmon) que entre cualquiera de ellos y RSEM. Esto es debido a la similitud que tienen las medidas de expresión que se han obtenido con estos 3 programas cuyos coeficientes de correlación eran superiores a 0.990.
En el caso de genes el coeficiente de correlación aumenta hasta casi un 0.900 para las comparaciones entre RSEM y el resto de programas. Las medidas de expresión obtenidas a nivel genético son más parecidas entre sí que a nivel de transcritos.

- **Número de transcritos/genes distintos reconocidos por cada programa.**

RSEM es el programa que detecta un menor número de transcritos y, por tanto, de genes. Esto puede estar relacionado con el porcentaje de *reads* mapeados. STAR proporciona más de un 90% de *reads* mapeados únicamente a un transcrito o gen, lo que se traduce en una gran precisión en el mapeado. Esto puede repercutir en el nº de observaciones reconocidas, ya que una mayor

precisión significa que una mayor proporción de *reads* están asignados a un solo fragmento, y, por tanto, habrá menos fragmentos a los que se ha detectado expresión. Sailfish y Salmon, que tienen un porcentaje de observaciones mapeado muy parecido y más bajo que el de STAR, reconocen una mayor proporción de transcritos y genes y en una cantidad muy parecida entre ellos y también Kallisto.

- **Análisis por tamaños**

Los resultados obtenidos reduciendo el tamaño de los archivos no aportan ningún resultado distinto a los ya reportados en los apartados de porcentaje de *reads* mapeados, coeficientes de correlación de la expresión medida por cada programa, desviación estándar o coeficiente de variación. Los únicos cambios que se han obtenido están directamente relacionados con el propio tamaño de los archivos.

El tiempo de ejecución de la fase de cuantificación de cada programa depende en parte del tamaño de los archivos a analizar, por lo que se esperaba una disminución del tiempo de ejecución al eliminar ciertos porcentajes de *reads* de los archivos originales. RSEM sigue siendo el programa que más tiempo emplea en realizar el análisis debido a su dependencia del mapeado anterior por STAR. Este mapeado es el que hace que el tiempo que se necesita con RSEM para obtener los resultados sea mucho mayor que con los otros 3 programas. Kallisto, Sailfish y Salmon presentan tiempos muy parecidos, siendo siempre Salmon el más rápido, aunque las diferencias entre ellos son mucho más pequeñas que las que presenta cualquiera de ellos con RSEM.

La otra diferencia encontrada al disminuir el tamaño de los archivos se encuentra en el número de transcritos y genes distintos detectados por cada programa. Al eliminar al azar parte de la información analizada con anterioridad estamos eliminando transcritos que lógicamente no serán contabilizados. Esta es la razón por la que se obtienen porcentajes más bajos con todos los programas al tratar con archivos más pequeños.

En definitiva, teniendo en cuenta que todas estas diferencias están producidas por el tamaño de los archivos, se puede concluir que ningún programa disminuye su rendimiento al trabajar con archivos más pequeños.

Por último, es necesario destacar el aumento en la correlación de los resultados obtenidos analizando los resultados obtenidos por gen, debido a que al sumar las estimaciones de expresión de los distintos transcritos de un gen, se eliminan las diferencias existentes entre los programas al reconocer transcritos diferentes. Esta observación también había sido reportada anteriormente[36].

Teniendo en cuenta todos los resultados, en caso de querer reducir los tiempos de computación, el programa que escogería sería Salmon. Es el más estable en el tiempo de análisis conforme cambia el tamaño de los archivos analizado e incorpora información durante la estimación de la expresión (fases *online* y *offline*) que no tienen en cuenta Sailfish ni Kallisto. Además, también permite usar un algoritmo EM o un EM bayesiano en la estimación según el objetivo del estudio.
En el caso de que no tuviera importancia el tiempo de computación (archivos de pequeño tamaño o poca cantidad de archivos) o se necesitaran resultados muy precisos, seguiría utilizando la combinación STAR+RSEM ya que el mapeado de STAR asegura esa precisión.

3. Conclusiones

- Las diferencias entre los programas de cuantificación de datos de RNA-seq aquí analizados están ocasionadas principalmente por el método que utiliza cada uno (mapeado vs pseudoalineamiento o *quasi-mapping*).
- Kallisto, Sailfish y Salmon reducen notablemente el tiempo de ejecución necesario para efectuar el análisis respecto a RSEM y son menos susceptibles al aumento en el número de *reads* de los archivos fastq.
- Realizar un mapeado previo no tiene demasiada influencia en los resultados de cuantificación: los coeficientes de correlación son altos entre los programas.
- Los resultados de expresión son mucho más parecidos cuanto mayor es la similitud entre el funcionamiento de cada programa.
- Los resultados de expresión son mucho más parecidos cuando se trabaja a nivel de gen y no de transcrito.
- Las reducciones en el número de *reads* que contienen los archivos fastq no alteran los resultados obtenidos más allá del tiempo en ser analizadas o del nº de transcritos y genes reconocidos por cada programa.

3.1 Cumplimiento de los objetivos

Respecto al cumplimiento de los objetivos del trabajo, estos se consideran totalmente cumplidos. Los 5 programas mencionados en el objetivo a.1 han sido descargados y ejecutados, comprendiendo el uso que tiene cada uno y su funcionamiento interno. También se ha ejecutado un pipeline apropiado para cada programa para poder automatizar en la medida de lo posible el análisis de las 17 muestras por los 5 programas y, posteriormente, las 3 reducciones de cada una de las muestras por cada programa. Por tanto, una vez construidos los pipelines, fue posible analizar los datos completamente y obtener los resultados, realizando su análisis y comparación de los resultados.

3.2 Seguimiento de la planificación

El seguimiento de la planificación ha sido prácticamente el previsto al inicio del proyecto. Aunque hubo cierto retraso en la descarga de las muestras para

analizar, se comenzó la redacción de la introducción y materiales y métodos del proyecto final antes de lo previsto para compensar el retraso. El tiempo necesario para la descarga de las muestras también hizo imposible repetir los análisis con muestras single-end. Respecto al análisis de las mismas muestras, pero con un menor tamaño, lo ideal habría sido ir reduciendo el tamaño en un 10% pero debido a los tiempos necesarios de computación no fue posible hacerlo y se optó por realizar solamente 3 reducciones (del 40, 60 y 80%).

3.3 Vistas al futuro

Respecto hacia dónde se podría avanzar en un proyecto posterior que siga las líneas aquí planteadas, lo primero sería repetir el análisis con muestras single-end. Esto estaba previsto en este proyecto, pero por una limitación temporal no ha sido posible realizarlo. Debido a que algunos estudios encuentran limitaciones en los programas que no utilizan un mapeado de las muestras, sería interesante analizar las diferencias en la detección de los distintos tipos de RNA según el software utilizado para su análisis. Por último, también se podría realizar un análisis de expresión diferencial con los distintos programas para comprobar si se obtienen los mismos resultados.

4. Glosario

Cola poly-A: secuencia de varias adeninas que se añaden al final del ARNm maduro para aumentar su estabilidad.

Cuantificación: cálculo de cuántos fragmentos se unen a cada zona del genoma.

Formato BAM: Binary Alignment/Map - versión comprimida del archivo SAM con al menos 11 campos de información sobre el alineado.

Formato FASTQ: formato con información sobre secuencias biológicas (ID y secuencias nucleotídicas) y el correspondiente índice de calidad para cada base.

Hashing: generación de un valor para cada elemento de una cadena de caracteres para acelerar el procesado de dichos datos.

k-mer: todas las sub-secuencias de longitud *k* dentro de una secuencia.

Mapeado: asociación de cada fragmento de ADN complementario a la zona del genoma a la que corresponde.

Microarray: técnica de laboratorio para detectar la medida de expresión genética de miles de genes simultáneamente.

Reads: o lecturas de secuencias que a su vez es información sobre billones de bases nucleotídicas.

RNA-seq: la aplicación de cualquier técnica de secuenciación de nueva generación al estudio del ARN.

Secuenciación paired-end: secuenciación de los 2 extremos de un fragmento de ADN.

Secuenciación single-end: secuenciación de un solo extremo de un fragmento de ADN.

Splicing: proceso post-transcripcional mediante el cual se eliminan ciertos fragmentos de del ARN durante su maduración.

Transcripción inversa: conversión de ARN a ADNc.

Transcriptasa: enzima que realiza el proceso de transcripción inversa.

Transcriptómica: conjunto completo de transcritos que hay en una célula y su cantidad en cierto momento del desarrollo o bajo condición fisiológica.

5. Bibliografía

1. Chu, Y. & Corey, D. R. RNA Sequencing: Platform Selection, Experimental Design, and Data Interpretation. *Nucleic Acid Ther.* **22**, 271–274 (2012).
2. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).
3. Yang, I. S. & Kim, S. Analysis of Whole Transcriptome Sequencing Data: Workflow and Software. *Genomics Inform.* **13**, 119–125 (2015).
4. Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. & Gilad, Y. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* **18**, 1509–1517 (2008).
5. Kukurba, K. R. & Montgomery, S. B. RNA Sequencing and Analysis. *Cold Spring Harb. Protoc.* **2015**, 951–969 (2015).
6. Griffith, M., Walker, J. R., Spies, N. C., Ainscough, B. J. & Griffith, O. L. Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud. *PLOS Comput. Biol.* **11**, e1004393 (2015).
7. Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K. & Liu, X. Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. *PLOS ONE* **9**, e78644 (2014).
8. Jiang, H., Lei, R., Ding, S.-W. & Zhu, S. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics* **15**, 182 (2014).
9. Andrews, S. FastQC A Quality Control tool for High Throughput Sequence Data. Available at:

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. (Accessed: 30th December 2018)

10. Index of /projects/fastqc/Help/3 Analysis Modules. Available at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/. (Accessed: 26th December 2018)
11. Zhao, S. Assessment of the Impact of Using a Reference Transcriptome in Mapping Short RNA-Seq Reads. *PLOS ONE* **9**, e101374 (2014).
12. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
13. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17**, 13 (2016).
14. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
15. Yoon, S. & Nam, D. Gene dispersion is the key determinant of the read count bias in differential expression analysis of RNA-seq data. *BMC Genomics* **18**, 408 (2017).
16. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
17. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference. *Nat. Methods* **14**, 417–419 (2017).
18. Patro, R., Mount, S. M. & Kingsford, C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat. Biotechnol.* **32**, 462–464 (2014).

19. Zhang, C., Zhang, B., Lin, L.-L. & Zhao, S. Evaluation and comparison of computational tools for RNA-seq isoform quantification. *BMC Genomics* **18**, 583 (2017).
20. Srivastava, A., Sarkar, H., Gupta, N. & Patro, R. RapMap: A Rapid, Sensitive and Accurate Tool for Mapping RNA-seq Reads to Transcriptomes. *bioRxiv* 029652 (2016). doi:10.1101/029652
21. Homo sapiens (ID 473142) - BioProject - NCBI. Available at: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA473142/. (Accessed: 16th December 2018)
22. Babraham Bioinformatics - Trim Galore! Available at: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/. (Accessed: 26th December 2018)
23. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinforma. Oxf. Engl.* **25**, 1105–1111 (2009).
24. RStudio Team. *RStudio: Integrated Development for R. RStudio, Inc.* (2015).
25. R Development Core Team. *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. (2008).
26. Wu, D. C., Yao, J., Ho, K. S., Lambowitz, A. M. & Wilke, C. O. Limitation of alignment-free tools in total RNA-seq quantification. *bioRxiv* 246967 (2018). doi:10.1101/246967
27. Li, H. *Toolkit for processing sequences in FASTA/Q formats: lh3/seqtk*. (2018).
28. Li, P., Piao, Y., Shon, H. S. & Ryu, K. H. Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data. *BMC Bioinformatics* **16**, 347 (2015).

29. Patro, R., Duggal, G. & Kingsford, C. Salmon: Accurate, Versatile and Ultrafast Quantification from RNA-seq Data using Lightweight-Alignment. *bioRxiv* 021592 (2015). doi:10.1101/021592
30. Jin, H., Wan, Y.-W. & Liu, Z. Comprehensive evaluation of RNA-seq quantification methods for linearity. *BMC Bioinformatics* **18**, 117 (2017).
31. Sun, S. *et al.* Differential expression analysis for RNAseq using Poisson mixed models. *Nucleic Acids Res.* **45**, e106 (2017).
32. Pachter, L. Models for transcript quantification from RNA-Seq. *ArXiv11043889 Q-Bio Stat* (2011).
33. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
34. Di, Y., Schafer, D. W., Cumbie, J. S. & Chang, J. H. The NBP Negative Binomial Model for Assessing Differential Gene Expression from RNA-Seq. *Stat. Appl. Genet. Mol. Biol.* **10**, (2011).
35. Zhang, Q. Classification of RNA-Seq Data via Gaussian Copulas. *bioRxiv* 116046 (2017). doi:10.1101/116046
36. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research* **4**, 1521 (2015).

6. Anexos

6.1 Otros

A la hora de construir el índice de STAR se usó 100 como valor del argumento *--sjdbOverhang* ya que el manual del programa aconseja 100 como valor por defecto que funciona como valor ideal en la gran mayoría de los casos. Para el mapeado del mismo programa se utilizó la opción *BAM Unsorted* en el argumento *–outSAMtype*.

En la construcción del índice de RSEM se utilizó el argumento *–star* para construir el índice usando el programa STAR. Para realizar la cuantificación, se utilizaron los archivos .bam generados en el mapeado de STAR, junto con las opciones *–paired-end* (en este análisis se han usado muestras paired-end) y *–no-bam-output* (no eran necesarios para este estudio).

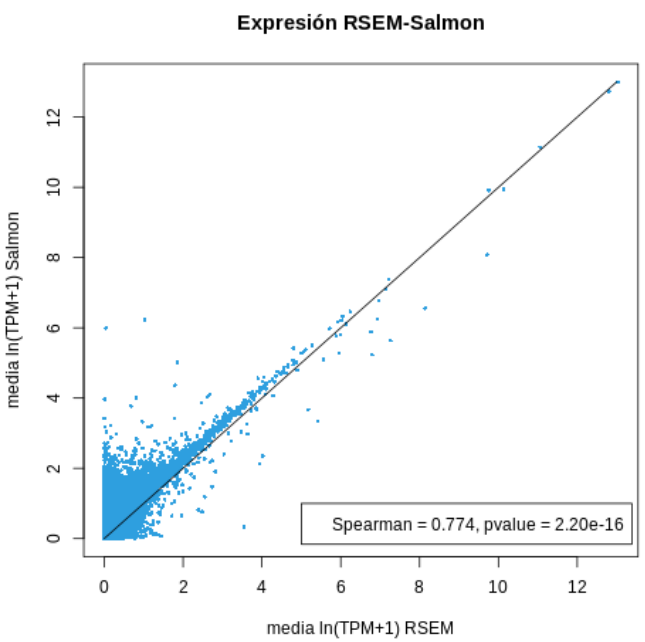
En el caso de la construcción del índice del programa Kallisto no se han usado más opciones aparte de las imprescindibles. En el paso de cuantificación tampoco se han usado argumentos distintos a los imprescindibles.

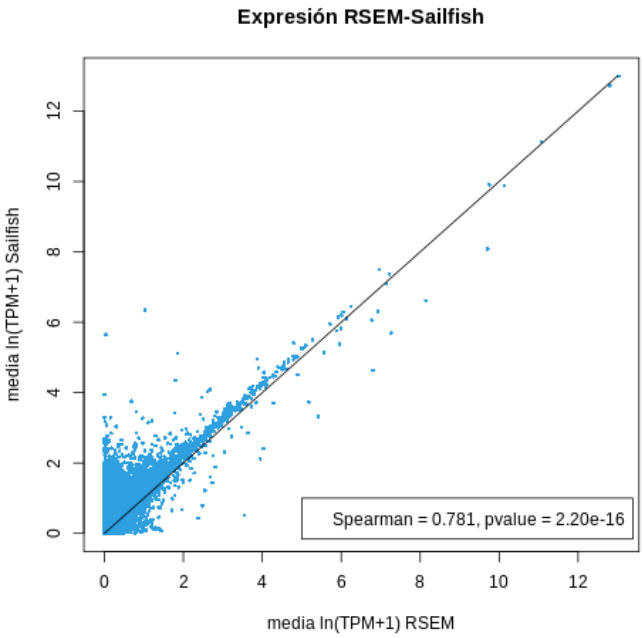
Construyendo el índice de Sailfish se usó 31 como valor de *-k* y se especificó que los datos eran *paired-end undstranded,* en el paso de cuantificación, mediante la opción IU en el argumento *-l*.

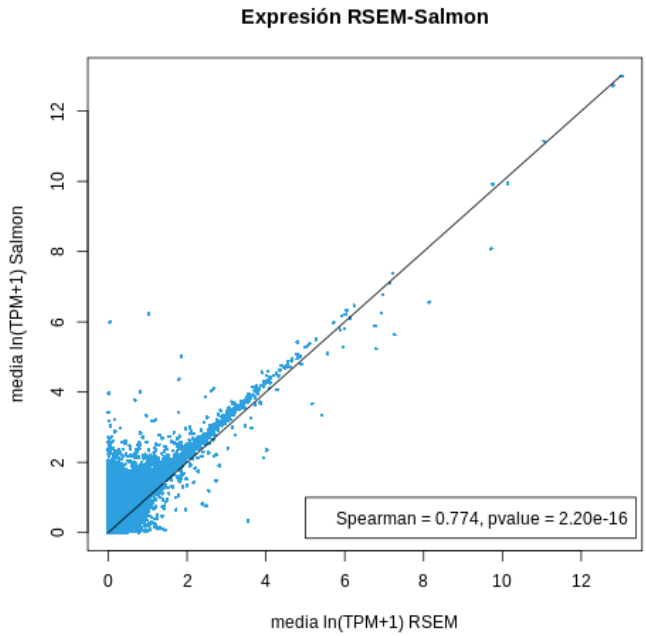
Por último, en la construcción del índice de Salmon se usó también 31 como valor de *-k* y se añadió la opción *–keepDuplicates* para que el programa no eliminase las secuencias que son totalmente idénticas y poder obtener el mismo número de lecturas que el resto de programas. El mapeado realizado es *quasi mapping*. En el paso de cuantificación se utilizó la opción *IU* en el argumento *-l* para especificar que los datos eran *paired end unstranded*.

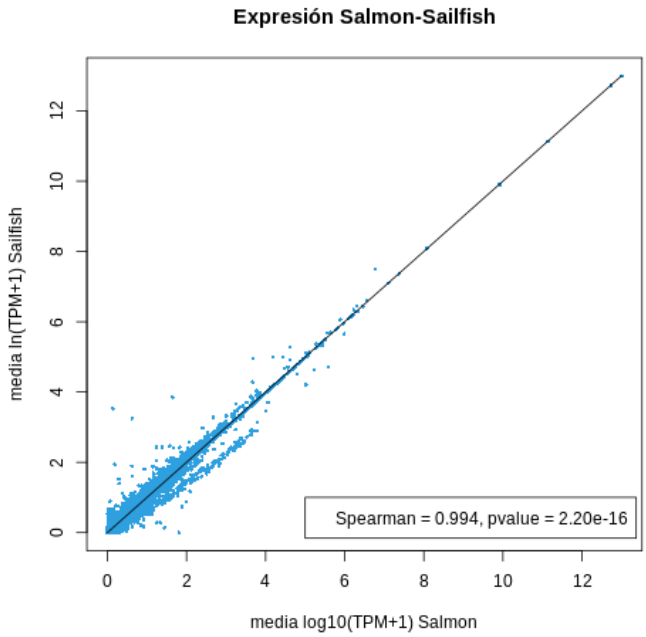
6.2 Figuras

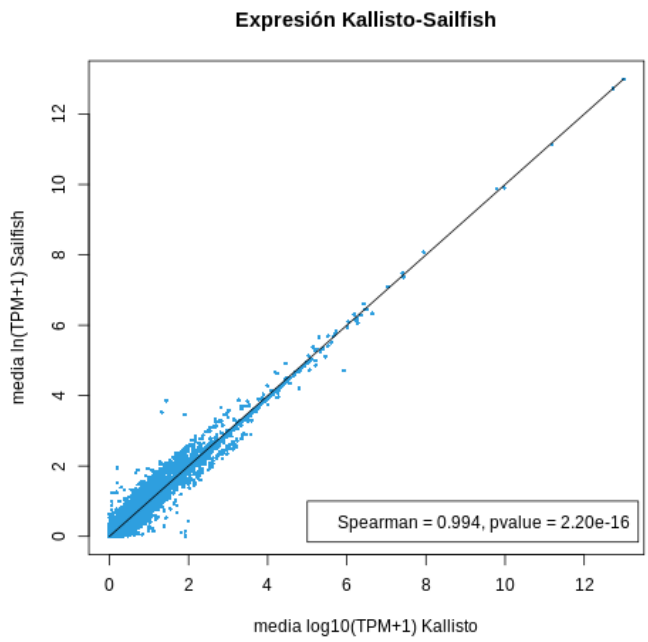
1- Correlación entre la media de la expresión de cada transcrito entre RSEM-Salmon, RSEM-Sailfish, Kallisto-Sailfish, Kallisto-Salmon y Sailfish-Salmon. La diagonal observada no se corresponde con la línea de regresión de los datos, solamente se ha añadido para una mejor interpretación.



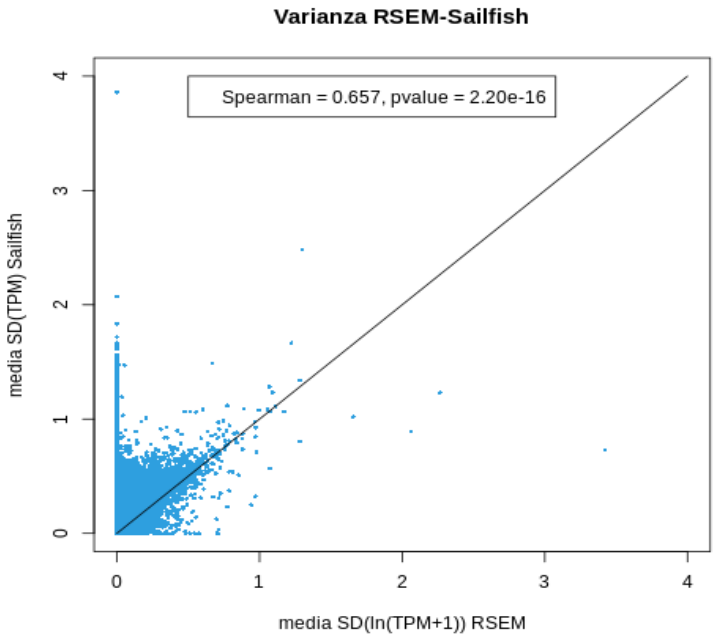


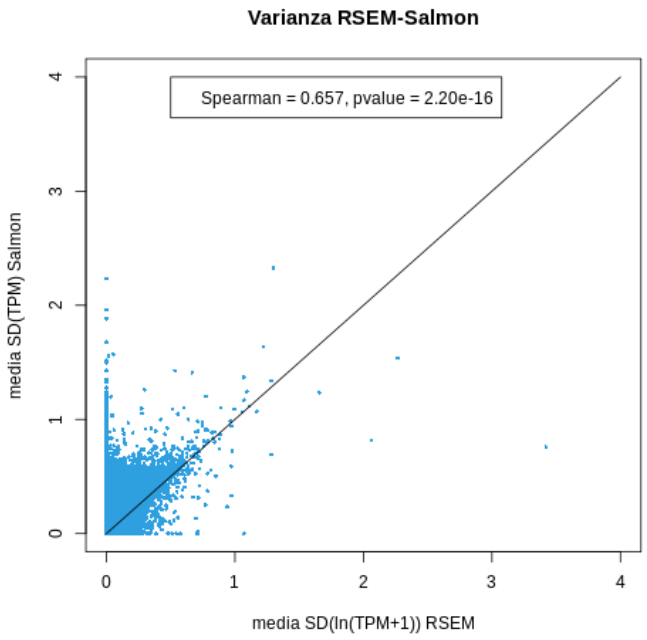


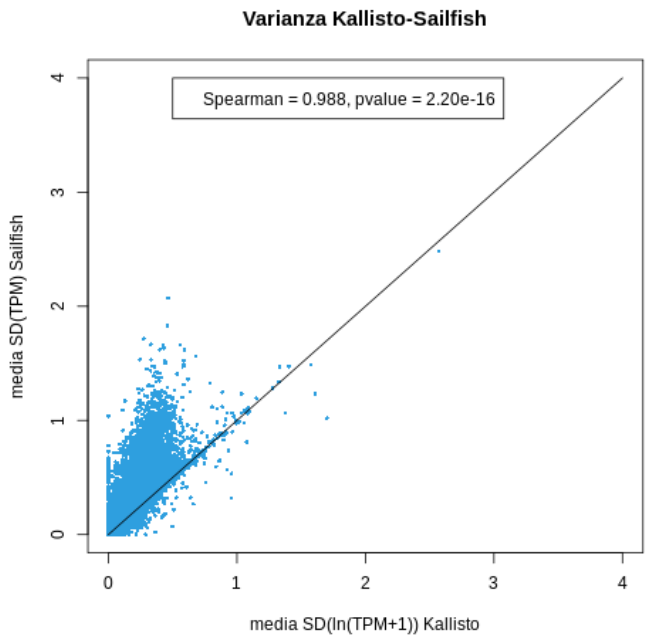


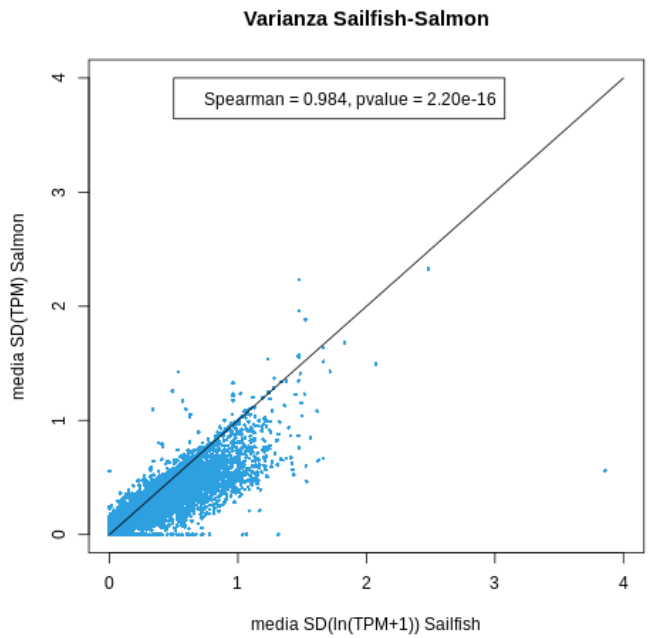


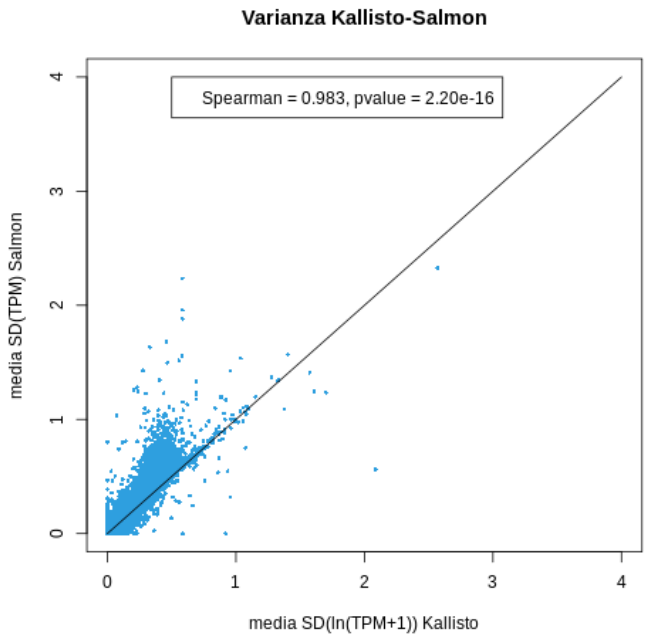
2- Correlación entre la media de la desviación estándar de cada transcrito obtenida por RSEM-Salmon, RSEM-Sailfish, Kallisto-Salmon, Kallisto-Safilsh y Sailfish-Salmon. La diagonal observada no se corresponde con la línea de regresión de los datos, solamente se ha añadido para una mejor interpretación.



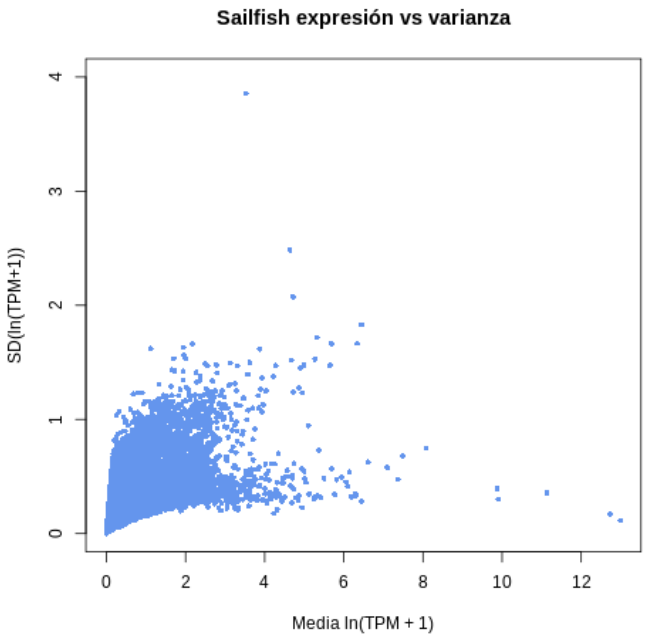


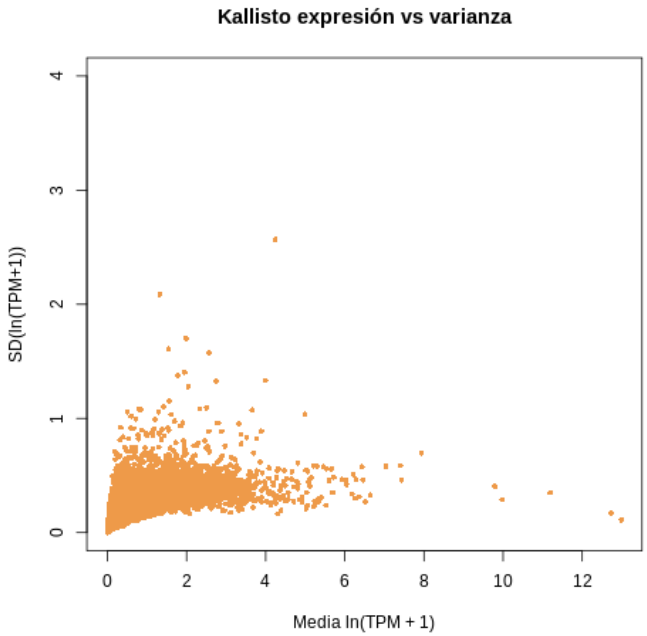

Varianza Kallisto-Sailfish
Spearman = 0.988, pvalue = 2.20e-16
media SD(TPM) Sailfish
media SD(ln(TPM+1)) Kallisto
0
1
2
3
4


Varianza Sailfish-Salmon
Spearman = 0.984, pvalue = 2.20e-16
media SD(TPM) Salmon
media SD(ln(TPM+1)) Sailfish
0
1
2
3
4

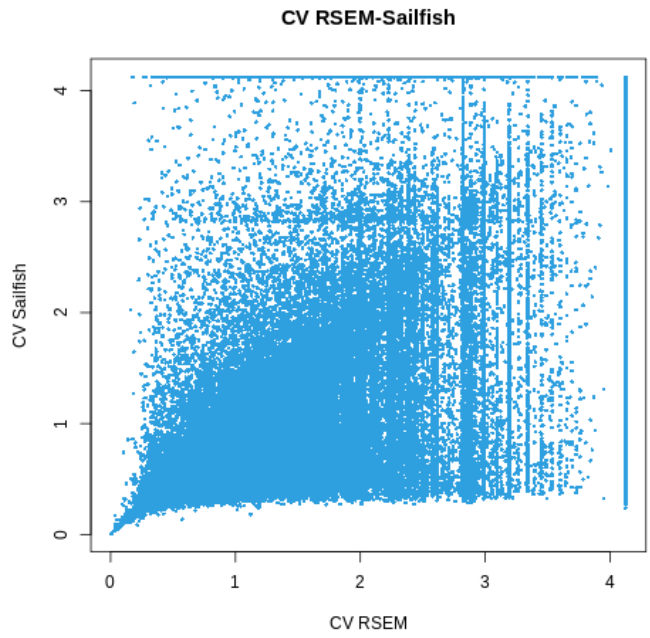

Varianza Kallisto-Salmon
Spearman = 0.983, pvalue = 2.20e-16
media SD(TPM) Salmon
media SD(ln(TPM+1)) Kallisto
0
1
2
3
4

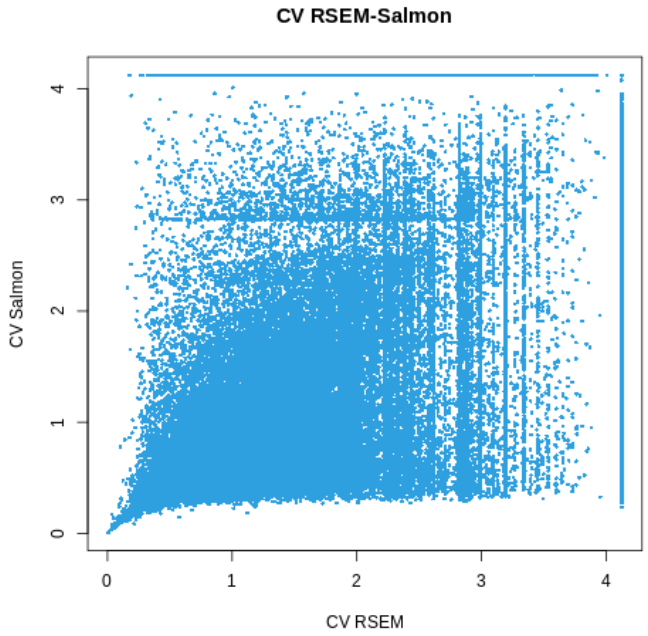
3- Relación entre la media del logaritmo de la expresión de cada transcrito con la media del logaritmo de la desviación estándar de dicha expresión en Kallisto, Sailfish y Salmon.

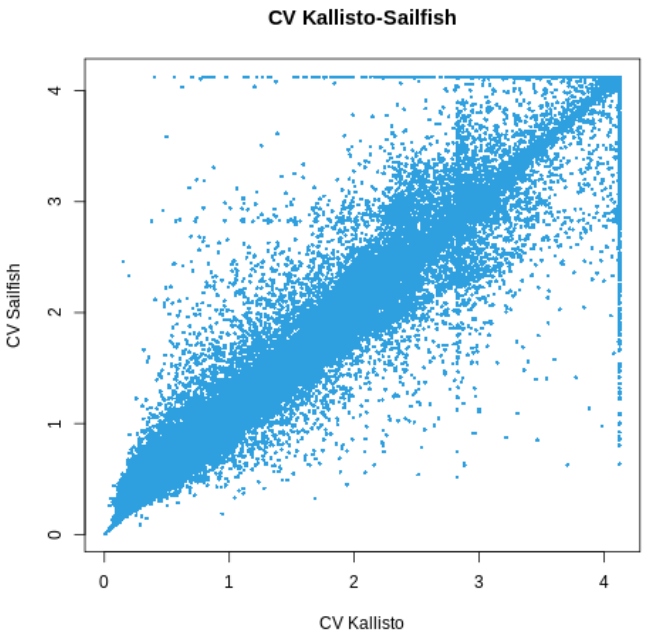


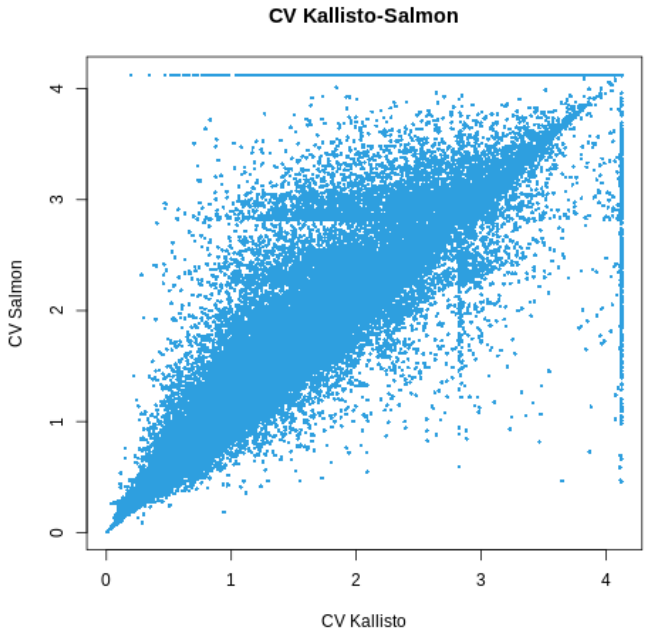


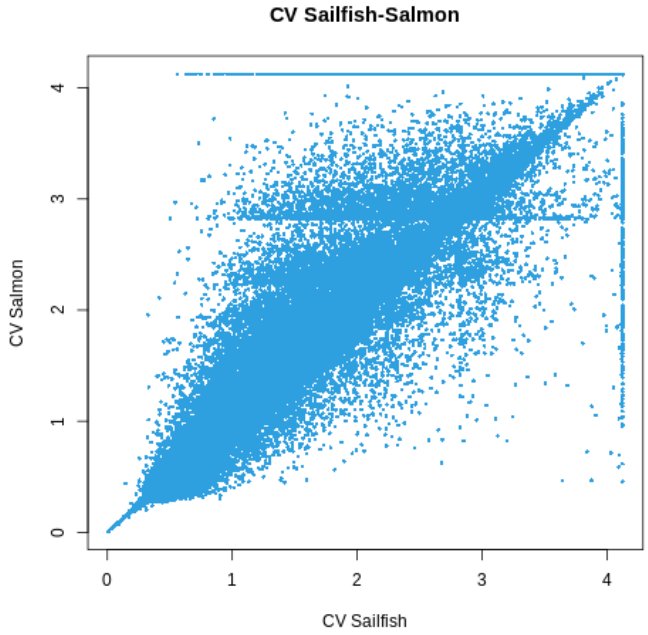
4- Correlación entre los CV obtenidos para cada transcrito entre RSEM-Sailfish, RSEM-Salmon, Kallisto-Sailfish, Kallisto-Salmon y Sailfish-Salmon.



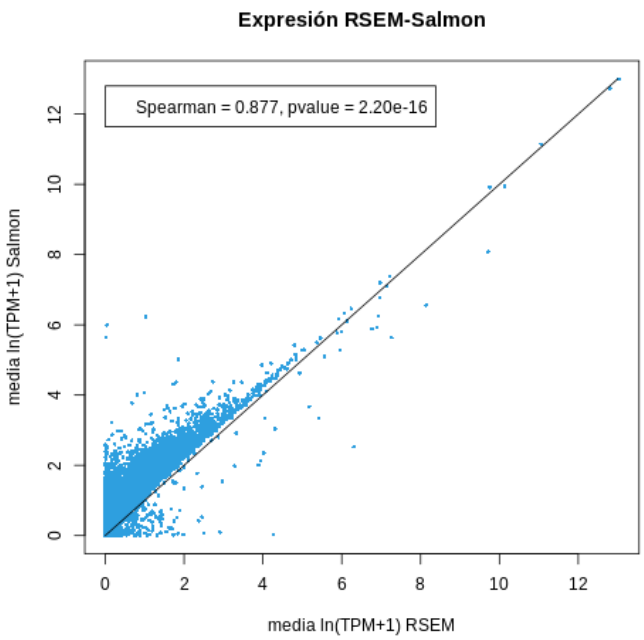


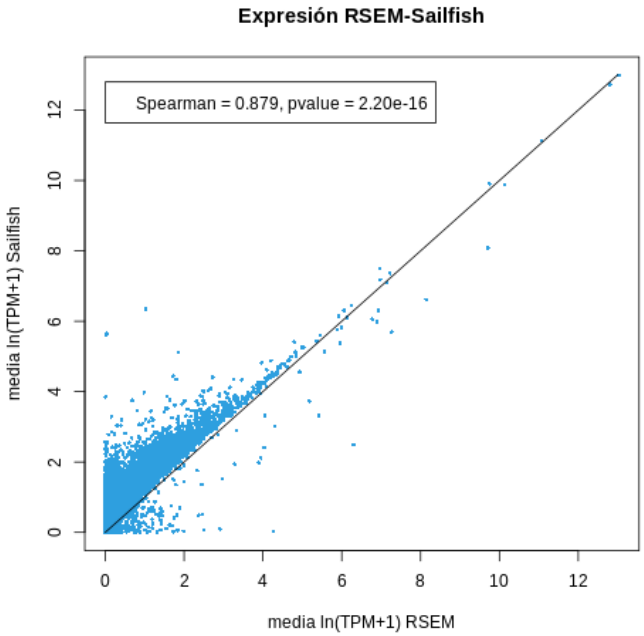


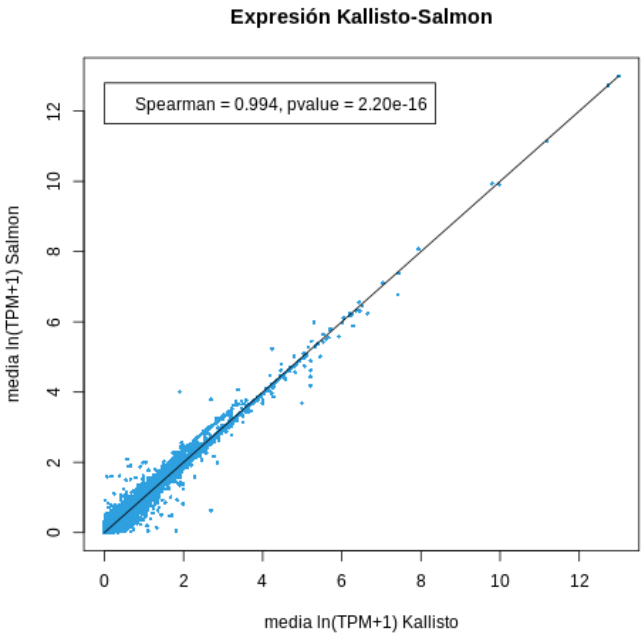


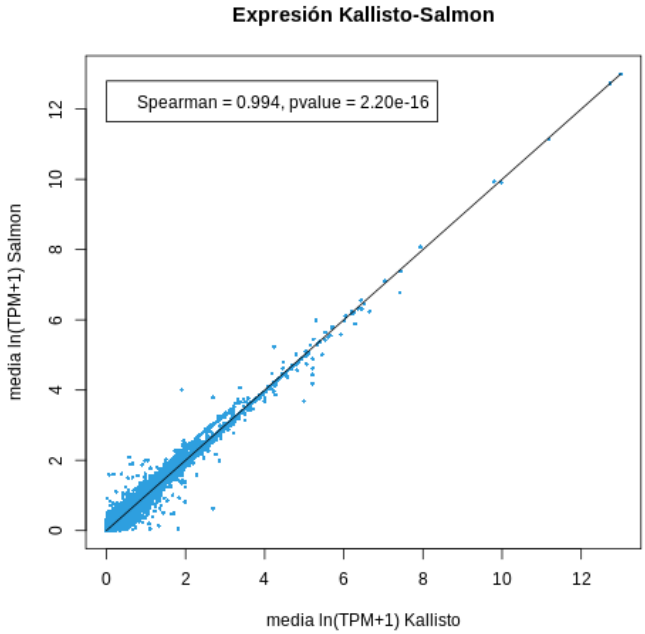


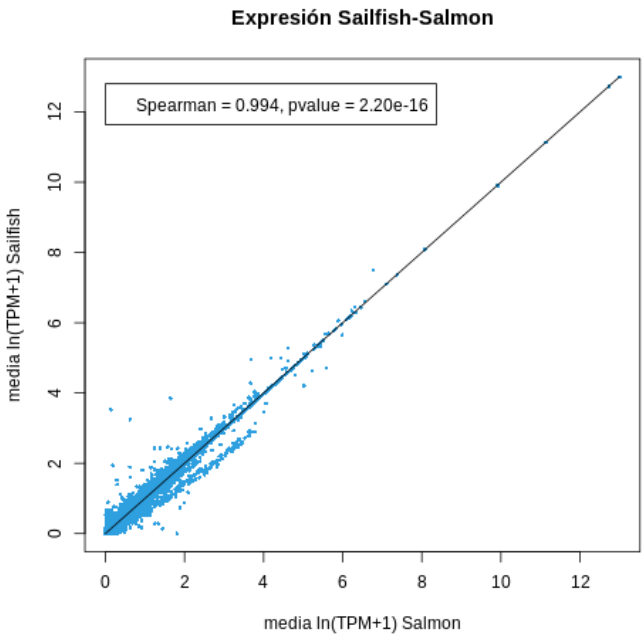
5- Correlación entre la media de la expresión de cada gen entre RSEM-Salmon, RSEM-Sailfish, Kallisto-Sailfish, Kallisto-Salmon y Sailfish-Salmon. La diagonal observada en cada gráfico se ha añadido para una mejor visualización y análisis de los datos y no corresponde a la correlación entre ellos.



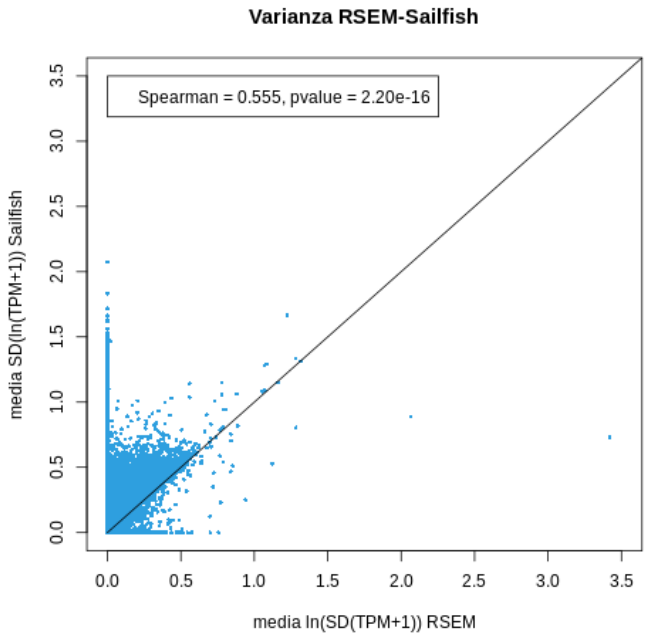


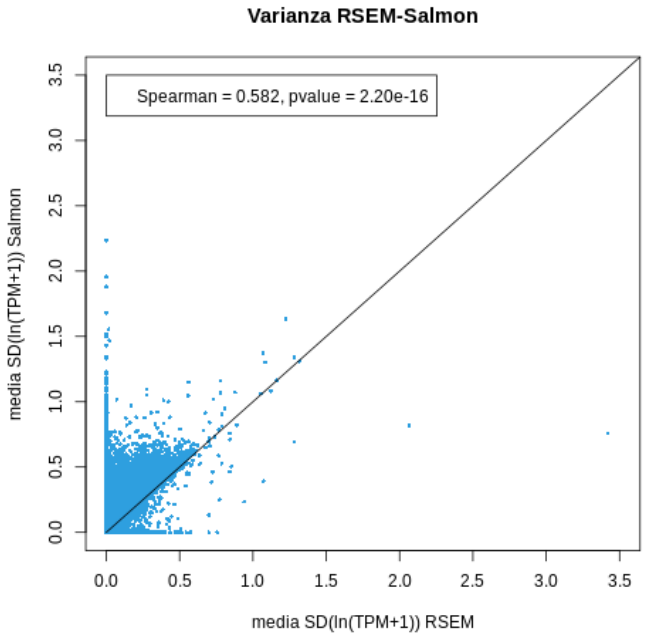

Expresión Kallisto-Salmon
Spearman = 0.994, pvalue = 2.20e-16
media ln(TPM+1) Salmon
media ln(TPM+1) Kallisto
0
2
4
6
8
10
12

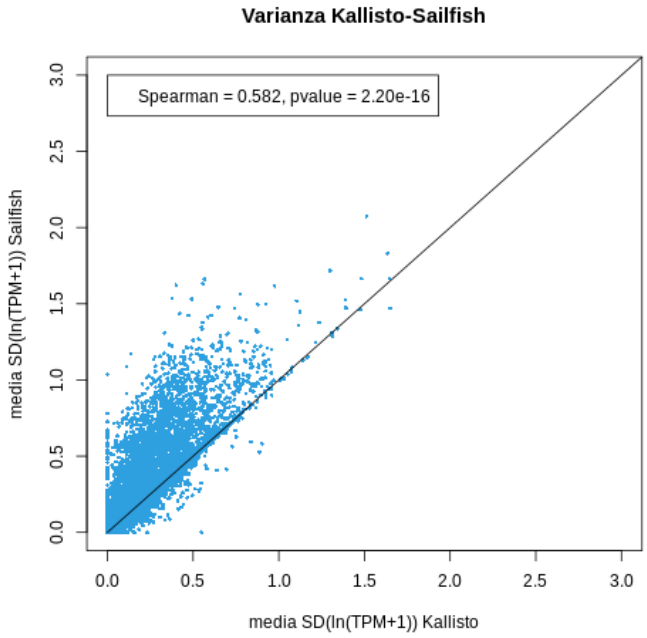

Expresión Kallisto-Salmon
Spearman = 0.994, pvalue = 2.20e-16
media ln(TPM+1) Salmon
media ln(TPM+1) Kallisto
0
2
4
6
8
10
12

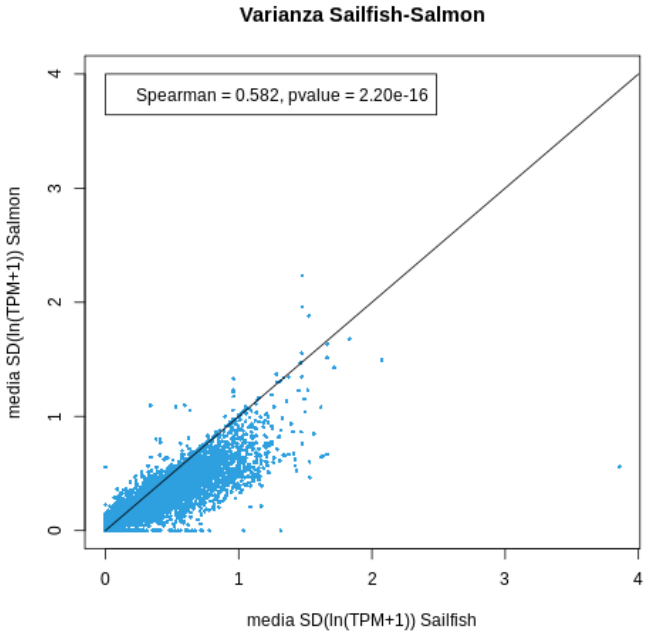

Expresión Sailfish-Salmon
Spearman = 0.994, pvalue = 2.20e-16
media ln(TPM+1) Sailfish
media ln(TPM+1) Salmon
0
2
4
6
8
10
12

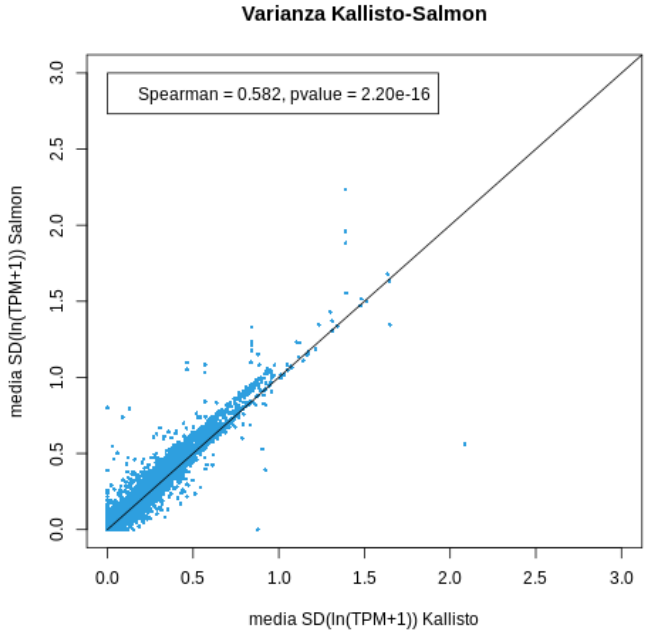
6- Correlación entre la media de la desviación estándar de cada gen obtenida por RSEM-Sailfish, RSEM-Salmon, Kallisto-Sailfish, Kallisto-Salmon y Sailfish-Salmon. La diagonal observada en cada gráfico se ha añadido para una mejor visualización y análisis de los datos y no corresponde a la correlación entre los datos.



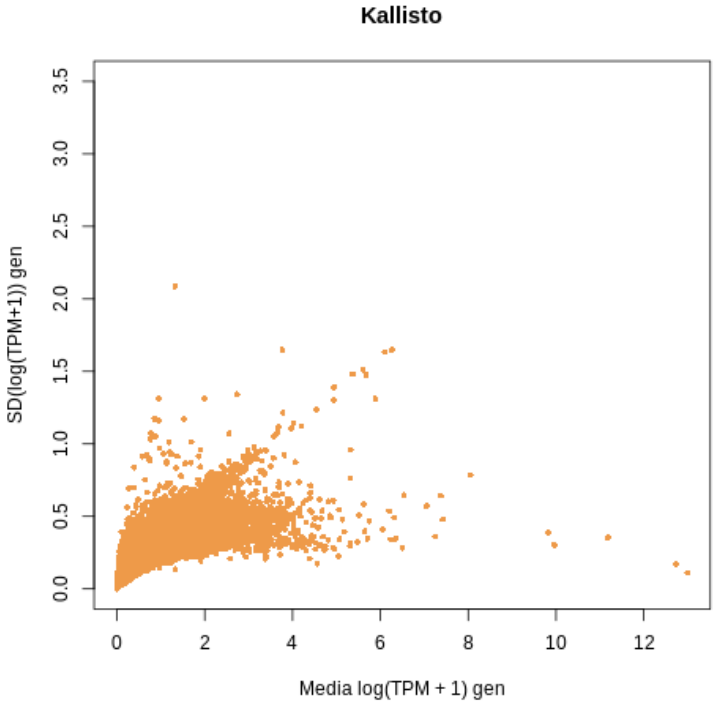


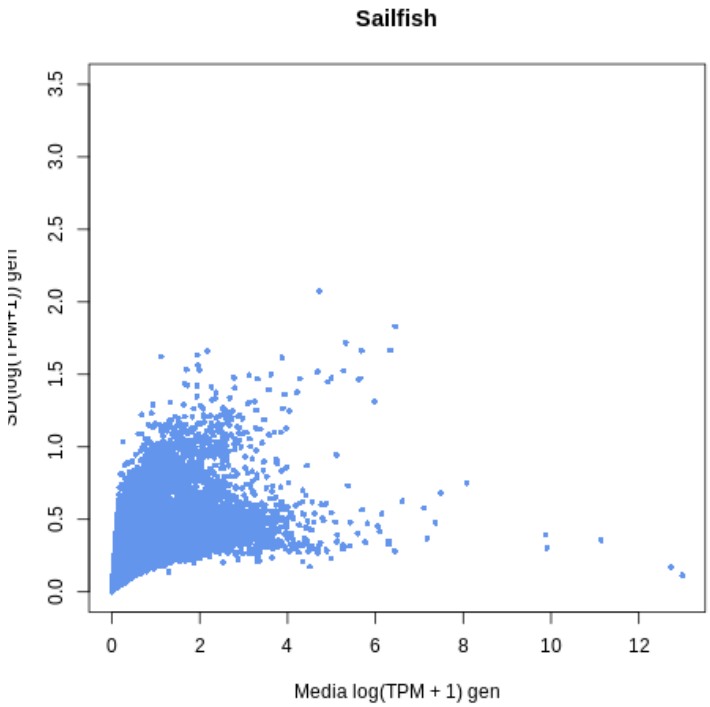


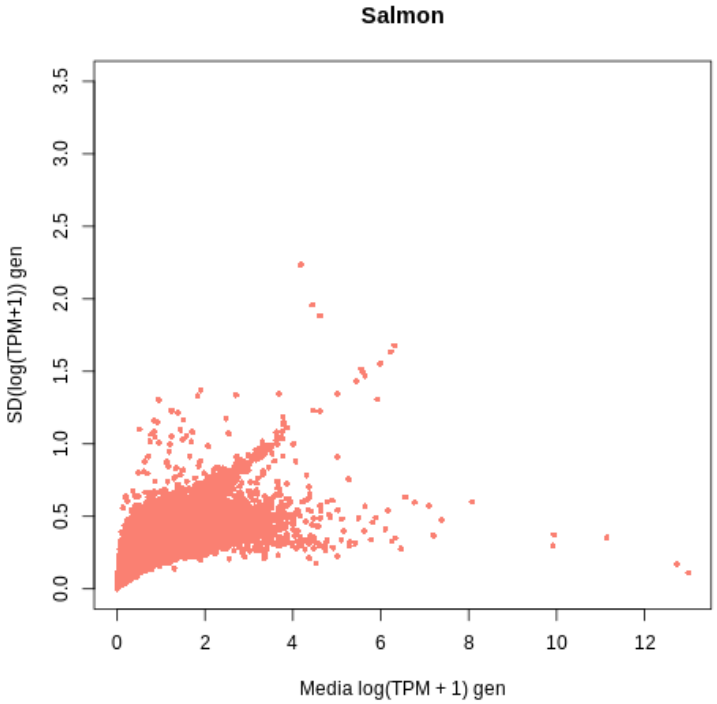




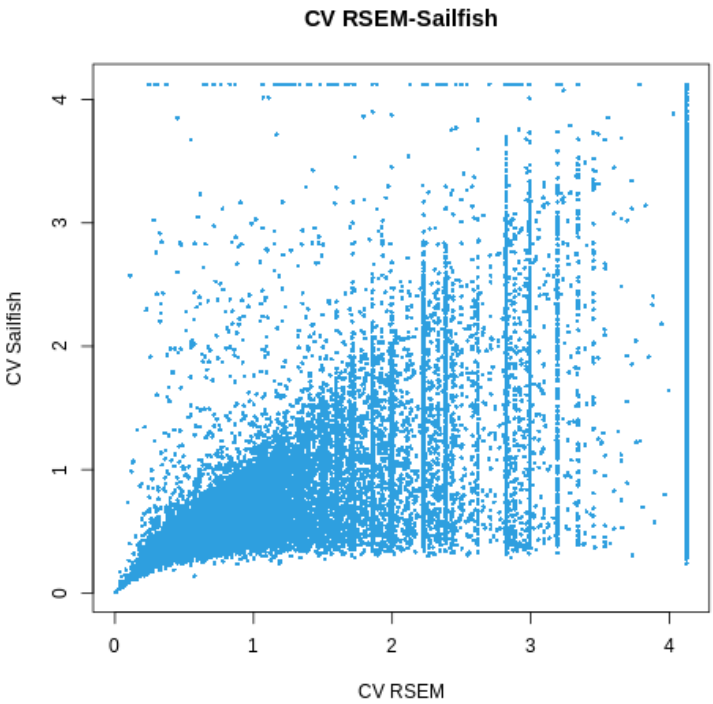
7- Relación entre la media del logaritmo de la expresión de cada gen con la media del logaritmo de la desviación estándar de dicha expresión en Kallisto, Sailfish y Salmon

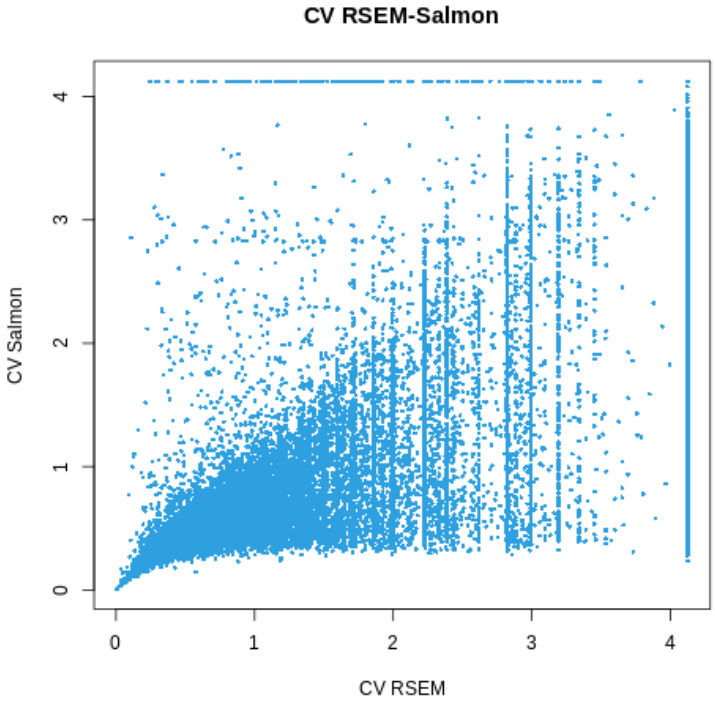


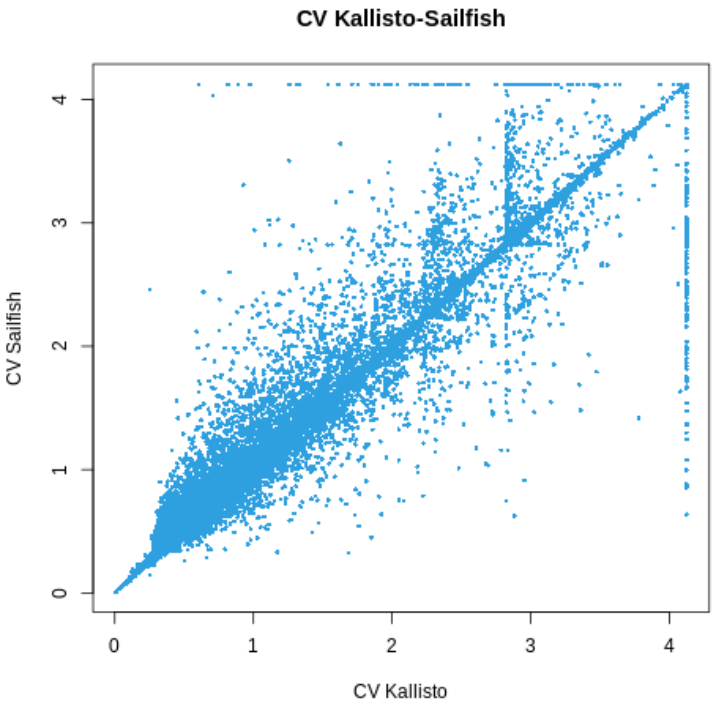


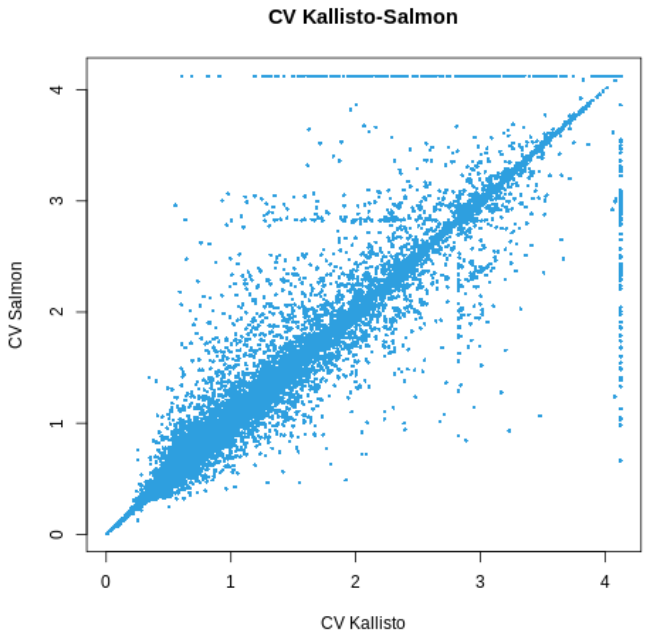


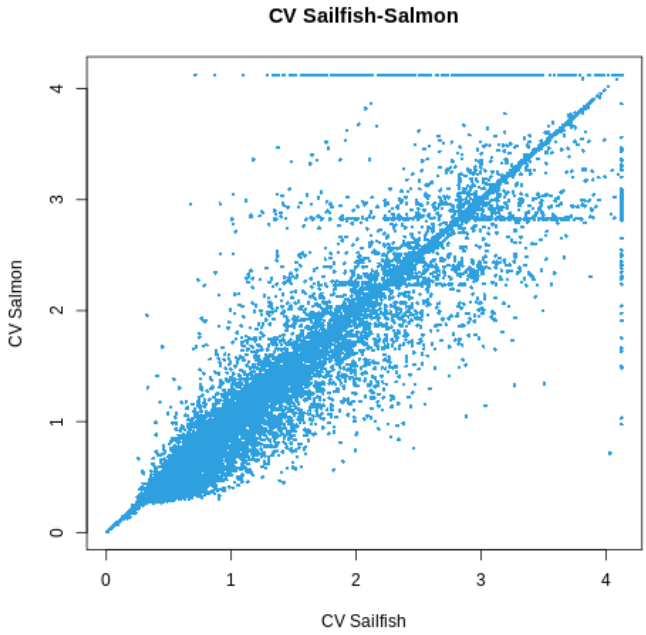
8- Correlación entre los CV obtenidos para cada transcrito entre RSEM-Sailfish, RSEM-Salmon, Kallisto-Sailfish y Sailfish-Salmon.






CV Kallisto-Sailfish
CV Sailfish
CV Kallisto
0
1
2
3
4


CV Kallisto-Salmon
CV Salmon
CV Kallisto
0
1
2
3
4


CV Sailfish-Salmon
CV Salmon
CV Sailfish
0
1
2
3
4

6.3 Tablas

1- Tabla con los p-valores obtenidos en el test de la U de Mann-Whitney para la igualdad de medias y para el test de Kolmogorov-Smirnov de igualdad en la distribución.

Comparación	U de Mann – Whitney (p-valor)	Kolmogorov-Smirnov (p-valor)
Rsem - Kallisto	< 2,2 e-16	< 2,2 e-16
Rsem - Sailfish	< 2,2 e-16	< 2,2 e-16
Rsem - Salmon	< 2,2 e-16	< 2,2 e-16
Kallisto - Sailfish	< 2,2 e-16	< 2,2 e-16
Kallisto - Salmon	< 2,2 e-16	< 2,2 e-16
Sailfish - Salmon	< 2,2 e-16	< 2,2 e-16

Tabla X: p-valores obtenidos en la comparación de medias de la varianza de los programas con el test de Mann-Whitney y en la comparación de la distribución de la varianza con el test de Kolmogorov-Smirnov.

2- Tabla con los p-valores obtenidos en el test de la U de Mann-Whitney para la igualdad de medias y para el test de Kolmogorov-Smirnov de igualdad en la distribución.

Comparación	U de Mann – Whitney (p-valor)	Kolmogorov-Smirnov (p-valor)
Rsem - Kallisto	< 2,2 e-16	< 2,2 e-16
Rsem - Sailfish	< 2,2 e-16	< 2,2 e-16
Rsem - Salmon	< 2,2 e-16	< 2,2 e-16
Kallisto - Sailfish	< 2,2 e-16	2,92 e-12
Kallisto - Salmon	< 2,2 e-16	0.02176
Sailfish - Salmon	< 2,2 e-16	3.448e-07

Tabla X: p-valores obtenidos en el test de Mann-Whitney para la igualdad de medias y del test de Kolmogorov-Smirnov para la distribución de la varianza.