UNIVERSIDAD NACIONAL DE CORDOBA

Análisis multivariado aplicado a la representación de datos sintéticos de secuenciación de ARN

Para optar al grado de: Magister en Estadística Aplicada

Pablo Daniel Reeb

Ingeniero Agrónomo

2017

Comisión asesora de tesis:

Director: Dr. Sergio J. Bramardi

Co-Director: Mg. Sc. Julio Di Rienzo

Fecha de aprobación de tesis: 15 de noviembre de 2017

A mis padres

Agradecimientos:

Concretar los estudios de la maestría significó un esfuerzo mucho mayor al que originalmente tenía pensado y sin duda pude culminarlos gracias al apoyo de muchas personas.

Gracias a Sergio Bramardi por ser un continuo apoyo académico y humano y a todos los profesores de la maestría que fueron ejemplos en mi formación estadística.

Gracias a mis amigos, compañeros de trabajo, compañeros de maestría y familia por alentarme y escucharme en este largo camino de la educación continua.

Palabras clave: análisis multivariado exploratorio, análisis factorial múltiple, simulación, plasmodios, datos genómicos, RNA-seq, medidas de disimilaridad

Resumen

La secuenciación de alto rendimiento de ARN genera grandes bases de datos con información que puede ser utilizada con diferentes objetivos. Una de las aplicaciones más utilizada consiste en resumir las lecturas de las secuencias agregándolas en función de una unidad de interés tal como gen, exón o transcript . En este tipo de análisis se obtienen matrices con datos de conteos correspondientes a cada individuo en estudio (filas) y asignados a una particular unidad de interés (columnas). En general el número de individuos es muy pequeño en relación al número de variables y los conteos presentan un rango de dispersión muy amplio. En esta tesis se comparan técnicas de análisis multivariado exploratorio a 2 y 3 vías de clasificación que contemplan la naturaleza de los datos obtenidos en experimentos de secuenciación de ARN. Utilizando datos sintéticos generados con la técnica de plasmodios se comparan transformaciones a los datos y medidas de disimilaridad empleadas en el análisis de cluster jerárquico, análisis de escalamiento multidimensional métrico y no métrico y en el análisis factorial multiple. La transformación de los conteos originales a través de funciones que utilizan logaritmo o el uso de disimilaridades basadas en correlacion de Spearman o disimilaridad Poisson rescata la estructura natural de las muestras en todos los métodos de análisis utilizados. La mera estandarización o normalización de los conteos no genera representaciones confiables. La elección de la mejor medida debe considerar el nivel de relación señal-ruido ya que no todas las medidas muestran la configuración natural de la muestras en función de la cantidad de transcripts expresados o no diferencialmente. Este aspecto debe considerarse al momento de representar las muestras utilizando todos transcripts obtenidos o filtrando por expresión diferencial.

Title: Multivariate analysis applied to the representation of synthetic RNA sequencing data

Key words: explorative multivariate analysis, multi factor analysis, simulation, plasmode, genomic data, RNA-seq, dissimilarity measures

Summary

High-throughput sequencing of RNA generates large databases with information that can be used for different purposes. One of the most common applications is to summarize the readings of the sequences by adding them according to a unit of interest such as gene, exon or transcript. In this type of analysis, the data matrices contain counts for each individual under study (rows) and assigned to a particular unit of interest (columns). In general, the number of individuals is very small in relation to the number of variables, and the counts have a very wide range of dispersion. In this thesis, exploratory two and three way multivariate analysis techniques that contemplate the nature of the data obtained in RNA sequencing experiments are compared. Using synthetic data generated by the plasmode technique, we compare the performance of specific data transformations and common dissimilarity measures in hierarchical cluster analysis, multidimensional metric and nonmetric scaling analysis, and in multiple factorial analysis. Either the transformation of the original counts through functions that use logarithm or the use of dissimilarities based on Spearman correlation or the Poisson dissimilarity rescues the natural structure of the samples in all the analysis methods. The mere standardization or normalization of counts does not generate reliable representations. The choice of the best measurement should consider the level of signal-to-noise ratio since not all measurements show the natural configuration of the samples depending on the number of differentially expressed and non-differentially expressed transcripts. This aspect must be considered when representing samples by using all the transcripts or by using only differentially expressed transcirpts.

CONTENIDOS

INDICE DE FIGURAS Y TABLAS

Capítulo 1: INTRODUCCION

La interpretación de datos genéticos ha constituido, desde la publicación de los experimentos fundacionales de Gregor Mendel en 1866, un estímulo para la propuesta y el desarrollo de teorías y técnicas estadísticas. Ambas ciencias, Genética y Estadística, evolucionaron conjuntamente durante el siglo XX y afrontan desde el comienzo del siglo XXI nuevos desafíos en el análisis de datos.

Las tecnologías de microarrays y la de secuenciación conocida como Next Generation Sequencing (NGS) (Metzker 2010), han constituido una revolución sin precedentes en el estudio de la constitución y expresión de los genes. En particular, la tecnología de NGS ha experimentado un desarrollo sin igual (Neafsey and Haas 2011) ya que permite la secuenciación completa de un genoma o el muestreo del transcriptoma de un individuo o una población de manera eficiente y económica. Tradicionalmente, los recursos han sido priorizados hacia organismos modelos, es decir, aquellos que son ampliamente estudiados por su facilidad de cría, porque ocupan una posición pivotante en la evolución o porque su genoma tiene características especiales (Twyman 2002). Actualmente, además de la profundización del estudio de los organismos modelos la implementación de técnicas de NGS y RNA-seq en particular, posibilita la expansión del estudio a organismos considerados no modelos. A su vez, estos estudios pueden llevarse a cabo en grupos de investigación de diversa escala por lo que es de prever que la expansión en la aplicación continuará en forma masiva. Estas características han servido para considerar que nos encontramos en un proceso de democratización de los estudios genéticos (Ekblom and Galindo 2011).

Dentro de las tecnologías de NGS, la secuenciación de ARN (RNA-seq) promete desplazar a otras técnicas en el estudio de la expresión génica (Z. Wang, Gerstein, and Snyder 2009; Mutz et al. 2012). Los datos generados por esta técnica, se enmarcan en el contexto de las grandes bases de datos con la particularidad de contar en el orden de miles de observaciones (genes)

correlacionadas, escasa cantidad de repeticiones (muestras o condiciones experimentales) y distribuciones de probabilidades no normales.

1.1. Secuenciación de ARN (RNA-seq)

La secuenciación de ARN conocida como RNA-seq es una tecnología desarrollada para el estudio del transcriptoma basada en la secuenciación de alto rendimiento de ADN (High Throughput Sequencing)(Olena Morozova, Hirst, and Marra 2009).

El transcriptoma es el conjunto de todas las transcripciones del ADN presentes en una célula en un determinado estadio de desarrollo o condición fisiológica. Su estudio es esencial para dilucidar el funcionamiento del genoma y comprender, por ejemplo, la respuesta de células y tejidos ante una enfermedad.

El proceso de secuenciación con RNA-seq incluye etapas de recolección y procesamiento de muestras, procesamiento informático y aplicación de herramientas de análisis según los objetivos del experimento (Alicia Oshlack, Robinson, and Young 2010). En la Figura 1 se presenta un esquema general de las etapas involucradas en la secuenciación, algunas variaciones pueden presentarse según la plataforma de análisis utilizada. En primera instancia se debe aislar el ARN de las muestras y acondicionarlo para someterlo a secuenciación (preparación de librerías). De este último proceso se obtienen millones de lecturas (reads) de secuencias cortas (35 a 300 pares de bases) de ADN complementario (cDNA) que deben someterse a un intenso procesamiento bioinformático. Finalmente, estas lecturas cortas de ARN pueden ser utilizadas para estudiar la expresión génica (Figura 5).

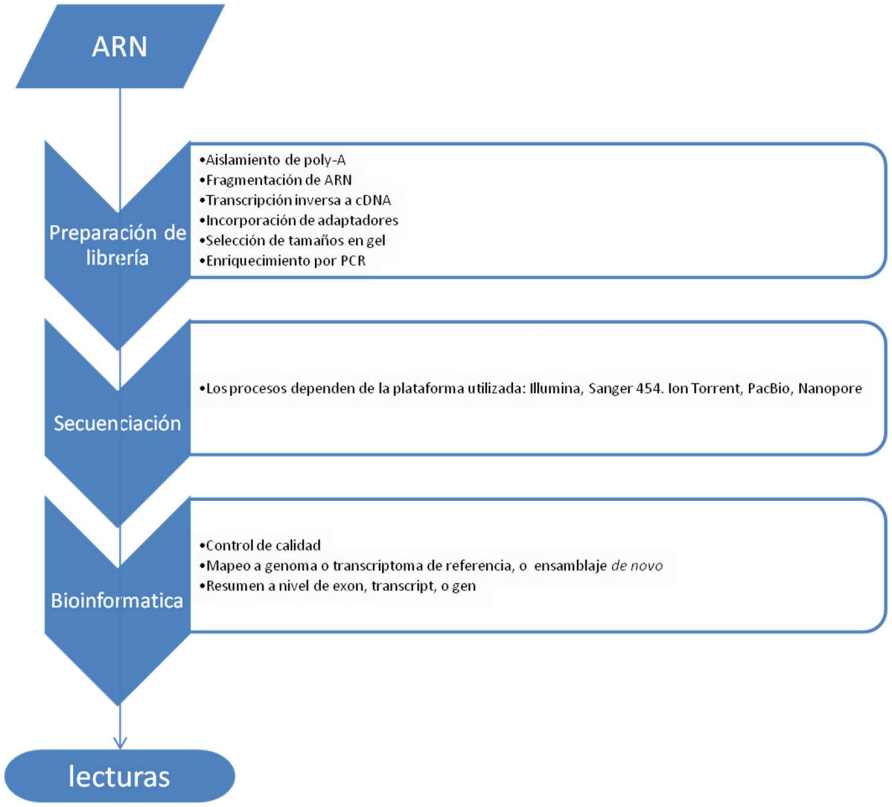

Figura 1. Esquema de las principales etapas en la secuenciación de ARN

1.1.1. Extracción de ARN

La obtención de una muestra confiable de ARN constituye un prerrequisito básico para cualquier análisis posterior. El ARN es sensible a enzimas endógenas y exógenas que pueden degradar rápidamente el ácido nucleico y alterar el perfil de expresión inmediatamente luego de la extracción de muestra (Adiconis et al. 2013). Existen diferentes protocolos para aislar, conservar y depurar adecuadamente el ARN (Camacho-Sanchez et al. 2013). La elección del método más conveniente depende de un balance entre disponibilidad de recursos económicos y de laboratorio, logística y objetivos del proyecto (Schwochow et al. 2012).

1.1.2. Preparación de librería

Una vez extraídas y aisladas, las moléculas de ARN deben someterse a una serie de procesos con el fin de acondicionarlas para su posterior secuenciación. Este conjunto de todas las moléculas de ARN preparadas para ser ingresadas a un secuenciador se conoce como

librería. La construcción de la librería ha sido reportada como una de las principales fuentes de error y sesgo en el análisis de ARN (Levin et al. 2010). Los pasos involucrados dependen de la tecnología de secuenciación y los objetivos del experimento (Zhong et al. 2011; Kumar et al. 2012; Raz et al. 2011) pero se puede describir algunos pasos comunes a todas las plataformas siguiendo el protocolo más utilizado (Parkhomchuk et al. 2009; Borodina, Adjaye, and Sultan 2011).

El proceso comienza con el aislamiento de la fracción poly-A (mRNA) del total de ARN y su fragmentación al azar. Posteriormente, por transcripción inversa se obtiene ADN complementario (cDNA) del que se elimina las colas de poly-A y se ligan adaptadores en los extremos. Finalmente, luego de filtrar las moléculas y seleccionar un tamaño específico, por ejemplo una longitud de 200 pares de bases, se utiliza PCR para amplificar la muestra. De esta manera, las moléculas de ARN son convertidas en fragmentos de cDNA listos para ingresar al secuenciador (Figura 2).

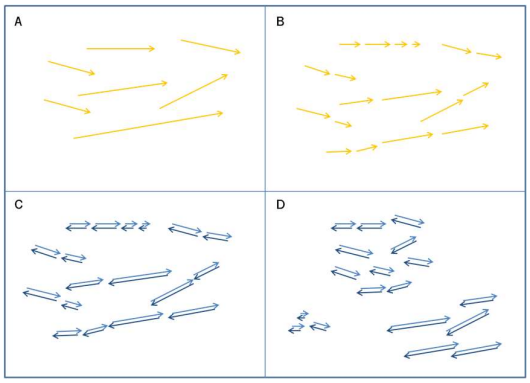

Figura 2. Etapas básicas de la preparación de una librería.

A: ARN de una muestra
B: ARN fragmentado al azar
C: Fragmentos de cDNA generados por transcripción inversa a partir de los fragmentos de ARN
D: Los fragmentos de cDNA son filtrados por tamaño para ingresar al secuenciador

1.1.3. Secuenciación

Las plataformas disponibles para secuenciar utilizando tecnología NGS son Illumina®, Roche 454®, IonTorrent ®, PacBio® y Nanopore®.

La tecnología base de cada plataforma es diferente y la preferencia por una u otra también está condicionada por el objetivo del experimento además del costo. Calidad de la

secuenciación, largo de las lecturas y profundidad de lectura de una librería son las principales características técnicas a considerar al momento de elegir una plataforma. Una comparación de las plataformas puede consultarse en (O Morozova and Marra 2008; Olena Morozova, Hirst, and Marra 2009) y, dados los rápidos avances en tecnología (Ozsolak and Milos 2011) y competencia entre proveedores, se recomienda actualizar las características de cada marca comercial en las respectivas páginas web (http://www.illumina.com , http://www.roche-applied-science.com , http://www.lifetechnologies.com , http://www.pacificbiosciences.com , http://www.nanoporetech.com).

Para ejemplificar, se describen los pasos (Figura 3) que utiliza la tecnología de mayor difusión (Van Verk et al. 2013) en la actualidad: Illumina (Illumina 2010). La librería de cDNA se coloca en uno de los ocho carriles de una celda de flujo. Cada fragmento de cDNA se adhiere a la superficie del carril y se somete a una etapa de amplificación en la que se convierten en clusters de doble hebra de ADN. Posteriormente, la celda de flujo se coloca en la máquina secuenciadora donde cada cluster es secuenciado en paralelo. Específicamente, en cada ciclo, se adicionan los cuatros nucleótidos con etiquetas fluorescentes y se mide la señal emitida por cada cluster. El proceso se repite un determinado número de ciclos, por ejemplo 100 veces, y la intensidad de la fluorescencia se transforma en lectura de bases. El número de ciclos determina la longitud de las lecturas y la cantidad de clusters determina el número total de lecturas.

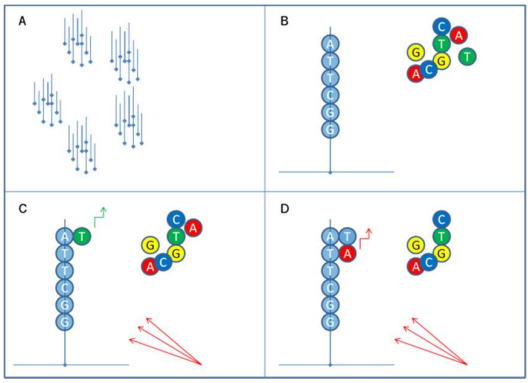

Figura 3. Etapas básicas de la tecnología de secuenciación con Illumina™

A: Se generan los clusters a partir de los fragmentos de cDNA, solo una copia se adhiere a la superficie. Esta es amplificada por un proceso puente, y finalmente queda adherida solo la copia forward.
B: se colocan las 4 bases fluorescentes que por afinidad se adhieren a la primera base
C: un láser excita las moléculas y la fluorescencia es captada, se registra así la primera base
D: el ciclo se repite n veces según la longitud deseada de las lecturas. Los fragmentos se pueden invertir para leer el otro extremo.

Independientemente de la tecnología utilizada, el resultado de la secuenciación son millones de lecturas de secuencias cortas de ADN (generalmente entre 35 y 300 pares de bases) leídos desde uno (lecturas simples) o ambos (lecturas pareadas) extremos de los fragmentos de cDNA (Figura 4).

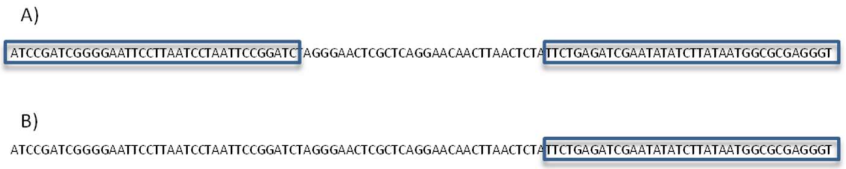

Figura 4. Fragmento de cDNA y lecturas desde ambos (A) o uno (B) de los extremos.

1.1.4. Bioinformática

Las lecturas generadas en el secuenciador deben someterse a un intensivo procesamiento informático, generalmente en una computadora o centro de alta performance, debido a los requerimientos de memoria RAM y espacio de almacenamiento permanente.

Las procedimientos a los que son sometidas las lecturas incluye (Alicia Oshlack, Robinson, and Young 2010; Givan, Bottoms, and Spollen 2012) : 1) control de calidad, 2) mapeo contra un genoma o transcriptoma de referencia, o ensamblaje *de novo* en caso de no contar con un genoma de referencia, 3) agregación a nivel de gen, exón o transcript según los objetivos del experimento.

Control de calidad: cada base de una lectura tiene asociado un parámetro que mide la calidad de su medición en el proceso de secuenciación. Se calcula utilizando el algoritmo Phred (Ewing et al. 1998) que predice la probabilidad de determinar incorrectamente una determinada base (Ewing and Green 1998) expresada en escala logarítmica (Q_{phred}=-10 $\log_{10}$ P(error)). Utilizando este parámetro de calidad se pueden filtrar las lecturas o segmentos de las mismas que sean poco confiables. Por ejemplo, una práctica común consiste en eliminar aquellas bases con score menor a 20, es decir bases determinadas con probabilidad de error mayor a 0.01 (1 en 100). Además se realiza eliminación de etiquetas de codificación (barcoding) que puedan

haberse anexado al cDNA para identificar muestras al preparar la librería y se controla la longitud de las lecturas.

Mapeo o alineación: el objetivo de esta etapa es encontrar la mejor posición en la que la secuencia de una lectura pueda atribuirse al genoma o transcriptoma de referencia (Langmead and Salzberg 2012; C Trapnell, Pachter, and Salzberg 2009) .

En caso de no contar con una referencia, es posible alinear las lecturas a un transcriptoma generado por la misma u otras muestras. Este mecanismo se conoce como ensamblaje *de novo* (Martin and Wang 2011; Grabherr et al. 2011)

Agregación: en un paso siguiente las lecturas pueden resumirse o agregarse en función de una unidad biológica de interés tal como gen, exón o transcript (C Trapnell, Pachter, and Salzberg 2009; Simon Anders 2013)

1.1.5. Lecturas

Las lecturas obtenidas con este primer procesamiento están disponibles para extraer información y responder a las preguntas de interés de la investigación científica. Las áreas más comunes de investigación (Figura 5) incluyen la determinación de expresión diferencial de alelos (Skelly et al. 2011; Fontanillas et al. 2010), estimación de abundancia de genes y de formas alternativas (Blekhman et al. 2010; Simon Anders, Reyes, and Huber 2012), edición de ARN (Hassan and Saeij 2014) y descubrimiento de nuevos transcripts (Cole Trapnell et al. 2010). Sin embargo, la aplicación más frecuente es el estudio de perfiles de expresión génica entre muestras (Alicia Oshlack, Robinson, and Young 2010).

El análisis de expresión diferencial génica es utilizado para establecer los genes que se encuentran diferencialmente expresados en muestras de distinto origen o sometidas a distintos tratamientos (por ejemplo, células normales versus células de tumores, o células hepáticas sometidas a diversas dosis de insecticida). Posteriormente, los resultados se integran con otros estudios con el fin de entender los procesos metabólicos involucrados.

El estudio de la expresión diferencial se basa en comparar los conteos que, según la agregación deseada de las lecturas, se obtienen para cada gen, exón o transcript. En esta tesis se aborda la exploración y representación de este tipo de datos.

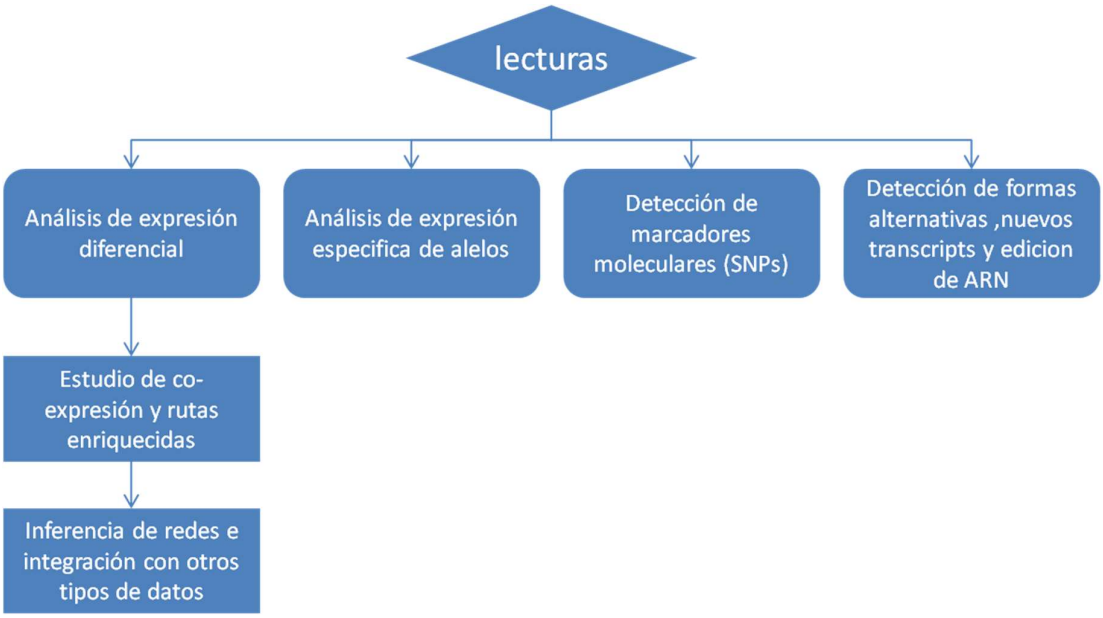

Figura 5. Tipos de estudios que pueden realizarse utilizando RNA-seq

1.1.6. Otras técnicas de análisis del transcriptoma

Otras tecnologías han sido desarrolladas para dilucidar y cuantificar el transcriptoma incluyendo aquellas que utilizan hibridación (microarrays) u otras tecnologías de secuenciación de bases (Sanger, serial analysis of gene expression (SAGE), cap analysis of gene expression (CAGE) y massively parallel signature sequencing (MPSS)). La tecnología de microarrays requiere de un genoma de referencia y el rango de detección está limitado por la saturación de las señales. Además, la comparación de los niveles de expresión entre experimentos requiere de métodos complicados de normalización. Otras tecnologías de secuenciación tienen limitaciones como la imposibilidad de distinguir entre formas alternativas de un gen y están basadas en procedimientos generalmente más costosos. Estas y otras limitaciones han sido

superadas por la secuenciación a través de RNA-seq (Z. Wang, Gerstein, and Snyder 2009) y la han posicionado como la tecnología preferida para el estudio del transcriptoma.

Desde el punto de vista de la expresión génica, RNA-seq no tiene límites en la detección, y provee una cuantificación que se correlaciona con el número de lecturas obtenidas. Esta medición directa del nivel de expresión de un gen ha hecho que se conozca a RNA-seq como una tecnología "digital" en contraste con la cuantificación "analógica" que provee la intensidad de señal en un microarray (Fang, Martin, and Wang 2012).

1.2. Características de los datos de RNA-seq

1.2.1. Naturaleza y dimensiones

Los datos generados por RNA-seq para el estudio del transcriptoma se pueden ordenar en una matriz de conteos en la que las filas están constituidas por muestras y las columnas por transcripts. La variable observada *X* corresponde al número de lecturas alineadas a cada transcript.

$$\begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & \ddots & & & \vdots \\ x_{i1} & & x_{ij} & & x_{in} \\ \vdots & & & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix}$$

donde: x_{ij}= número de lecturas en la *i*-ésima muestra asociadas al *j*-ésimo transcript, con *i*= i-ésima muestra *i*= 1, … , *n* y *j*= i-ésimo transcript *j*= 1, … , *p*

En análisis de datos de alta dimensionalidad es común encontrar la forma transpuesta de esta matriz, es decir usualmente tiene individuos en las columnas y variables en las filas.

Generalmente el número de muestras (*n*) es limitado y no sobrepasa las decenas en el mejor de los casos, mientras que el número de transcripts (*p*), ya sea genes o isoformas, se encuentra en el orden de los miles o decenas de miles ($p >> n$). El número de lecturas (x_{ij}) tiene un rango amplio que varía entre cero y miles.

Dada la naturaleza de la variable, diferentes distribuciones de probabilidades discretas ha sido utilizadas en modelos de análisis tales como multinomial (Nicolae et al. 2011), Poisson (Marioni et al. 2008) y binomial negativa (S Anders and Huber 2010; Robinson, McCarthy, and Smyth 2010), siendo esta ultima la más utilizada. Algunos estudios también han utilizado modelos Gausianos aplicados a la variable transformada (Langmead, Hansen, and Leek 2010).

Ya que las moléculas se muestrean aleatoriamente, la tecnología de RNA-seq no mide la abundancia absoluta de un transcript sino más bien la abundancia relativa del mismo (Pachter 2011).

1.2.2. Normalización

Debido a la presencia de fuentes de variabilidad sistemática propias de la tecnología de RNA-seq, una etapa reconocida en el procesamiento de los datos es la normalización (Bullard et al. 2010). Las fuentes incluyen variabilidad entre muestras como diferencias en el tamaño total de la librería (profundidad de secuenciación)(Mortazavi et al. 2008), o diferencias dentro de cada muestra tales como la longitud del gen (A Oshlack and Wakefield 2009) o el contenido en bases GC (Risso et al. 2011). Diferentes procedimientos han sido propuestos para corregir estas fuentes y la elección de un método puede condicionar los resultados del análisis estadístico posterior (Dillies et al. 2012).

1.3. Análisis estadístico

Diversos métodos estadísticos se aplican tanto en las etapas de pre- o pos-procesamiento de las lecturas. Los elementos de diseño experimental son especialmente útiles al momento de planificar los experimentos (Auer and Doerge 2010; J. Lee, Ang, and Xiao 2013). En la etapa de intensa utilización de herramientas bioinformáticas, son de particular importancia los métodos de control de calidad y modelos para medir la incertidumbre en la alineación o mapeo de las lecturas (Cole Trapnell et al. 2012; B. Li et al. 2010; Leng et al. 2013) . Posteriormente, los métodos estadísticos utilizados varían según los objetivos de estudio

(Figura 5). Aquí, tienen particular importancia los modelos estadísticos para normalizar datos (Bullard et al. 2010) y realizar inferencias (J. Li et al. 2012; Fang, Martin, and Wang 2012; Robinson and Smyth 2007; Van De Wiel et al. 2013). Las técnicas de análisis multivariado (AM) son también muy útiles en esta etapa fundamentalmente desde el punto de vista de exploración y representación de resultados. En esta tesis se desarrolla el uso de este último grupo de técnicas. La adaptación de métodos estadísticos aplicados a tecnologías como microarrays, al igual que el desarrollo de nuevos métodos específicos para RNA-seq es motivo de activa investigación en los últimos tres años. Nuevas herramientas bioinformáticas y metodologías de análisis estadístico, tales como técnicas de agrupamiento y análisis de redes, que consideren las propiedades de los datos de RNA-seq son necesarias para el estudio de expresión génica basada en secuenciación (Ma and Wang 2012).

1.3.1. Análisis exploratorio y data mining

El análisis exploratorio de datos (Tukey 1977) constituye un enfoque especialmente útil para abordar el desarrollo de técnicas estadísticas adaptadas a nuevos conjuntos de datos. Este abordaje permite resumir las principales características sin estar restringidos a plantear un modelo o hipótesis específicas, e incentiva la formulación de hipótesis, evaluación de supuestos para justificar métodos inferenciales y provee bases para planificar la recolección de nuevos datos. Muchos de los métodos son esencialmente gráficos y han sido adoptados y adaptados por la minería de datos al contexto de alta dimensionalidad (Izenman 2008).

La minería de datos es una metodología multidisciplinaria para extraer conocimiento de los datos. Es un proceso iterativo que analiza grandes bases de datos para generar modelos descriptivos y predictivos a partir de tendencias y patrones previamente desconocidos (Z. A. Zhou and Liu 2011). Constituye uno de los pasos del proceso de Descubrimiento de Conocimiento en Bases de Datos (KDD: Knowledge Discovery in Databases) propuesto por Fayyad et al. (1996) que es utilizado en varias disciplinas y ha tenido una aplicación creciente en bioinformática (Dua and Chowriappa 2013). Dentro de los métodos de aprendizaje supervisado y no supervisado que se utilizan en la minería de datos (Dua and Chowriappa

2013; Hastie, Tibshirani, and Friedman 2009) son particularmente útiles las técnicas de análisis multivariado.

1.3.2. Análisis Multivariado

Los datos multivariados consisten en múltiples medidas, observaciones o respuestas obtenidas para una colección particular de variables. En la última década, los avances en tecnología para la recolección y almacenamiento de datos han hecho que las bases de datos multivariados de grandes dimensiones constituyan una regla más que una excepción (Izenman 2008). La estructura de los datos generados por la secuenciación de ARN se ubica en este contexto (ver 1.2.1).

En el marco expuesto en el punto anterior (ver 1.3.1) los métodos de análisis de datos multivariados que se aplican tienen como objetivo la clasificación, agrupamiento y visualización de individuos o variables.

1.3.2.1. Desafíos de los datos de grandes dimensiones

El análisis multivariado para datos de grandes dimensiones en las que el número de variables (p) es mucho mayor al número de observaciones (n), $p >> n$, requiere modificaciones o procedimientos totalmente nuevos a los aplicados en un escenario con $n > p$ (Hastie, Tibshirani, and Friedman 2009). En particular, muchas de las variables pueden ser irrelevantes o redundantes y por lo tanto afectar negativamente los algoritmos de cálculo y los resultados de las diferentes técnicas estadísticas. Por ejemplo se puede desvirtuar la similaridad a través de la incorporación de variables espurias o resultar en una estimación no eficiente de una matriz de covarianzas.

Por esta razón, se reconoce que la aplicación de cualquier técnica de análisis multivariado debe estar precedida de una selección de variables (Izenman 2008; Hastie, Tibshirani, and Friedman 2009). Este procedimiento, puede ejecutarse a través de la eliminación de las variables irrelevantes o redundantes (selección de atributos), o por obtención de un nuevo conjunto de variables que son combinación de las originales (extracción de atributos) (Dua and Chowriappa

2013; Z. A. Zhou and Liu 2011). En los estudios de expresión génica se han aplicado técnicas de selección de atributos (Romero 2009); sin embargo, el procedimiento más utilizado en los estudios que aplican micromatrices o RNA-seq , es el empleo de análisis multivariado utilizando aquellas variables que resultan estadísticamente significativas en el estudio de expresión diferencial (Hastie, Tibshirani, and Friedman 2009) .

1.3.2.2. Naturaleza discreta y relaciones no lineales

La naturaleza discreta constituye otro desafío en el análisis multivariado (Bishop, Fienberg, and Holland 2007) de datos de secuenciación de ARN. La relación lineal media-varianza que modelan las técnicas tradicionales no se cumple en este tipo de datos. Por ejemplo, la representación de la matriz de varianzas-covarianzas o de correlación utilizando Análisis de Componentes Principales (ACP) puede no reflejar las verdaderas relaciones entre variables. Otras medidas de asociación deben ser adaptadas para captar relaciones no lineales en grandes bases de datos (Speed 2011; Reshef et al. 2011). La metodología más frecuente recurre a la transformación logarítmica o raíz cuadrada de las variables, pero la transformación no siempre refleja la verdadera estructura de los datos (Witten 2011) y la adición de una constante para generar pseudoconteos que contemplen la presencia de ceros y heterocedasticidad debe ser cuidadosamente contemplada (Love, Huber, and Anders 2014; Paulson et al. 2013; Costea et al. 2014). Otras medidas de asociación han sido propuestas y deben ser validadas y adaptadas para reflejar los patrones de dispersión y las relaciones de coregulación entre genes (Ma and Wang 2012; Witten 2011).

1.3.2.3. Tipos de aplicación

Desde el punto de vista exploratorio las técnicas multivariadas contribuyen en 4 importantes etapas del análisis: control de calidad, exploración de relaciones, presentación de resultados, y comparación de plataformas o integración entre distintos tipos de estudios. Debe rescatarse que la posibilidad de visualizar en un espacio reducido las relaciones

multidimensionales, tanto entre individuos como entre variables, constituye el principal atractivo del AM en el estudio de los sistemas biológicos (Gehlenborg et al. 2010).

El análisis de conglomerados en todas sus formas (Izenman 2008; Dua and Chowriappa 2013) es la técnica más utilizada generalmente junto a heatmaps (Zapala and Schork 2012). Se aplica fundamentalmente para presentar resultados luego del análisis de expresión diferencial, aunque podría ser utilizado en otras etapas. Recientemente, se han propuesto modificaciones de las medidas de distancias para datos de RNA-seq (Ma and Wang 2012; Witten 2011). Numerosas investigaciones han propuesto la adaptación de los algoritmos de agrupamiento para datos de secuenciación (N. Wang et al. 2013; Rau, Celeux, and Maugis-rabusseau 2011; Wei et al. 2012; Witten 2011; Si 2011; Xu et al. 2013).

Las técnicas de ordenación o reducción de las dimensiones han tenido menos investigación y uso. Se puede mencionar el uso de escalamiento multidimensional (Robinson, McCarthy, and Smyth 2010), la descomposición en valores singulares Poisson (S. Lee et al. 2013) y el uso de análisis de componentes principales sobre las variables transformadas (Love, Anders, and Huber 2014). El desarrollo de estas técnicas podrían ser de especial interés para complementar el estudio de relaciones entre variables (genes) y la integración con otras técnicas multivariadas como análisis de conglomerados o análisis factorial múltiple (Escofier and Pagès 1994).

Otro conjunto de técnicas multivariadas de especial interés son aquellas que posibilitan la integración de resultados y comparación entre estudios complementarios (Nie et al. 2008). El análisis de correlaciones canónicas ha sido propuesto como una alternativa para indagar la co-expresión de redes de genes (Hong et al. 2013) en un experimento. La integración de resultados de expresión génica con, por ejemplo, la abundancia de proteínas (proteomics) u otros metabolitos puede explicar de una manera más acabada la manifestación de un particular fenotipo. Si bien hay numerosos estudios que integran la expresión génica estudiada mediante micromatrices, éstas técnicas deben ser validadas o adaptadas para incorporar la expresión génica obtenida por secuenciación de ARN. Entre las técnicas disponibles, podemos mencionar el uso de análisis factorial múltiple y otros análisis multivariados a tres vías (de

Tayrac et al. 2009), el análisis de bosques aleatorios para el estudio de metagenomas (Touw et al. 2012; Dinsdale et al. 2013) y otras técnicas de machine learning (Dua and Chowriappa 2013).

El análisis biplot para modelos lineales generalizados también constituye una herramienta potencialmente útil (Greenacre 2010).

Objetivo general

Realizar un estudio comparativo de técnicas de análisis multivariado exploratorio para la representación de resultados de experimentos de secuenciación de ARN mensajero.

Objetivos específicos

1. Evaluar medidas de disimilaridad aplicadas a técnicas de agrupamiento no supervisado
2. Evaluar medidas de disimilaridad aplicadas a métodos de ordenación a dos vías
3. Aplicar métodos de análisis multivariado a tres vías en función de medidas de disimilaridad apropiadas

Capítulo 2: ANALISIS MULTIVARIADO

2.1. Conceptos generales. Clasificación.

El análisis multivariado engloba un conjunto de numerosas técnicas que pueden ser aplicadas con diversos objetivos. No existe una única forma de clasificar a todas las técnicas de análisis multivariado y tradicionalmente se las ha dividido en función del alcance (descriptivo o inferencial) y/u objetivos (clasificación u ordenación) (Johnson and Wichern 2002).

El análisis de datos genómicos, entre ellos los datos de secuenciación de ARN, se encuadran en el contexto de datos de alta dimensionalidad por lo tanto es pertinente enmarcar la aplicación de análisis multivariado en el contexto de una clasificación moderna. La necesidad de crear nuevos métodos y terminología para analizar conjuntos de datos complejos y de alta dimensionalidad ha llevado a que investigadores de muchas disciplinas (estadística, redes neuronales, análisis simbólico, computación e inteligencia artificial) trabajen en conjunto en la definición de nuevas teorías de análisis. En estadística, muchas técnicas de interés, tales como estimación de densidades, regresión, redes neuronales, análisis discriminante, árboles de clasificación, bosques aleatorios, análisis de cluster y métodos de reducción de la dimensionalidad, aplicadas particularmente a datos complejos y de alta dimensionalidad, actualmente se refieren en forma global como aprendizaje estadístico (Izenman 2008; Hastie, Tibshirani, and Friedman 2009; James et al. 2013). Vapnik (1998) relaciona estadística con la teoría de aprendizaje de la siguiente manera:

> *"El problema de aprendizaje es tan general que casi toda cuestión que ha sido discutida en la ciencia estadística tiene su analogía en la teoría del aprendizaje. Aún más, algunos resultados generales importantes se fundaron*

primero en el marco de la teoría de aprendizaje y luego se formularon en términos estadísticos."

Esta teoría es compatible con el enfoque propuesto en el capítulo anterior (ver 1.3) por lo que se considera la clasificación de las técnicas de análisis multivariado en i) aprendizaje supervisado y ii) aprendizaje no supervisado. La primer categoría incluye las técnicas con enfoque predictivo cuyos algoritmos tienen un conjunto de variables de entrada y otro/s de resultados de referencia que es observado o provisto por un "supervisor", se incluyen aquí técnicas de regresión y clasificación. La segunda categoría tiene objetivos descriptivos y no existe un conjunto de referencia explicito, comprende técnicas como análisis de conglomerados y representación de mapas de proximidades. Existen algunas técnicas que pueden ser catalogadas en ambas categorías según la aplicación, tal como es el caso del análisis de componentes principales que puede ser utilizado como técnica de reducción de las dimensiones pero también puede ser utilizado como una técnica de regresión.

A continuación se presenta la teoría estadística de técnicas de análisis multivariado que se encuentran en la categoría de aprendizaje no supervisado donde el objetivo es descubrir estructuras en las bases de datos, relaciones entre variables, tendencias y agrupamiento de individuos y presentar resultados.

Además se dividen según el número de vías o modos en dos grupos: a) técnicas de AM a 2 vías, y b) técnicas de AM a 3 vías. El primer grupo se utiliza para estudiar matrices de datos con *n* individuos y *p* variables, incluye técnicas tradicionales (Johnson and Wichern 2002) tales como análisis de componentes principales, análisis factorial de correspondencias, análisis de coordenadas principales, análisis de escalamiento multidimensional y análisis de conglomerados. El segundo grupo se emplea para estudiar matrices con *n* individuos, *p* variables y *q* condiciones, incluye métodos como análisis factorial múltiple (Escofier and Pagès 1994), meta biplot (Martín-Rodríguez, Galindo-Villardón, and Vicente-Villardón 2002), análisis de componentes principales a tres modos (Tucker 1966), escalamiento multidimensional a tres vías (McGee 1968), y análisis de procrustes generalizado (Gower 1975), entre otros.

2.2. Análisis multivariado a dos vías.

2.2.1. Análisis de agrupamientos o conglomerados (AC)

El análisis de agrupamientos o análisis de cluster, también llamado análisis de segmentación en minería de datos y descubrimiento de clases en machine learning, es la técnica multivariada más conocida y empleada de aprendizaje no supervisado. Su principal objetivo es agrupar un conjunto de objetos en subconjuntos o "clusters" de manera tal que aquellos objetos que pertenecen a un mismo cluster estén más relacionados entre sí que los objetos asignados a clusters diferentes. Adicionalmente, otro objetivo puede incluir el arreglo de los clusters en una jerarquía natural tal que sucesivos agrupamientos de clusters entre sí en cada nivel de jerarquía, agrupe los clusters que más se asemejen. No existe garantía de que los algoritmos encuentren más de un grupo, sin embargo, en cualquier aplicación práctica, la hipótesis subyacente es que los datos forman un conjunto heterogéneo que puede ser separado naturalmente en grupos lógicos para el dominio de estudio del investigador. Es oportuno recordar que las técnicas de análisis de conglomerados deben ser utilizadas como herramientas exploratorias para descubrir patrones en los datos o para complementar la visualización de resultados acompañando a otras técnicas confirmatorias.

Muchos de los resultados del AC son de naturaleza visual y se muestran como diagramas de dispersión, árboles, dendrogramas, gráficos de siluetas o heatmaps.

En los estudios de expresión génica, puede ser de interés agrupar genes, muestras o ambos a la vez (bi-clustering). El objetivo de agrupar genes radica en identificar aquellos que tienden a comportarse de manera similar en los individuos estudiados, mientras que también puede ser de interés estudiar el agrupamiento de muestras basados en patrones de expresión génica, o identificar simultáneamente qué grupos de genes está más relacionado a qué grupo de muestras.

2.2.1.1. Medidas de disimilaridad y similaridad

Los métodos para agrupar objetos (observaciones o variables) dependen de cuan semejantes o diferentes son los objetos entre sí, por lo tanto una paso importante consiste en establecer la mejor medida para representar la similaridad, o inversamente la distancia o disimilaridad, entre objetos. La medida depende de la naturaleza cuali o cuantitativa de las variables en estudio. Dado que la expresión de los genes se determina en términos cuantitativos, se presentan las medidas de similaridad más utilizadas en estos experimentos. Sean $x_s = (x_{s1}, \dots, x_{sr})$ y $x_t = (x_{t1}, \dots, x_{tr})$ dos puntos en $\Re^r$. Entonces, se puede definir las disimilaridades y similaridades entre los puntos tal como se expresa en las Tabla 1 y Tabla 2, respectivamente. Considere que x_s y x_t pueden ser dos genes que se encuentran en *r* muestras, o dos muestras en las que se determinaron *r* genes.

Distancia	$d(x_s, x_t) =$
Euclídea	$\sqrt{\sum_{h=1}^{r} (x_{sh} - x_{th})^2}$
Manhattan (City-block)	$\sum_{h=1}^{r} \lvert x_{sh} - x_{th} \rvert$
Canberra	$\sum_{h=1}^{r} \frac{\lvert x_{sh} - x_{th} \rvert}{\lvert x_{sh} \rvert + \lvert x_{th} \rvert}$
Minkowski	$\left(\sum_{h=1}^{r} \lvert x_{sh} - x_{th} \rvert^m \right)^{1/m}$
Chi-cuadrado	$\sum_{h=1}^{r} \frac{1}{x_{h.}} \left(\frac{x_{sh}}{x_{.s}} - \frac{x_{th}}{x_{.t}} \right)^2$

Tabla 1. Medidas de disimilaridad

Las disimilaridades de la Tabla 1 están relacionadas y la familia de métricas de Minkowski incluye a la Euclídea (m=2) y Manhattan (m=1). La métrica Euclídea es muy utilizada pues brinda una interpretación geométrica de distancia pero es sensible a la presencia de outliers. La métrica de Manhattan es menos sensible a la presencia de datos raros pero ambas deben considerarse como poco robustas ante outliers. La disimilaridad de Canberra esta escalada y su rango va de 0 a 1. La distancia chi-cuadrado entre perfiles columna (variables) estandariza la distancia Euclídea por los marginales columna y pondera en función de los marginales fila. Similarmente se puede calcular la distancia entre perfiles fila (observaciones). Es una distancia

muy utilizada para datos de frecuencia y constituya la base de la técnica de análisis de correspondencias (François Husson, Le, and Pages 2011).

Aplicando una función monótona decreciente apropiada, es posible convertir medidas de similaridad en medidas de disimilaridad. De esta forma, se puede calcular 1 menos cualquiera de las correlaciones de la Tabla 2 y obtener las respectivas medidas de disimilaridad.

Similaridad	$sim(x_s, x_t) =$
Correlación de Pearson	$\frac{\sum_{h=1}^{r}(x_{sh}-\bar{x_s})(x_{th}-\bar{x_t})}{\sqrt{\sum_{h=1}^{r}(x_{sh}-\bar{x_s})^2 \sum_{h=1}^{r}(x_{th}-\bar{x_t})^2}}$
Correlación no centrada	$\frac{\sum_{h=1}^{r} x_{sh}x_{th}}{\sqrt{\sum_{h=1}^{r} {x_{sh}}^2 \sum_{h=1}^{r} {x_{th}}^2}}$
Correlación de Spearman	$\frac{\sum_{h=1}^{r}(rg(x_{sh})-\overline{rg})(rg(x_{th})-\overline{rg})}{\sqrt{\sum_{h=1}^{r}(rg(x_{sh})-\overline{rg})^2\,(rg(x_{th})-\overline{rg})}}$

Tabla 2. Medidas de similaridad

La correlación de Pearson es muy utilizada y es equivalente a utilizar la distancia Euclidea cuadrada si los datos han sido previamente estandarizados por variables:

$$\frac{1}{2(r-1)}\sum_{h=1}^{r}(x_{sh}-x_{th})^2 = 1-\rho_{st} \in [0,2]$$

donde ρ_{st} es la correlación entre x_s y x_t .

Esta medida es particularmente adecuada si x_s y x_t tiene relaciones fuertemente lineales pero es sensible a outliers. Es la medida preferida cuando el objetivo es agrupar variables. Muchos investigadores utilizan el valor absoluto de la correlación basándose en que tanto la correlación negativa como positiva implican similaridad, mientras que valores cercanos a cero indican que no hay correlación.

La correlación de Pearson no centrada difiere en que considera la magnitud de los valores observados.

La correlación de Spearman considera el ranking u orden de los datos (*rg()*) en lugar de los valores observados. Es muy robusta a la presencia de outliers y a desviaciones de relaciones lineales. Si la relación ente x_s y x_t es aproximadamente lineal, la correlación de Pearson y

Spearman producirán resultados similares. Algunos investigadores también prefieren emplear el valor absoluto de la correlación de Spearman.

En general, es más importante encontrar patrones de expresión génica similares sin importar los niveles de expresión, por lo que la correlación de Pearson y Spearman (o sus valores absolutos) pueden considerarse más útiles que las distancia tipo o la correlación no centrada.

2.2.1.2. Medidas de disimilaridad para RNA-seq

Además de las medidas mencionadas en la sección anterior, recientemente se han propuesta las siguientes medidas para datos de RNA-seq:

Poisson

Esta disimilaridad se basa en calcular la distancia utilizando un cociente de verosimilitud modificado para un modelo log-lineal Poisson (Witten 2011). Considérese el siguiente modelo:

$$X_{ij} \sim Poisson\left(N_{ij} d_{ij}\right),\ X_{i'j} \sim Poisson\left(N_{i'j} d_{i'j}\right)$$

$$N_{ij} = s_i g_j\ ,\ N_{i'j} = s_{i'} g_j$$

donde: X_{ij} es el conteo correspondiente a la *i*-ésima observación del *j*-ésimo transcript, con *i= 1* a *n*, y *j=1* a *p. s*$_i$=número total de lecturas por muestra, *g*$_i$= número total de lecturas por transcript.

Para la hipótesis nula H_0: $d_{ij} = d_{i'j} = 1$, *j=1,…,p*, el logaritmo del cociente de verosimilitud puede ser usado como medida de distancia adicionando una constante para evitar calcular

$\widehat{d_{ij}} = 0\ si\ X_{ij} = 0$

$\widehat{d_{ij}} = \frac{X_{ij}+\beta}{\widehat{N_{ij}}+\beta}$, $\widehat{d_{i'j}} = \frac{X_{i'j}+\beta}{\widehat{N_{i'j}}+\beta}$

La disimilaridad entre dos objetos puede ser medida como:

$$\sum_{j=1}^{p}\left(\widehat{N_{ij}} + \widehat{N_{i'j}} - \widehat{N_{ij}}\widehat{d_{ij}} - \widehat{N_{i'j}}\widehat{d_{i'j}} + X_{ij} log\widehat{d_{ij}} + X_{i'j} log\widehat{d_{i'j}}\right)$$

1-Correlacion de Gini (GCC)

Esta correlación es ampliamente utilizada por ejemplo en economía (Yitzhaki and Schechtman 2013) y ha sido propuesta recientemente para análisis del transcriptoma (Ma and Wang 2012). Tal como otras medidas de correlación su rango está acotado en el intervalo [-1,1]. El valor cero indica independencia entre dos variables, mientras que los valores extremos indican una relación monotónicamente decreciente o creciente. A diferencia de las correlaciones de Pearson, Spearman o Kendal considera simultáneamente la información referida al rango y al valor de las variables. La correlación de Gini es más robusta sobre datos de distribución no Gaussiana y también a la presencia de outliers.

La correlación de Gini utiliza recíprocamente la información del valor de una variable y la información del rango de la otra variable, por lo que para un par de variables (*X, Y*), por ejemplo un par de transcripts, se pueden utilizar las siguientes formulas:

$$GCC(X,Y) = \sum_{i=1}^{n}(2i - n - 1)x(i,Y) \,/\, \sum_{i=1}^{n}(2i - n - 1)x(i,X)$$

donde n es el tamaño de muestra (por ejemplo el número de transcripts), x(i,X) es el i-ésimo valor para el perfil de valores de X ordenados en forma creciente y x(i,Y) es el valor correspondiente de X en el par de transcripts (X,Y) para el i-ésimo valor del perfil de Y ordenados en forma creciente.

En forma similar se puede definir a GCC como:

$$GCC(Y,X) = \sum_{i=1}^{n}(2i - n - 1)y(i,X) \,/\, \sum_{i=1}^{n}(2i - n - 1)y(i,Y)$$

2.2.1.3. Propiedades

La mayoría de las disimilaridades satisfacen las siguientes primeras tres propiedades:

1. $d(x_s, x_t) \geq 0$; “distancia no negativa”
2. $d(x_s, x_s) = 0$; “distancia 0 entre objetos iguales”
3. $d(x_s, x_t) = d(x_t, x_s)$. “distancia simétrica entre objetos”
4. $d(x_s, x_t) \leq d(x_s, x_u) + d(x_u, x_t)$. “distancia entre dos objetos no es mayor a la suma de las distancias entre esos objetos y un tercero: desigualdad triangular”

5. $d(x_s, x_t) \leq \max\{d(x_s, x_u), d(x_u, x_t)\}$

Si además satisface la propiedad 4 se consideran métricas, mientras que si satisface la propiedad 5 se consideran ultramétricas.

Las disimilaridades ultramétrica pueden ser representadas sin distorsión en un dendrograma. Si bien estas propiedades son importantes en contextos analíticos, en la práctica se aceptan como válidas medidas que no satisfacen están propiedades. Por ejemplo, 1 menos las similaridades de la Tabla 2 son ampliamente usada aunque no satisfacen las propiedades 1 y 4.

2.2.1.4. Matrices de disimilaridad o proximidad

Dadas *n* observaciones, $x_1, \ldots, x_n \in \Re^r$, el punto de inicio de los algoritmos de agrupamiento es calcular los pares de disimilaridades entre observaciones y ordenarlos en una matriz simétrica (*n* x *n*), $\mathbf{D} = (d_{st})$, donde $d_{st} = d(x_s, x_t)$, con ceros en la diagonal. Si por el contrario se utiliza una correlación entre variables, la matriz $\mathbf{D} = (d_{st})$, es simétrica de dimensión (*p* x *p*), donde la *st*-ésima disimilaridad es $d_{st} = 1 - \rho_{st}$.

2.2.1.5. Métodos jerárquicos

Los métodos de AC jerárquicos proporcionan una visualización relativamente concisa de los estudios de expresión génica, permitiendo ver qué genes o muestras tienden a agruparse. De allí que sean los más populares en genética, aunque otras métodos de AC pueden ser igualmente útiles (Garrett, Irizarry, and Zeger 2003).

Hay dos paradigmas en la construcción de agrupamientos jerárquicos: aglomeración (agnes) o división (diana). La estrategia más popular, aglomerativa o de abajo hacia arriba, comienza con r clusters singulares y en sucesivos pasos agrupa el par de clusters más parecido en un nuevo cluster hasta obtener el paso r-1 un solo grupo con todas los entidades. Contrariamente, la estratégica divisiva, o de arriba hacia abajo, inicia con todas las entidades y en cada paso separa uno de los clusters existentes en dos nuevos clusters de tal forma que tengan la mayor disimilaridad entre grupos.

En ambos paradigmas se obtienen r-1 niveles de jerarquía donde cada nivel representa un particular agrupamiento en particiones disjuntas. El investigador debe decidir qué nivel, si es que existe alguno, representa el agrupamiento natural en el sentido de que las observaciones dentro de un grupo son suficientemente más parecidas entre sí que las observaciones asignadas a grupos distintos de esa jerarquía.

2.2.1.6. Métodos aglomerativos

A fin de agrupar los clusters en cada nivel de jerarquía, debe definirse una medida de disimilaridad entre clusters. Sean *G* y *H* dos clusters, en la tabla 3 se definen las disimilaridades entre *G* y *H d(G,H)* del conjunto de disimilaridades d_{st} donde *s* pertenece a *G* y *t* pertenece a *H*. Estas medidas se conocen como métodos de encadenamiento siendo los tres más conocidos el método del mínimo, completo y de la media.

Método de Encadenamiento	$\boldsymbol{d(G,H)} =$
Mínimo o vecino más cercano	$\min\ (d_{st})$
Completo o vecino más lejano	$\max\ (d_{st})$
Media	$\frac{1}{N_G N_H} \sum_{s \in G} \sum_{t \in H} d_{st}$

Tabla 3. Métodos de encadenamiento en AC jerárquico aglomerativo

Si las disimilaridades entre objetos d_{st} presentan una tendencia de agrupamientos muy marcada, los tres métodos de encadenamiento presentaran resultados similares. Caso contrario, los métodos pueden presentar resultados diferentes y de allí la importancia de la elección del método de encadenamiento. El método del mínimo es espacio contractivo y suele generar dendrogramas encadenados que pueden no reflejar fielmente a los cluster naturalmente compactos, tiende a generar agrupamientos de gran diámetro. En el extremo opuesto, el método completo tiende a generar cluster compactos de menores diámetros. Sin embargo es posible que individuos asignados a un cluster se encuentren más cercanos a miembros de otros clusters que a miembros del propio cluster. El método de la media

representa un compromiso entre ambos extremos y puede producir clusters relativamente compactos y alejados entre sí. Sin embargo, sus resultados dependen de la escala numérica en la que se miden las disimilaridades. Si se aplica una transformación estrictamente monótona decreciente $h(\cdot)$ a d_{st} , $h_{st}=h(d_{st})$, se pueden cambiar los resultados producidos con encadenamiento de la media. Por el contrario, tanto el método del mínimo como el completo, dependen sólo del orden de d_{st} y por lo tanto es invariante a la transformación. Por esta razón, suelen preferirse ante métodos que utilizan la media.

Algoritmo de construcción de AC jerárquico aglomerativo:

1. Seleccionar una medida de disimilaridad o similaridad. Por ej. una de las presentadas en la Tabla 1 o Tabla 2.
2. Calcular los $\binom{r}{2} = r(r-1)/2$ pares de disimilaridades y tratar a cada observación como un cluster singular.
3. Seleccionar un método de encadenamiento para calcular las distancias entre clusters. Por ej. uno de los definidos en la Tabla 3.
4. Calcular las distancias entre clusters y unir aquellos que tienen la menor distancia. La disimilaridad entre estos dos clusters indica la altura en el dendrograma a la cual se fusionaron.
5. Calcular nuevamente las distancias entre los nuevos clusters.
6. Repetir los pasos 4 y 5 hasta obtener un solo agrupamiento. Total de r-1 iteraciones.

2.2.1.7. Dendrograma

El dendrograma representa los resultados de la aplicación de un particular algoritmo. Generalmente en el eje *Y* se presenta la medida de distancia y en eje *X* se ordenan los individuos. Realizando un corte horizontal a una altura particular del eje Y se particionan los individuos en cierta cantidad de clusters disjuntos.

La estructura de un dendrograma puede representarse a través de la distancia cofenética entre individuos que es la disimilaridad intergrupo a la que dos individuos *s* y *t* se unen en un mismo cluster. Si se compara con las r(r-1)/2 disimilaridades calculadas originalmente se obtiene una medida de bondad de la representación del dendrograma. Esta medida se denomina

correlación cofenéticas. Dado que las distancias cofenéticas obedecen a una propiedad ultramétrica, y a que la estructura natural de los datos puede no tener la misma cantidad de clusters que los propuestos por el dendrograma, es de esperar que la matriz cofenéticas no represente fielmente las disimilaridades originales. Por lo tanto, debe recordarse el carácter descriptivo de la técnica y valorar si es natural esperar la estructura impuesta a los datos analizados por el algoritmo.

2.2.1.8. Selección de muestras y variables

En el contexto de alta dimensionalidad donde el número de variables es muy superior al número de individuos resulta importante considerar el filtrado sobre todo de las variables o atributos, en general las muestras son escasas y se consideran representativas de la población. Para el caso de las variables, en genética se espera que la mayoría de los genes en un experimento no estén relacionados a la estructura de agrupamientos buscada por lo tanto la inclusión de esos genes introduce ruido al análisis. El análisis de cluster puede considerarse un tipo de análisis de variables latentes, donde la variable latente puede pensarse como la verdadera pertenencia a un grupo. Se asume que la matriz de correlación (o disimilaridad) definida provee la estructura de agrupamientos latente. Si en el análisis se incluyen variables, genes, que no están relacionados con la estructura de agrupamientos de interés, los resultados serán posiblemente diferentes a si se los excluye. Si bien no es posible saber a priori que variables están relacionadas con los grupos de interés, es posible realizar un filtrado de los genes que muestran evidencias de expresión diferencial (Garrett, Irizarry, and Zeger 2003), complementar con técnicas de análisis de componentes principales (Hastie, Tibshirani, and Friedman 2009), complementar con el uso de modelos (Izenman 2008) o recurrir al uso de filtros específicos en el contexto de data mining (Dua and Chowriappa 2013).

2.2.2. Análisis de componentes principales (ACP)

Como ya se mencionó, un camino para analizar datos de grandes dimensiones en un espacio menor al original es recurrir a la selección de variables (atributos), mientras que otra

posibilidad es la extracción de atributos. Este último enfoque consiste en obtener un número reducido de variables creadas por transformación de las variables originales y el ACP propuesto por Hotelling (1933) es la técnica más utilizada.

2.2.2.1. Obtención de componentes principales

Sea: $\mathbf{X} = (X_1, \dots, X_r)^T$ un vector aleatorio con media μ_X y matriz de varianza $\mathbf{\Sigma}_{XX}$ de dimensión (*p* x *p)*.

El ACP trata de reemplazar al conjunto de *p* variables originales correlacionadas $X_1, X_{2,\dots,} X_p$, por un conjunto de *p'* combinaciones lineales no correlacionadas $\xi_1, \xi_{2,\dots,} \xi_t$ $(p' < p)$, de las variables originales,

$\xi_j = \mathbf{b}_j^T \mathbf{X} = b_{j1} X_1 + \cdots + b_{jp} X_p$, con *j=1,2,…,p'*,

donde la pérdida de información, en términos de la variación total de las variables originales, sea mínima.

$$\sum_{j=1}^{p} var\left(X_j\right) = tr(\mathbf{\Sigma}_{XX})$$

Utilizando el teorema de descomposición espectral, se puede escribir

$$\mathbf{\Sigma}_{XX} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathrm{T}}, \mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}_p \,,$$

donde la matriz diagonal $\mathbf{\Lambda}$ tiene como elementos de su diagonal los autovalores, $\{\lambda_j\}$, de Σ_{XX}, y las columnas de **U** son los autovectores de $\mathbf{\Sigma}_{XX}$. Por lo tanto, la variación total es $tr(\mathbf{\Sigma}_{XX}) = tr(\mathbf{\Lambda}) = \sum_{j=1}^{p} \lambda_j$.

El *j*-ésimo vector de coeficientes $\mathbf{b}_j = b_{1j}, \cdots, b_{rj}$, es calculado de tal manera que:

- Las *p'* combinaciones lineales ξ_j, *j=1,2,…,p'*, de **X** están ordenadas en forma decreciente según la magnitud de sus varianzas: $var(\xi_1) \geq var(\xi_2) \geq \cdots \geq var\left(\xi_{p'}\right)$.
- ξ_j no está correlacionada con $\xi_{j'}, j' < j$. $(cov\left(\xi_j, \xi_{j'}\right) = \mathbf{b}_j^T \mathbf{\Sigma}_{XX} \mathbf{b}_{j'} = 0)$

Una única solución de $\{\xi_j\}$ se obtiene imponiendo una restricción de normalización tal que $\mathbf{b}_j^T\mathbf{b}_j = 1$, para todo *j=1,2,...,p'*.

Las combinaciones lineales constituyen las componentes principales de **X**, los vectores de coeficientes $\{\mathbf{b}_j\}$ se conocen como las cargas asociadas a cada componente.

2.2.2.2. Interpretación de las componentes principales

La primera pregunta que surge cuando se ha realizado un ACP es cuantas componentes principales se deben retener tal que la proyección de las observaciones en el nuevo subespacio preserven la mayor cantidad de información. Para determinar la contribución de las componentes principales a explicación de la variación total deben evaluarse los autovalores (λ_j) de la descomposición. Los métodos más difundidos son el test de scree plot (Cattell 1966), la regla de la traza del rango de las componentes principales () y la regla de Kaiser (Izenman 2008). Estas reglas u otras son útiles para determinar el número de CP cuando el objetivo es resumir los datos en un espacio reducido pero si el objetivo es obtener una visualización, generalmente se utilizan solo las 2 o 3 primeras componentes. Si en estos espacios no se encuentran patrones de interés, la inclusión de nuevas dimensiones difícilmente aportará más información. Por el contrario, si se encuentran relaciones interesantes, se adicionan nuevas componentes hasta que no se obtienen nuevos patrones.

Los coeficientes de los vectores de cargas proporcionan información sobre la importancia relativa de las variables originales en la formación de las componentes principales. A su vez, los vectores de cargas definen las direcciones en el espacio de las observaciones, donde estas tienen mayor variabilidad. Si se proyectan las observaciones originales en estas direcciones, se obtienen sus nuevas coordenadas (scores) en el sistema de componentes principales y es posible analizar por ejemplo las distancias entre individuos. A su vez se pueden representar los vectores de cargas y analizar geométricamente la importancia de las variables en la formación de las componentes principales y las correlaciones entre aquellas en el nuevo espacio como el

ángulo entre los vectores. La representación simultánea de las observaciones y variables se puede establecer en un biplot.

2.2.2.3. Invariancia y escalamiento

El ACP no es invariante al escalamiento de las variables originales. Si se centran y escalan las variables originales por su media y varianza:

$$\mathbf{Z} \leftarrow (diag\{\mathbf{\Sigma}_{XX}\})^{-1/2}(\mathbf{X} - \boldsymbol{\mu}_X)$$

los resultados que se obtienen son equivalentes a analizar la matriz de correlaciones en lugar de la matriz de covarianzas pero la falta de invariancia implica que los resultados obtenidos pueden ser muy distintos.

La estandarización tiene ventajas en situaciones en que las unidades de medida de las variables originales son muy heterogéneas o cuando los rangos de sus valores difieren considerablemente. En estos casos, si no se estandarizan las variables, aquellas con mayores varianzas tenderán a sobre evaluar su contribución en las componentes en detrimento de las restantes variables. La estandarización es muy oportuna cuando el ACP se utiliza con fines exploratorios aunque tiene restricciones en las propiedades distribucionales si se quiere utilizar con fines inferenciales.

2.2.2.4. Otros usos del ACP

Además del uso primario como técnica de reducción de la dimensionalidad, y como método de visualización de relaciones entre observaciones y variables, el ACP puede utilizarse con otros objetivos.

Las primeras componentes principales pueden revelar si los datos realmente se encuentran en un subespacio lineal de $\Re^r$ y puede ser utilizado para identificar outliers, particularidades distribucionales y agrupamiento de observaciones. Las ultimas componentes principales muestras las proyecciones lineales que tienen menor varianza y son por lo tanto virtualmente constantes y pueden ser usadas para detectar colinealidad y outliers que puedan alterar la dimensionalidad percibida de los datos.

Las componentes principales proveen el mejor resumen de los datos dado que proporcionan la recta o plano más cercanos a las observaciones en términos de distancia Euclídea cuadrada. Esta característica es explotada por otras técnicas estadísticas de regresión, clasificación y agrupamiento que utilizan los scores de las primeras $c << p$ componentes principales en lugar de utilizar la matriz completa de datos originales. Esta puede llevar a resultados más limpios de ruido, ya que generalmente la señal (como oposición al ruido) de los datos se concentra en las primeras componentes principales.

2.2.2.5. Otras técnicas de ACP

Sobre la base de la teoría del ACP se han propuesto modificaciones para aplicar ante cierto tipo de datos.

Por ejemplo, dado que el ACP es sensible a la presencia de outliers se han propuesto versiones robustas de ACP. En otras situaciones, la linealidad del ACP pude ser un obstáculo para reducir exitosamente los datos si estos no se encuentran en un subespacio lineal (o aproximadamente lineal). En esos casos puede ser útil explorar generalizaciones no lineales, como análisis de curvas o superficies no lineales, o colectores de aproximación no lineal, tales como ISOMAP, LLE (local linear embedding), o escalamiento multidimensional local (Hastie, Tibshirani, and Friedman 2009).

Por ejemplo, el ACP de núcleo (kernel) (Scholkopf, Smola, and Muller 1998) , expande la aplicación del ACP tradicional emulando lo que se obtendría si se expandieran las variables con transformaciones no lineales y luego se aplicara ACP a ese espacio de variables transformadas. Este tipo de análisis está relacionado al escalamiento multidimensional métrico si se aplica una función de núcleo isotrópica (Williams 2001).

2.2.3. Escalamiento multidimensional (EM)

Bajo el nombre de escalamiento multidimensional se conoce a una familia de algoritmos que dada una matriz de proximidad intentan representar en un espacio reducido a las entidades (individuos o variables) preservando de manera óptima las proximidades entre

ellas. Su principal uso es como método de visualización para identificar agrupamientos por cercanía entre objetos en la representación final.

Los métodos pueden clasificarse en: a) escalamiento clásico, también llamado geometría de distancia en bioinformática, y b) escalamiento de distancias que se divide según el tipo de disimilaridades en: i) métrico o ii) no métrico.

2.2.3.1. Matrices de proximidad

La medida para determinar la proximidad entre dos entidades tiene al igual que en el AC un rol fundamental en el EM. Para ejecutar un análisis de EM no se precisa tener las variables originales sino que basta con determinar las proximidades entre las entidades con una medida conveniente. En esta tesis utilizaremos las medidas de similaridad o disimilaridad presentadas en la Tabla 1 y Tabla 2 utilizando una transformación monótona decreciente para transformar las similaridades en disimilaridades.

Sea δ_{st} la disimilaridad entre las entidades *s*-ésima y *t*-ésima. Las *r* disimilaridades, $\{\delta_{st}\}$, pueden arreglarse en una matriz cuadrada (*r*x*r*), que suele representarse como la matriz triangular inferior $\triangle = (\delta_{st})$.

2.2.3.2. Algoritmos de calculo

Dadas *r* entidades y su matriz de disimilaridades $\triangle = (\delta_{st})$, el escalamiento multidimensional clásico encuentra una configuración de puntos en un espacio reducido tal que las distancias entre los puntos satisfagan $d_{st} \approx \delta_{st}$. El escalamiento de distancia, en cambio, encuentra una configuración para $d_{st} \approx f(\delta_{st})$, donde f es una función monótona que transforma las disimilaridades en distancias. Si las disimilaridades son cuantitativas (por ejemplo de razón o intervalo) se utiliza escalamiento de distancia métrico, mientras que si las disimilaridades son cualitativas (por ejemplo ordinales) se utiliza escalamiento de distancia no métrico. En la literatura generalmente se refiere al escalamiento de distancia métrico como EM métrico y al escalamiento de distancia no métrico como EM no métrico, aunque actualmente se tiende a usar el término EM como sinónimo de EM no métrico. Para aclarar otro término

utilizado en la literatura, se puede mencionar que el escalamiento clásico puede ser visto como un caso especial del escalamiento de distancia métrico que se ajusta por mínimos cuadrados, si la disimilaridad que se utiliza es la Euclídea y f es la identidad. De hecho, frecuentemente el termino escalamiento de distancia métrico es visto como sinónimo de escalamiento clásico.

El algoritmo del EM clásico consiste en:

- Dada una matriz *r* x *r* de distancias entre puntos $\triangle = (\delta_{st})$, formar la matriz *r* x *r* $\mathbf{A} = (a_{st})$, donde $a_{st} = -\frac{1}{2}\delta_{st}^2$
- Calcular la matriz doble centrada, simétrica, *r* x *r,* **B=HAH**, donde $\mathbf{H} = \mathbf{I}_r - r^{-1}\mathbf{J}_r$ y $\mathbf{J}_r = \mathbf{1}_r\mathbf{1}_r^T$ es una matriz de unos *r* x *r.*
- Calcular los autovalores y autovectores de **B.** Aplicando el teorema de descomposición espectral, $\mathbf{B} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T$, donde $\boldsymbol{\Lambda}$ es la matriz diagonal con los autovalores de **B,** y $\mathbf{V}$ es la matriz con los autovalores de **B** en las columnas.
- Si **B** es definida no negativa con rango=*r*(**B**)=*r'*<*r*, el autovalor más grande será positivo y los restantes *r-r'* autovalores serán cero. Sea $\boldsymbol{\Lambda}_1 = diag(\lambda_1, \dots, \lambda_{r'})$ la matriz diagonal t x t de los autovalores positivos de **B** y sea $\mathbf{V}_1 = (\mathbf{v}_1, \dots, \mathbf{v}_t)$ la correspondiente matriz de autovecotres de **B**. Entonces,

  $\mathbf{B} = \mathbf{V}_1\boldsymbol{\Lambda}_1\mathbf{V}_1^T = \left(\mathbf{V}_1\boldsymbol{\Lambda}_1^{1/2}\right)\left(\boldsymbol{\Lambda}_1^{1/2}\mathbf{V}_1\right) = \mathbf{Y}\mathbf{Y}^T$, donde $\mathbf{Y} = \mathbf{V}_1\boldsymbol{\Lambda}_1^{1/2} = \left(\sqrt{\lambda_1}\mathbf{v}_1, \dots, \sqrt{\lambda_{r'}}\mathbf{v}_{r'}\right) = (\mathbf{y_1}, \dots, \mathbf{y}_r)^T$
- Las coordenadas principales, que son las columnas $\mathbf{y_1}, \dots, \mathbf{y}_r$ de la matriz $\mathbf{Y}^T$ *r'* x *r,* proporcionan los *r* puntos en el espacio de dimensión *r',* cuyas distancias son iguales a las disimilaridades de la matriz $\triangle$.
- Si los autovalores de **B** no son todos no negativos, entonces se puede ignorar los autovalores no negativos y sus respectivos autovectores o adicionar una constante a las disimilaridades y comenzar nuevamente. Si

r' es muy grande para objetivos prácticos, se pueden tomar los primero *c* autovalores para construir las coordenadas principales. En este caso las distancias se aproximan a las disimilaridades de la matriz Δ .

Las coordenadas principales que se obtienen en el escalamiento multidimensional clásico cuando se utiliza como disimilaridad la distancia Euclídea son equivalentes a las coordenadas obtenidas por el ACP. Por esta razón, también se la conoce como Análisis de Coordenadas Principales (Gower 1966).

.

2.2.3.3. Escalamiento multidimensional de distancias

El escalamiento clásico se resuelve con descomposición en autovalores y por lo tanto es equivalente al ACP si el objetivo es reducir dimensionalidad. Los métodos de escalamiento de distancia minimizan una función de pérdida (o stress) utilizando procesos iterativos para encontrar la solución.

$$L_f(\mathbf{y_1}, \dots, \mathbf{y}_r, \mathbf{W}) = \sum_{s<t} w_{st}\left(d_{st} - f_{(\delta_{st})}\right)^2$$

Los diferentes métodos varían en la propuesta de la matriz **W,** así como en la función aplicada a las disimilaridades e incluso reemplazan las distancias d_{st} por una una función de estas.

Por ejemplo, el método de mapeo no lineal de Sammon utiliza como matriz de pesos **W,** $w_{st} = (\delta_{st}^{-1}\{\sum_{s\prime<t\prime} \delta_{s\prime t\prime l}\}^{-1})$ y función f identidad. Esto hace que se preserven las distancias pequeñas y resulte particularmente útil para identificar clusters.

Por otro lado, el escalamiento no métrico (Shephard-Kruskal), minimiza una función de perdida en la que f preserva el orden de rango de las disimilaridades.

2.2.4. Descomposición en valores singulares Poisson

Esta técnica fue propuesta recientemente para datos de conteos de RNA-seq (S. Lee et al. 2013) y considera simultáneamente la reducción de las dimensiones y la normalización de las muestras del siguiente modelo lineal generalizado:

$$\begin{cases} y_{ij} \sim Poisson(\lambda_{ij}) \\ \lambda_{ij} = T_i p_{ij} \\ log(\lambda_{ij}) = log(T_i) + \beta_{i1} f_{j1} + \cdots + \beta_{iv} f_{jv} \end{cases}$$

donde: $i=1$ a n, $j=1$ a p, l=1 a v, T_i es un parámetro offset para la i-ésima muestra, p_{ij} es la proporción normalizada y β_{il} es el l-ésimo factor de scores par el i-ésimo perfil y $\mathbf{f}_l=(f_{1l},\ldots,f_{pl})^T$ es el k-ésimo factor.

Brevemente el algoritmo comienza calculando los primeros $\mathbf{f}_l$ extrayendo los vectores de la derecha de la descomposición en valores singulares al logaritmo de la matriz de conteos centrada por filas. Luego iterativamente ajusta modelos log-lineales para cada fila y columna obteniendo valores para **B** y **F**, centrando nuevamente las filas de $\mathbf{BF}^T$ y aplica descomposición en valores singulares para obtener los vectores singulares de la izquierda. El proceso se repite hasta lograr convergencia y es llamado PSVDOS (Poisson Single Value Decomposition with offset parameters).

2.3. Análisis multivariado a 3 vías.

Si la matriz de datos presenta dos direcciones de clasificación, *n* individuos y *p* variables, entonces técnicas de análisis multivariado a 2 vías como las presentadas en el punto 2.2 pueden ser aplicadas. Sin embargo es común encontrar experimentos en los que hay una tercera vía de clasificación de interés, por ejemplo *q* condiciones, y en estos casos deben aplicarse técnicas de análisis multivariado a 3 vías para analizar la matriz o tabla de datos que tiene dimensiones *n* x *p* x *q*.

En el análisis multi-vía las direcciones de clasificación (vías) pueden estar formadas por el mismo tipo de entidad (modo) o por uno diferente (Kroonenberg 2008; Carroll and Arabie 1980). Si se cuenta con un mismo conjunto de *n* individuos a los que se les midieron las mismas *p* variables en *q* condiciones, las tres vías están totalmente cruzadas y constituye la definición clásica de datos a tres vías (Kiers 1991), a estos datos también se los llama a tres vías y tres modos (Kiers 2000). Sin embargo, es frecuente contar con datos clasificados a tres

vías pero utilizando solo dos modos, por ejemplo cuando se tiene más de un conjunto de individuos a los que se les ha determinado las mismas variables, o cuando a un mismo conjunto de individuos se le observan diferentes conjuntos de variables. Este tipo de datos se denominan de conjuntos múltiples, o datos de tres vías y dos modos, y la tercera vía de clasificación surge al agrupar sobre uno de los modos (Kiers 1991). En la Figura 6 se representan los tipos de datos a tres vías.

En el experimento más típico de RNA-seq, se observa la expresión de los mismos transcripts (variables) a individuos que reciben diferentes tratamientos por lo que constituye un caso de conjuntos múltiples con datos de tres vías y dos modos donde la tercera vía surge al agrupar los individuos por tratamiento Figura 6-b.

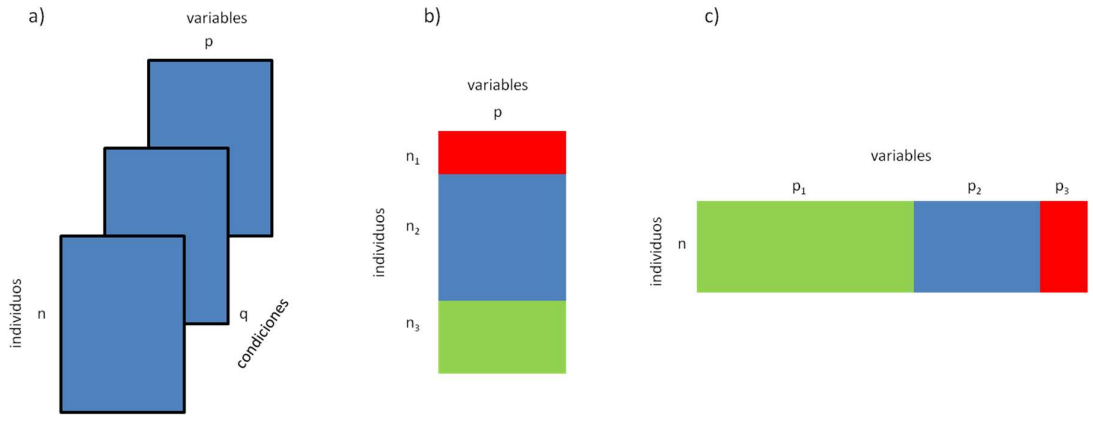

Figura 6. Tipos de datos a tres vías

a) Datos a tres vías totalmente cruzadas, a un mismo conjunto de individuos se le determinan las mismas variables en un mismo conjunto de condiciones, b) datos de conjuntos múltiples con tres conjuntos de individuos sobre los que se determinan el mismo conjunto de variables,, y c) datos de conjuntos múltiples donde a un mismo conjunto de individuos se le determinan tres conjuntos de variables.

2.3.1. Análisis factorial múltiple (AFM)

El AFM propuesto por Escofier y Pages (1990; 1994) es una generalización del ACP que tiene por objetivo analizar varios conjuntos de variables medidas sobre un mismo conjunto de individuos. En la versión dual (Lê and Pagès 2010) también se pueden analizar varios conjuntos de observaciones medidas sobre un mismo conjunto de variables.

Los grupos de variables pueden ser cuanti o cualitativas pero siempre del mismo tipo en un mismo grupo. En RNA-seq las variables son siempre cuantitativas por lo que se desarrolla el análisis para este tipo de variables.

La idea general del AFM (Figura 7) consiste en normalizar cada grupo de variables de manera tal que sus primeras componentes principales tengan la misma longitud según la medida del primer autovalor de cada grupo, y luego combinar las tablas de datos en una representación común llamada compromiso o consenso. Este compromiso se obtiene a partir de una ACP sin normalizar de una gran tabla que concatena todas las tablas normalizadas. Este ACP permite visualizar las observaciones en el espacio compromiso. También pueden obtenerse las representaciones parciales, es decir las observaciones según cada conjunto de variables, y visualizarlas en el mapa compromiso. De igual forma, las cargas de las variables pueden ser calculadas para estudiar la importancia de las variables en la construcción de las componentes principales y estudiar la asociación entre variables.

2.3.2. Pasos del análisis factorial múltiple

1. Sean $\mathbf{X}_1$, $\mathbf{X}_2$, … $\mathbf{X}_q$, q matrices, tablas de datos originales donde cada tabla $\mathbf{X}_k$ *tiene dimensiones* n x p_k, es decir contiene p_k variables medidas sobre las mismas observaciones. El AFM comprende tres grandes pasos: Se calcula un ACP de cada tabla de datos y se toma el primer autovalor de cada tabla.

   $\mathbf{X}_k = \mathbf{U}_k \boldsymbol{\Gamma}_k \mathbf{V}_k$ donde $\mathbf{U}_k$ y $\mathbf{V}_k$ contienen los autovectores izquierdo y derecho respectivamente, y $\boldsymbol{\Gamma}_k$ los respectivos autovalores.

   El peso de cada tabla se obtiene de su primer autovalor:

   $$\alpha_k = \frac{1}{\lambda_{1,k}^2}$$

   Los pesos obtenidos para cada tabla pueden ordenarse en una matriz diagonal $\mathbf{A} = diag\{[\alpha_1 \mathbf{1}_1^T, \dots, \alpha_k \mathbf{1}_k^T, \dots, \alpha_q \mathbf{1}_q^T]\}$ donde $\mathbf{1}_k$ representa un vector de unos P_k .

2. Se genera una gran matriz $\mathbf{X}$ concatenando todas las tablas y se calcula un ACP sin normalizar descomponiendo la gran matriz con una descomposición en valores

singulares generalizada donde los pesos de las columnas se obtienen del primer valor singular al cuadrado de cada tabla de datos. O una forma alternativa consiste en dividir todos los elementos de cada tabla por su respectivo primer valor singular, concatenar las tablas y luego calcular un ACP sin normalizar.

$\mathbf{X} = \mathbf{P\Delta Q}^T$ con $\mathbf{P}^T\mathbf{MP} = \mathbf{Q}^T\mathbf{AQ} = \mathbf{I}$

Si hacemos $\mathbf{F} = \mathbf{P\Delta}$, $\mathbf{F}$ contiene las coordenadas de los factores que describen las observaciones y $\mathbf{Q}$ contiene las cargas que describen las variables.

Como la matriz $\mathbf{X}$ concatena *k* tablas, cada una con p_k variables, la matriz $\mathbf{Q}$ puede expresarse como una matriz bloque $\mathbf{Q} = \left[\mathbf{Q}_1^{\mathrm{T}}| \ldots |\mathbf{Q}_k^{\mathrm{T}}|\mathbf{Q}_q^{\mathrm{T}}\right]^{\mathrm{T}}$ donde cada sub-bloque contiene los vectores singulares de la derecha de cada grupo de variables de la matriz $\mathbf{X}_k$ **.** Por lo tanto se puede expresar a

$$\mathbf{X} = \left[\mathbf{P\Delta Q}_1^{\mathrm{T}}| \ldots \left|\mathbf{P\Delta Q}_k^{\mathrm{T}}\right| \ldots |\mathbf{P\Delta Q}_q^{\mathrm{T}}|\right]$$

3. En el tercer paso, se obtienen las coordenadas parciales de las observaciones para cada factor proyectando cada tabla de datos en el espacio común.

   Las coordenadas de los factores contenidas en $\mathbf{F}$ se pueden usar para obtener la representación de las observaciones en el espacio compromiso, y recordando la estructura en bloque de $\mathbf{Q}$ se pueden obtener las coordenadas parciales para cada tabla proyectando cobre los vectores singulares contenidos en $\mathbf{Q}_k$.

   Los elementos de $\mathbf{Q}$ son las cargas y también pueden ser representadas solas en conjunto con las observaciones en un biplot. Las variables también pueden representarse en el círculo de correlación.

   La simplicidad teórica y computacional hace del AFM una herramienta ideal para integrar los grandes conjuntos de datos de la ciencia moderna.

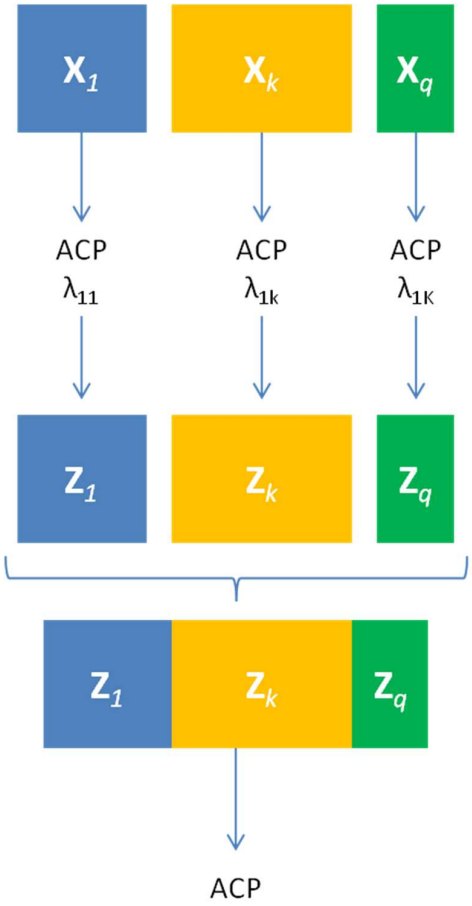

Figura 7. Pasos del Análisis Factorial Múltiple

Las $\mathbf{X}_k$ tablas originales se analizan cada una con ACP y se dividen todos los elementos de cada tabla por su respectivo primer valor singular. Luego se concatenan las tablas normalizadas (Z_k) y se calcula un ACP.

2.3.3. Análisis Factorial Múltiple Dual

Esta técnica constituye una extensión del AFM (Lê and Pagès 2010) para el caso en que se cuentan con *q* conjuntos de observaciones medidas sobre el mismo conjunto de variables. En este caso se computan *q* productos cruzados entre las matrices de covarianzas, una por cada conjunto de individuos, y se aplican los mismos pasos de AFM. Se obtiene un espacio compromiso para las variables y cargas parciales para cada tabla.

Esta versión del AFM permite analizar las correlaciones de las variables en diferentes grupos de individuos en una representación simultánea de los gráficos de dispersión para cada grupo de individuos. Las correlaciones entre genes no necesariamente son las mismas entre grupos de individuos que reciben por ejemplo diferentes tratamientos por lo que esta técnica resulta de utilidad para estudiar las correlaciones inducidas por cada tratamiento.

Debe tenerse particularmente en cuenta como se realizar el centrado de las variables. Si se centran considerando las medias de las observaciones de toda la tabla, el análisis será sensible a diferencias de medias entre los grupos de individuos (efectos principales). Contrariamente, si las diferencias son sólo importantes dentro de cada tabla, las variables deben centrarse separadamente para cada grupo de individuos.

2.3.4. Otras técnicas de análisis de tablas múltiples relacionadas a AFM

El AFM tiene otras versiones como AFM Jerárquico y AFM de procrustes . Además forma parte de una familia de ACP para tablas múltiples que comprende técnicas como STATIS, análisis de correspondencia multibloque (MUDICA), ACP consenso y SUM-PCA. Otros métodos relacionados son el análisis de correlaciones canónicas generalizado, DISTATIS, INDSCAL y análisis de Tucker (Abdi, Williams, and Valentin 2013).

2.4. Validación y comparación

La validación de métodos de aprendizaje no supervisado es compleja y no existe consenso en como evaluar la calidad de los resultados obtenidos. A diferencia de las técnicas de aprendizaje supervisado (por ejemplo análisis discriminante, arboles de clasificación o support vector machines) donde técnicas como validación cruzada o la validación utilizando un conjunto de datos independientes constituyen herramientas claramente reconocidas de validación, en las técnicas de aprendizaje no supervisado (como análisis de conglomerados o análisis de componentes principales), la respuesta verdadera no se conoce, los objetivos de análisis no siempre están claramente determinados pues son frecuentemente parte de un

análisis exploratorio y por lo tanto las herramientas de validación tienden a ser más subjetivas (James et al. 2013).

Una serie de índices puede utilizarse para comparar resultados obtenidos con técnicas de análisis multivariados (Xiong and Li 2013). Según los objetivos del estudio uno u otro índice resulta más conveniente (Jiang, Tang, and Zhang 2004). En aplicaciones de expresión génica existen numerosos trabajos que evalúan el comportamiento de estos índices sobre todo para resultados de análisis de conglomerados aunque la mayoría utiliza una partición de grupos preestablecida como referencia (Dalton, Ballarin, and Brun 2009). Por ejemplo el índice de Rand (Rand 1971) puede ser usado para evaluar la tasa de clasificación errónea pero requiere una partición estricta de las observaciones. Otros índices como compactación son útiles para evaluar la calidad interna de los conglomerados.

La correlación entre matrices de distancia, ya sean matrices cofenéticas de dendrogramas o entre matrices de distancia Euclidea de una configuración también puede ser utilizada como medida de comparación sin la restricción de definir los grupos en forma estricta (Handl, Knowles, and Kell 2005). El coeficiente RV (Escoufier 1973) es una generalización de la correlación entre matrices particularmente utilizado al comparar configuraciones multivariadas de un mismo número de individuos incluso entre matrices de diferentes dimensiones y obtenidas con distintas disimilaridades (Robert and Escoufier 1976; Abdi 2007; Mayer, Lorent, and Horgan 2011).

2.5. Generación de datos

La generación de datos con fines de evaluar técnicas de análisis en RNA-seq incluye el desafío de la multidimensionalidad (T. S. Mehta et al. 2006) y la incidencia de fuentes de variación aún no dilucidadas al igual que sucede con otras técnicas "–omics" (proteomics, metabolomics, etc.) (Sloutsky et al. 2013; G. Gadbury, Garrett, and Allison 2009; Allison et al. 2009)

Una de las técnicas más utilizadas para estudiar la validación de métodos estadísticos es la simulación de datos en base a modelos pre-establecidos. Estas simulaciones paramétricas

permitan conocer íntegramente a priori el verdadero estado de naturaleza de los datos tales como la pertenencia a un grupo y la distribución de probabilidades que la origina, permitiendo así evaluar los métodos bajo distintos escenarios . En datos de secuenciación de ARN los modelos más utilizados consideran distribuciones Poisson o Binomial Negativa (Robinson, McCarthy, and Smyth 2010; S Anders and Huber 2010).

La complejidad de los datos genéticos hace que toda simulación paramétrica constituya una representación parcial de la realidad. Por ejemplo, la estructura de correlación entre genes es difícil de imitar ya que generalmente no se conoce. Una alternativa que considera una generación de datos más realista es la obtenida a través de plasmodios (T. Mehta, Tanik, and Allison 2004; T. S. Mehta et al. 2006). Un plasmodio es un conjunto de datos generado a partir de datos experimentales en los que algún aspecto verdadero es conocido de antemano (G. L. Gadbury et al. 2008). La idea de los plasmodio fue originalmente propuesta para análisis factorial (Cattell and Jaspers 1967) y ha sido aplicado en ciencias del comportamiento (Waller, Underhill, and Heather 2010) y en genética (G. L. Gadbury et al. 2008; Vaughan et al. 2009; Steibel et al. 2009). Recientemente han sido utilizados el contexto de experimentos de RNA-seq para la comparación de modelos de análisis de expresión diferencial (Reeb and Steibel 2013) y análisis de conglomerados (Reeb, Steibel, and Bramardi 2013) .

Los plasmodios son conjuntos de datos sintéticos que se generan sin asumir un modelo paramétrico en particular e imitan las condiciones experimentales asumiendo cierta estructura o propiedades de los datos. Por lo tanto no hay una solo forma de construir plasmodios, pues su formulación depende de los datos experimentales disponibles y de las técnicas estadísticas que se desean evaluar.

Capítulo 3: MATERIALES Y METODOS

3.1. Datos

A partir de datos experimentales obtenidos por Bottomly et al. (2011) y disponibles en el repositorio ReCount (Frazee, Langmead, and Leek 2011) se generaron 50 nuevos conjuntos de datos siguiendo la metodología de plasmodios.

Los datos experimentales corresponden a un ensayo en el que se secuenció el ARN mensajero de 21 muestras de tejido del *striatum* de dos líneas consanguíneas de ratones (C57BL/6J (B6), n=10; and DBA/2J (D2), n=11) en tres celdas de flujo de Illumina GAIIx (San Diego, Ca, USA). Luego de filtrar los transcripts que contenían ceros en todas las muestras, la matriz de conteos originales quedo constituida por 13932 filas (transcripts) y 21 columnas (muestras).

Los plasmodios se generaron siguiendo los pasos descriptos en la Figura 8 utilizando las muestras correspondientes a la línea consanguínea B6. Estas se separan en dos grupos y a uno de estos se le adicionan los efectos de algunos transcripts respecto a la línea consanguínea D2. El algoritmo inicia analizando la expresión diferencial de los transcripts en base a dos fuentes de variación, línea consanguínea y celda de flujo, cuya importancia ha sido evaluada previamente (Reeb and Steibel 2013; Law et al. 2014). El análisis de expresión diferencial incluyendo todas las muestras se ejecutó con el programa edgeR (Robinson, McCarthy, and Smyth 2010) y se consideraron diferencialmente expresados a los transcripts con q-valor < 0.05. Posteriormente, las muestras de la línea B6 se dividieron aleatoriamente en dos grupos (A y B) dentro de cada celda de flujo, y un conjunto de efectos seleccionados al azar entre los transcripts diferenciados se agregó a los transcripts correspondientes a las muestras etiquetadas como B. Por lo tanto, las muestras difieren entre los grupos A y B debido a los efectos de línea consanguínea adicionados, mientras que dentro de cada grupo las muestras difieren debido a los efectos de celda de flujo. Se generaron 50 plasmodios con un 10% de transcripts diferenciados. Según la cantidad de muestras disponibles, se asignaron una

muestra al grupo A y una o dos al grupo B en cada celda. De esta forma, cada plasmodio resultante está formado por 10 muestras clasificadas según dos tratamientos (A o B) y tres bloques (1, 2 o 3) que corresponden a las siguientes etiquetas: {(A_1, A_1, B_1, B_1),(A_2,B_2,B_2),(A_3,B_3,B_3)}.

Finalmente, para cada plasmodio se determinaron dos posibles escenarios. El escenario 1 ($ED_{100\%}$) contempla la inclusión solamente de los transcripts expresados diferencialmente, mientras que el escenario 2 ($ED_{10\%}$+$nED_{90\%}$) incorpora también transcripts no diferenciados.

(1) Input: datos experimentales: 21 muestras de dos líneas consanguíneas B6 (n=10) y D2 (n=11), y G transcripts
(2) Analizar los datos experimentales con un modelo lineal generalizado
    1. *filtrar los conteos de ser necesario*
    2. *modelo*: linea consanguinea + celda de flujo,
    3. Analizar diferencias entre lineas consanguineas
    4. *output*: G transcripts, con respectivos log-FC y q-valores.
(3) Definir
    1. p numero de plasmodios a generar
    2. π= proporcion de transcripts diferencialmente expresadost
(4) Generar efectos*:*
    1. Seleccionar G_1 transcripts con q-valor<0.05 a partir de G.
    2. Tomar una muestrasin reposicion de $T = \pi \times G$ transcripts a partir de G_1, $T < G_1$, y guardar los log-F en el conjunto S_1
(5) Generar particion de muestras:
    1. Seleccionar las 10 muestras de la línea consanguinea B6.
    2. Dentro de cada celda de flujo asignar aleatoriamente por lo menos una muestra a uno de dos grupos (A o B)
(6) Adicionar efectos al grupo B:
    1. Calcular el logaritmo de los conteos (c): $z= (log_2(c+1))$ para todas las muestras en el grupo B
    2. Adicionar el logFC del conjunto S_1 a los transcripts correspondientes en las muestras etiquetadas como B
(7) Output: obtener conteos del plasmodio
    - Aplicar $c=2^z-1$ a los calores obtenidos en (6)
(8) Generate nuevos plasmodios:
    - Repetir p veces los pasos 4 a 7

Figura 8. Algoritmo para generar plasmodios

3.2. Análisis Multivariado a dos vías

3.2.1. Análisis de Conglomerados

Para cada uno de los plasmodios se realizó un análisis de conglomerados jerárquico aglomerativo utilizando cada una de las 13 medidas de disimilaridad que se presentan en la Tabla 4. Dado que el principal objetivo es comparar las medidas de disimilaridad, se utilizó en

todos los casos el método de encadenamiento del máximo o también llamado encadenamiento completo. Este método es invariante bajo trasformaciones monótonas y por lo tanto medidas de disimilaridad que tienen un mismo ranking relativo, por ejemplo la distancia Euclidea estandarizada y 1- correlación de Pearson, generan la misma estructura en el dendrograma. Esta robustez posibilita la evaluación de las disimilaridades reduciendo el efecto que puede generar el método de encadenamiento.

Las disimilaridades entre muestras (columnas de la matriz) que se evaluaron (Tabla 4) incluyeron 5 variantes basadas en la distancia Euclidea, 5 basadas en correlaciones, la distancia chi-cuadrado, la disimilaridad Poisson y la basada en descomposición de valores singulares Poisson. Las distancias Euclideas se calcularon en base a: i) datos de conteos originales (*raw*), ii) datos de conteos previamente normalizados según el método propuesto por Anders et al. (2010) (*rnr*), iii) datos de conteos transformados por regularización logaritmo según la propuesta del software DESeq2 (Love, Huber, and Anders 2014) (*rlg*), iv) datos de conteos normalizados según Robinson et al. (2010) y transformados por la función `voom` (Law et al. 2014) (*voo*) , y v) datos normalizados y transformados según las funciones del inciso anterior pero considerando solo los 500 transcripts más variables según la propuesta del software edgeR (*edg*). Las disimilaridades basadas en correlaciones incluyeron: i) 1- correlación de Pearson utilizando los conteos originales (*pea*), ii) 1- correlación de Pearson utilizando los conteos originales +1 transformados por logaritmo en base 2 (*plg*), iii) 1- correlación de Pearson utilizando los conteos normalizados método propuesto por Anders et al. (2010) +1 y transformados por logaritmo en base 2 (*pln*), iv) 1- correlación de Spearman (*spe*), y v) 1 – correlación de Gini (*gcc*). Finalmente, también se incluyeron la distancia chi-cuadrado (*chi*) y dos disimilaridades propuestas para RNA-seq, la disimilaridad de Poisson (*poi*) y la 1- correlación de Pearson sobre los datos filtrados por descomposición en valores singulares Poisson (*psv*).

Etiqueta	Disimilaridad	Normalización	Transformación	Descripción

raw	Euclidea	No	No	Utiliza la distancia Euclídea sobre los datos sin modificar
rnr	Euclidea	Median ratio size factor	No	Normaliza los conteos según Anders et al. (2010)
rld	Euclidea	Median ratio size factor	regularización logaritmo_2	Normaliza los conteos según Anders et al. (2010) y utiliza la función rlog de DESeq2 (2014)para modelar la relación media-varianza y transformar los conteos por millón a la escala $\log_2$
voo	Euclidea	TMM	regularización logaritmo_2(cpm)	Normaliza los conteos según TMM de Robinson et al. (2010)y utiliza la función voom de Law et al. (2014)para modelar la relación media-varianza y transformar los conteos por millón a la escala $\log_2$
edg	Euclidea	TMM	regularización logaritmo_2(cpm)	Normaliza y transforma como en *voo*, pero considera solo los 500 transcripts con mayor varianza
pea	$1 - r_{pea}$	No	No	Calcula 1 - la correlación de Pearson
plg	$1 - r_{pea}$	No	$\text{logaritmo}_2(c+1)$	Calcula 1 - la correlación de Pearson a los conteos +1 transformados por logaritmo en base 2
pln	$1 - r_{pea}$	Median ratio size factor	$\text{logaritmo}_2(c_n+1)$	Calcula 1 - la correlación de Pearson a los conteos +1 transformados por logaritmo en base 2 y previamente normalizados según Anders et al. (2010)
spe	$1 - r_{spe}$	No	No	Calcula 1 - la correlación de Spearman
gcc	$1 - r_{gcc}$	No	No	Calcula 1 - la correlación de Gini
chi	d_{χ}^2	No	No	Calcula la distancia chi-cuadrado entre las muestras
poi	d_{poi}	Suma total por muestra	No	Calcula la disimilaridad Poisson según Witten (2011)
psv	$1 - r_{pea}$	offsets	No	Calcula conteos en base a modelos log-lineales con reducción de dimensiones (S. Lee et al. 2013)

Tabla 4. Disimilaridades evaluadas

TMM: ARN de una muestra
c: matriz de conteos, $\mathbf{c_n}$: matriz de conteos normalizados, **cpm**: matriz de conteos expresados por millón

La inclusión de *raw* y *rnr* permite comparar el efecto de normalización entre sí y a su vez permite evaluar el efecto de estandarización respecto de las medidas basadas en correlación de

Pearson. Estas últimas son ampliamente usadas en análisis de expresión génica de micromatrices y las variantes Spearman y Gini son candidatas para modelar relaciones no lineales como las encontradas en RNA-seq. Las propuestas de *voo*, *rld* y *edg* corresponden a dos de los paquetes de análisis más utilizados en análisis de expresión diferencial en RNA-seq. La disimilaridad Poisson también ha sido elaborada para datos de conteo de RNA-seq. Finalmente, la distancia chi-cuadrado es ampliamente utilizada en datos de frecuencia donde los conteos son estandarizados por marginales fila y columna por lo que resulta de interés evaluar su performance en datos de RNA-seq.

Además de las disimilaridades de la Tabla 4, se incluyó la descomposición en valores singulares Poisson , *psv*, que fue recientemente presentada como una alternativa para analizar datos de conteos de RNA-seq (S. Lee et al. 2013). Se obtuvo por cálculo directo sobre los datos originales ya que el modelo contempla el cálculo de offsets como factores de normalización. Se utilizó el algoritmo PSVDOS versión 4 provisto por los autores para obtener la descomposición en las tres primeras dimensiones. Posteriormente, se calculó la correlación de Pearson para obtener la matriz de disimilaridades y el dendrograma correspondiente siguiendo la metodología propuesta por los autores (S. Lee et al. 2013).

Los paquetes edgeR v.3.6.8, PoiClaClu v. 1.0.2, DESeq2 v. 1.4.5, lima v.3.20.8 y gcc 1.0.6 se utilizaron para los cálculos de *edg*, *poi*, *rld*, *voo* y *gcc* respectivamente. La distancia chi cuadrado se calculó con un programa R diseñado a tal efecto.

El análisis de cluster jerárquico aglomerativo se realizó con la función `hclust` utilizando la función `dist` para el cálculo de distancias Euclideas y 1-correlación.

Comparación y validación de agrupamientos

Existen numerosos índices para evaluar resultados en análisis de conglomerados (Halkidi, Batistakis, and Vazirgiannis 2001; Xiong and Li 2013) y la elección de un índice en particular depende del tipo de análisis y objetivos del estudio (Jiang, Tang, and Zhang 2004). Aquí se propuso la comparación de las estructuras de los árboles y no la partición estricta en número predeterminado de grupos por lo que se planteó el uso de la correlación entre matrices

cofenéticas (Handl, Knowles, and Kell 2005). Se compararon tanto los dendrogramas obtenidos con diferentes medidas de disimilaridad (comparación entre medidas) como los dendrogramas obtenidos con una medida en particular (comparación dentro de cada medida). La media y el desvío estándar de las correlaciones entre medidas de disimilaridad se utilizaron como medidas de concordancia mientras que la media y el desvío estándar de las correlaciones dentro de cada medida de disimilaridad se usó como medida de consistencia. La estructura de los cluster se comparó con la estructura natural de referencia conocida a priori en el proceso de generación del plasmodio.

La comparación de los dendrogramas se realizó con el paquete CLUE versión 0.3-48 (Hornik 2005)

3.2.2. Técnicas de ordenación

Las primeras doce medidas presentadas en la Tabla 4 se sometieron a escalamiento multidimensional métrico (*_m*) y no métrico (*_N*) resultando en un total de 24 procedimientos de ordenación. Se excluyó del análisis la disimilaridad *psv* ya que la descomposición en valores singulares que utiliza el métodos PSVDOS es en sí misma una técnica de reducción de dimensiones y aplicar un escalamiento mutidimensional a dicha matriz equivaldría a resumir nuevamente la información. Por otro lado, los autores (S. Lee et al. 2013) recomiendan el uso en conjunto con técnicas de cluster y el algoritmo no provee los resultados de la descomposición en valores singulares para calcular las coordenadas de las observaciones.

3.2.2.1. Escalamiento multidimensional (EM) métrico o análisis de coordenadas principales (ACoP)

Se utilizó la función `cmdscale` del lenguaje R versión 3.1.0 para obtener la representación de las muestras en el plano principal. Las coordenadas de las observaciones en las dos primeras dimensiones se utilizaron para calcular la matriz de distancias Euclideas entre las muestras utilizando la función `dist` del mencionado lenguaje.

Nótese que como el escalamiento clásico utiliza descomposición en autovalores y autovectores, la representación que se obtiene al utilizar la disimilaridades basadas en correlación de Pearson (*pea_m*) es equivalente a realizar un Análisis de Componentes Principales, y cuando se utiliza la distancia chi-cuadrado (*chi_m*) el análisis es equivalente al Análisis de Correspondencias.

3.2.2.2. Escalamiento multidimensional no métrico (nEM)

Se utilizó la función `isoMDS` del paquete MASS del lenguaje R versión 3.1.0 para obtener la representación de las muestras en dos dimensiones. Las coordenadas de las observaciones se utilizaron para calcular la matriz de distancias Euclideas entre las muestras utilizando la función `dist` del mencionado lenguaje.

Dado que las disimilaridades planteadas no son todas Euclideas, se consideró apropiado incluir una técnica de ordenación no métrica para comparar con métodos métricos usualmente utilizados por ejemplo por edgeR y DESeq.

3.2.2.3. Dimensiones, comparación y validación

Se decidió comparar la representación obtenida por las técnicas en dos dimensiones con el fin de facilitar la comparación y focalizar en la representación de resultados en la opción de visualización mayoritariamente utilizada por el investigador.

Para comparar las configuraciones obtenidas, se calcularon los coeficientes RV (Escoufier 1973) entre y dentro de cada procedimiento de ordenación. El coeficiente RV es una generalización del coeficiente de correlación especialmente utilizado para comparar matrices en configuraciones de análisis multivariado (Robert and Escoufier 1976; Abdi 2007). Al igual que en el análisis de conglomerados la media y desvío de las correlaciones calculadas por el coeficiente RV entre medidas de disimilaridad se utilizó como medida de concordancia, y la media y desvío de las correlaciones dentro de cada medida de disimilaridad como magnitud de consistencia. Las configuraciones se compararon con la configuración natural de referencia

conocida a priori en el proceso de generación del plasmodio. Además se reportan los valores de inercia explicados en el caso de EM y los valores de stress para nEM.

El coeficiente RV se calculó con la función `coeffRV` del paquete FactoMineR versión 1.26 (Le, Josse, and Husson 2008).

Alternativamente al coeficiente RV, se calcularon las correlaciones entre matrices de distancias Euclideas de las configuraciones en dos dimensiones tanto entre como dentro de cada procedimiento de ordenación. Los resultados obtenidos fueron concordantes con los obtenidos por el coeficiente RV.

3.3. Análisis Multivariado a tres vías

La estructura de los datos originales y de los plasmodios generados cuenta con un número pequeño de muestras y un número muy grande de las mismas variables (transcripts) medidas sobre todas las muestras que constituyen las dos vías del análisis tradicional. Como tercera vía de análisis se consideró el agrupamiento de las muestras por celda de flujo o tratamiento y en consecuencia el análisis factorial múltiple dual (AFMD) (Lê and Pagès 2010) resultó la alternativa conveniente de análisis multivariado a tres vías. El AFM proporciona un carácter más exploratorio que otras técnicas como INDSCALL (F. Husson and Pagès 2006). En la Figura 9 se presenta un esquema con la estructura de los datos y las 3 vías de análisis, nótese que los individuos no son los mismos en cada matriz y por ende corresponde un análisis de 2 modos a tres vías.

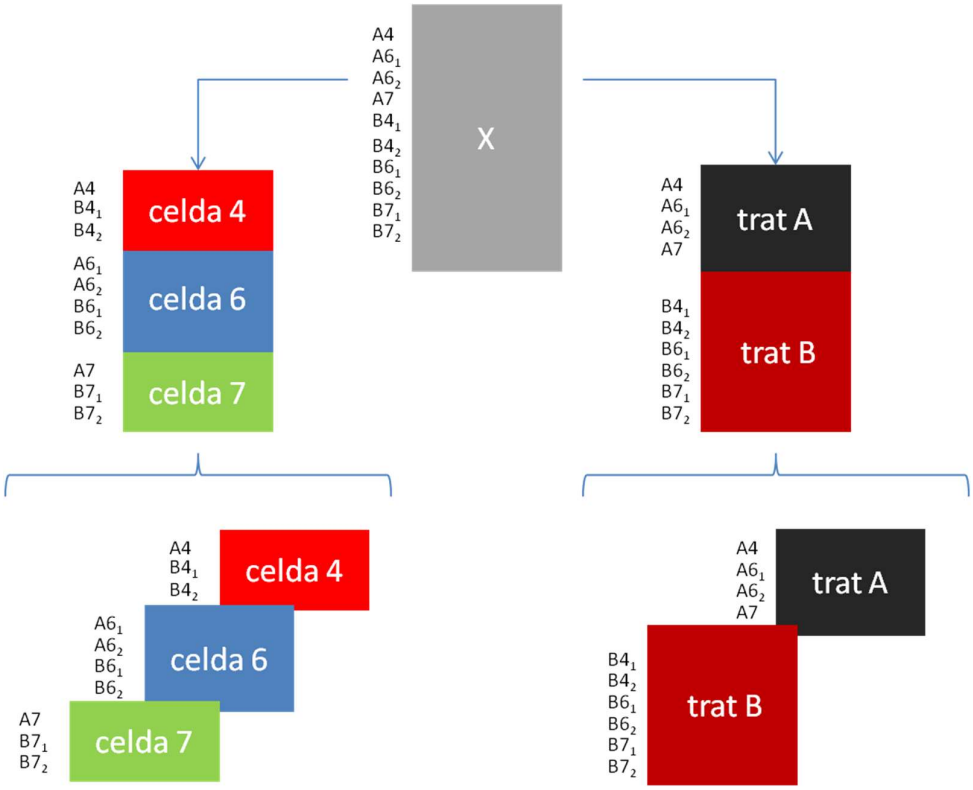

Figura 9. Estructura de los datos en el análisis a 3 vías.

Cada plasmodio cuenta con 10 individuos (primera vía) y un mismo conjunto de transcripts (segunda vía) medidos sobre todos los individuos y representados en el rectángulo gris X. Al agrupar los individuos por i) celda de flujo o ii)) tratamiento, se generan múltiples conjuntos de datos (con 3 o 2 matrices respectivamente) que constituyen la tercera vía de análisis. Nótese que los individuos son diferentes en cada matriz, por lo tanto las tres vías no están totalmente cruzadas.

De los resultados obtenidos en el análisis a dos vías, se seleccionaron las disimilaridades *rld* y *voo* para ser empleadas en el AFMD ya que fueron las que presentaron mejores valores de concordancia y consistencia en ambos escenarios.

Análogamente a lo realizado en la evaluación del análisis de ordenación a dos vías, se compararon las configuraciones consenso en dos dimensiones utilizando el coeficiente RV entre y dentro de ambas disimilaridades.

Se empleó la función `DMFA` del paquete FactoMineR versión 1.26 (Le, Josse, and Husson 2008)

3.4. Software

Todos los análisis se realizaron en el lenguaje R versión 3.1.0 (R Development Core Team 2014). Los paquetes y funciones específicos para cada análisis se mencionan en las respectivas secciones. Los códigos de programación utilizados se encuentran en el Apéndice.

Capítulo 4: RESULTADOS Y DISCUSION

4.1. Análisis multivariado a dos vías

4.1.1. Resultados del análisis de conglomerados

Las 13 medidas de disimilaridad evaluadas en el análisis de conglomerados se agruparon en 3 conjuntos según la concordancia entre todas ellas. Para el escenario en que solo se evaluaron transcripts con expresión diferencial, $ED_{[10\%]}$ (Figura 10a), se observó que el conjunto {*rld*, *voo*, *pln*, *plg*, *spe*, *edg*, *poi*, *gcc*, *psv*} formó un grupo con valores altos de concordancia mayores a 0.80, mientras que {*chi*, *rnr*} tuvieron valores intermedios entre 0.40 y 0.80, y {*pea*, *raw*} presentaron valores inferiores a 0.40. Un patrón similar se observó para el escenario con transcripts con y sin expresión diferencial, $ED_{[10\%]}+nED_{[90\%]}$, aunque las disimilaridades no fueron las mismas en los tres conjuntos (Figura 10b). Los tres conjunto correspondieron a {*poi*, *rld*, *spe*, *pln*, *gcc*, *plg*, *chi*, *psv*, *voo* },{*pea*, *rnr*} y {*edg*,*raw*}. En ambos escenarios los dendrogramas obtenidos con cualquiera de las disimilaridades del conjunto 1 fueron muy similares (correlaciones superiores a 0.80), los dendrogramas obtenidos con las disimilaridades del conjunto 2 son parecidas entre sí pero diferentes a los obtenidos con disimilaridades de otros grupos, y por último, los dendrogramas obtenidos con las similaridades del conjunto 3 no revisten similitud entre sí ni con otras disimilaridades. Se destaca además que en ambos escenarios la variabilidad aumentó a medida que la concordancia disminuyó.

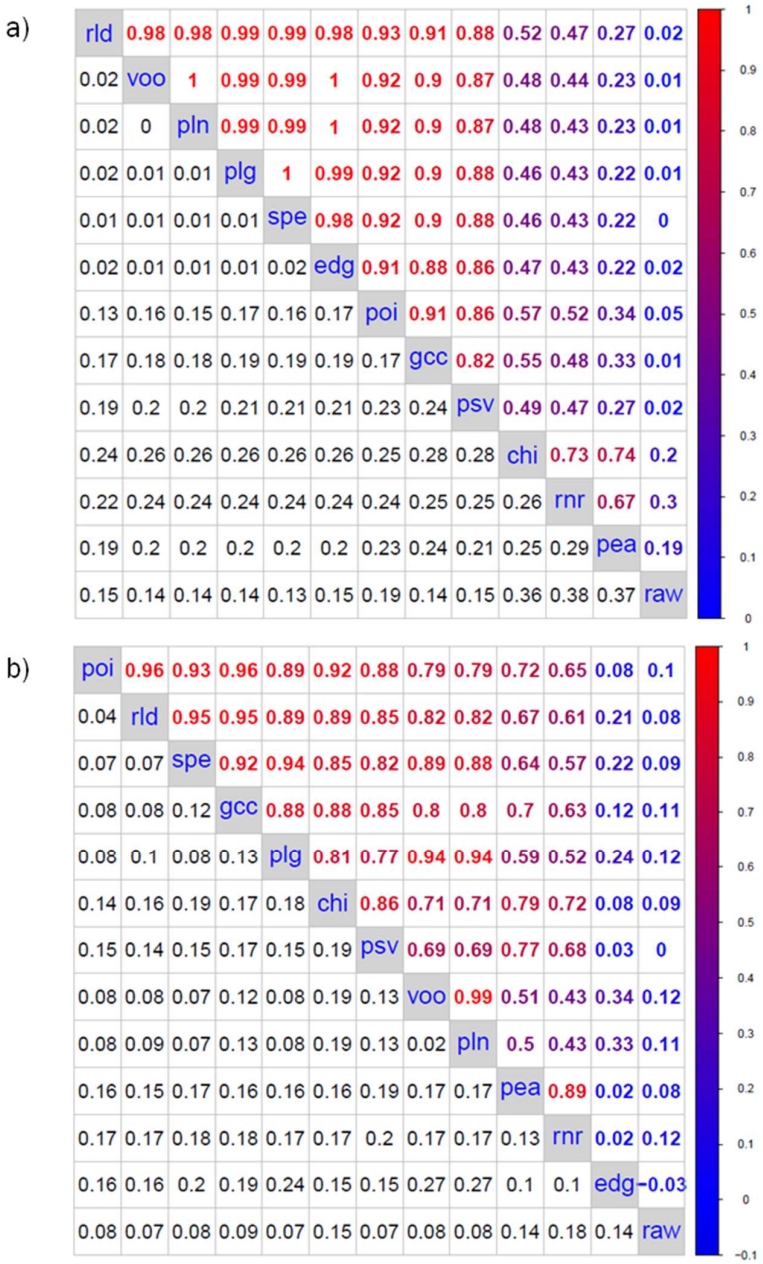

Figura 10 . Concordancia entre disimilaridades en el análisis de cluster

Matrices de concordancia entre disimilaridades utilizando a) solo los transcripts con expresión diferencial $ED_{[100\%]}$, o b) transcripts con y sin expresión diferencial. $ED_{[10\%]}$+$nED_{[90\%]}$. Cada matriz contiene en la parte superior la media y en la parte inferior el desvío estándar de las correlaciones entre matrices cofenéticas (n=50). *raw= distancia Euclidea sobre datos originales, rnr= distancia Euclidea sobre datos normalizados, rld= regularización logaritmo, voo= transformación logaritmo sobre conteros por millón, edg= distancia según edgeR, pea=correlación de Pearson, plg=correlación de Pearson sobre datos transformados por logaritmo, pln= correlación de Pearson sobre datos transformados por logaritmo y normalizados, spe=correlación de Spearman, gcc=correlación de Gini, chi=distancia chi-cuadrado, poi=disimilaridad Poisson, psv=descomposición en valores singulares*

Los valores de consistencia se presentan en la Tabla 5. Se consideró que una medida era consistente si las correlaciones entre matrices cofenéticas de todos los plasmodios tenía una media alta (>.80) y un desvío estándar bajo (<0.15). Se observaron diferencias tanto entre disimilaridades para un mismo escenario como entre escenarios para una misma disimilaridad. Por ejemplo *raw*, *rnr*, *pea* y *chi* presentaron medias bajas con desvíos estándares altos en el escenario $ED_{[10\%]}$ indicando que los dendrogramas obtenidos fueron muy distintos cuando se utilizó cada una de esas disimilaridades. Sin embargo, *raw* fue muy consistente en el escenario $ED_{[10\%]}+nED_{[90\%]}$. (media=0.91, desvío estándar=0.13) mientras que *rnr*, *pea* y *chi* se mantuvieron inconsistentes. Por otro lado, medidas como *rld* o *plg* fueron consistentes en ambos escenarios.

Disimilaridad	Consistencia	
	$ED_{[100\%]}$	$ED_{[10\%]}+nED_{[90\%]}$
raw	0.58 (0.22)	0.91 (0.13)
rnr	0.35 (0.27)	0.43 (0.28)
rld	0.98 (0.02)	0.90 (0.11)
voo	0.97 (0.02)	0.86 (0.13)
edg	0.97 (0.02)	0.63 (0.30)
pea	0.37 (0.31)	0.53 (0.29)
plg	0.98 (0.01)	0.89 (0.13)
pln	0.92 (0.18)	0.79 (0.25)
spe	0.99 (0.01)	0.86 (0.14)
gcc	0.82 (0.24)	0.88 (0.16)
chi	0.44 (0.29)	0.78 (0.21)
poi	0.86 (0.21)	0.93 (0.09)
psv	0.79 (0.25)	0.73 (0.20)

Tabla 5. Consistencia para disimilaridades en el análisis de cluster

Media y (desvío estándar) de las correlaciones entre matrices cofenéticas (n=50) para cada disimilaridad. *raw= distancia Euclidea sobre datos originales, rnr= distancia Euclidea sobre datos normalizados, rld= regularización logaritmo, voo= transformación logaritmo sobre conteros por millón, edg= distancia según edgeR, pea=correlación de Pearson, plg=correlación de Pearson sobre datos transformados por logaritmo, pln= correlación de Pearson sobre datos transformados por logaritmo y normalizados, spe=correlación de Spearman, gcc=correlación de Gini, chi=distancia chi-cuadrado, poi=disimilaridad Poisson, psv=descomposición en valores singulares*

A fin de comparar los dendrogramas obtenidos con la estructura esperada, en la Figura 11 se presentan las estructuras jerárquicas típicas para tres medidas de disimilaridad, *rld*, *rnr* y *raw,* una por cada uno de los 3 conjuntos encontrados en la Figura 10 en ambos escenarios.

Si se analizan solo los transcripts diferencialmente expresados, $ED_{[100\%]}$, se espera que muestras con igual letra (A o B) se agrupen en un mismo cluster debido al efecto de tratamiento. En contrapartida, al agregar un gran número de transcripts no diferenciados, se espera que las muestras con igual número (4, 6 o 7) tiendan a agruparse debido al efecto de celda de flujo (bloque) que ha sido demostrado anteriormente (Reeb and Steibel 2013; Law et al. 2014).

Para el escenario 1, donde se espera que las muestras se agrupen en dos cluster (A y B), se observa que solo la disimilaridad *rld* (Figura 11a) reconstruyo dicha estructura, conformando además subgrupos para cada celda de flujo (4,6 o7). Las disimilaridades *rnr* y *raw* presentaron otras estructuras no esperadas para los datos analizados (Figura 11 c y e).

Bajo el escenario 2, donde se espera que la estructura jerárquica representa principalmente aspectos no ligados a los tratamientos sino a otros factores relacionados a las celdas de flujo y otros desconocidos, se observa que las disimilaridades *rld y raw* (Figura 11b y f) correctamente agruparon las muestras por celda de flujo y luego por tratamientos. Las disimilaridad *rnr* presento una estructuras diferente a la esperadas (Figura 11 d).

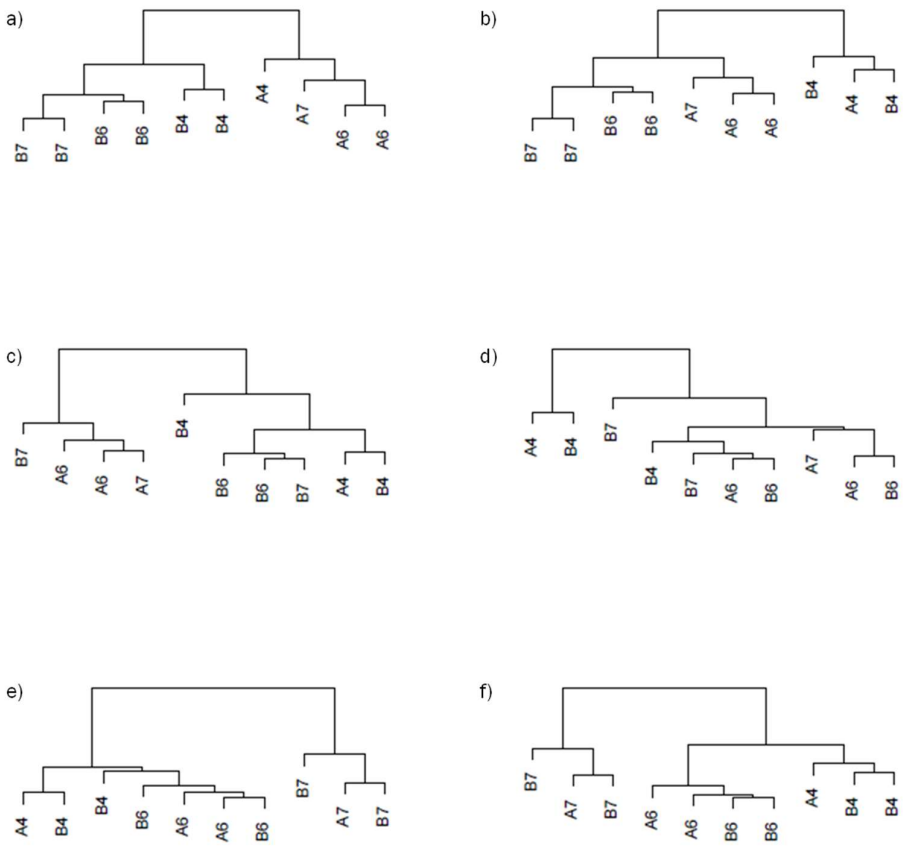

Figura 11. Dendrogramas típicos obtenidos con las disimilaridades *rld*, *rnr* y *raw*

Dendrogramas típicos obtenidos por análisis de cluster jerárquico aditivo con encadenamiento completo obtenidos con las siguientes disimilaridades: Arriba) *rld*, Centro) *rnr*, y Abajo) *raw*. En la izquierda están los dendrogramas obtenidos para el escenario 1, transcripts con expresión diferencial $ED_{[100\%]}$, y en la derecha los obtenidos para el escenario 2, transcripts con y sin expresión diferencial. $ED_{[10\%]}+nED_{[90\%]}$. Las etiquetas de la muestras corresponden al tratamiento principal (A o B) y a la celda de flujo (4, 6 o 7)

4.1.2. Discusión del análisis de conglomerados

La distancia Euclídea y disimilaridades basadas en la correlación de Pearson, generalmente sobre datos transformados por logaritmos con o sin normalización, son las medidas comúnmente utilizadas (Liu and Si 2014). Esta rutina es heredada del análisis de expresión de genes en micromatrices (Dalton, Ballarin, and Brun 2009). La disimilaridad basada en correlación de Pearson está relacionada con la distancia Euclídea estandarizada por media y varianza de modo tal que ambas medidas proporcionan dendrogramas equivalentes (Hastie, Tibshirani, and Friedman 2009). Por lo tanto, la comparación entre *raw* y *rnr* nos permite evaluar el efecto de normalización, mientras que la comparación de *raw* con *pea*, nos

permite evaluar el efecto de estandarización. Con ninguna de estas medidas se obtuvieron dendrogramas satisfactorios (Figura 11c,d y e , no se presentan los dendrogramas de *pea*), de estructura variable (Tabla 5) en ambos escenarios. La única excepción es la distancia Euclídea de los datos originales (*raw*) en el escenario 2 (Figura 11f). Esta elevada consistencia se debe a que en el escenario 2 el peso de los transcripts no diferenciados es de 9 a 1 respecto a los diferenciados y la diferencia en la profundidad de lectura de las librerías es muy marcada entre celdas de flujo (profundidad media celdas de flujo:4=4.770.406, 6=3.120.366, 7=6.648.946). Puesto que las variables no están estandarizas en *raw*, el valor absoluto de estas variables determina que la matriz de distancias esté dominadas por la magnitud y peso relativo de la expresión de estos transcripts. Una característica bien conocida de la distancia Euclídea que distorsiona las medidas entre objetos (Izenman 2008). Por lo tanto, si bien la medida *raw* en el escenario 2 es consistente para separar el efecto de celda, enmascara el análisis del efecto de tratamiento. De aquí se concluye que la mera estandarización o normalización de los datos originales no es suficiente para lograr una correcta representación.

Al utilizar la transformación logaritmo, ya sea sobre los datos originales (*plg*) o normalizados (*pln*), se logran estructuras muy similares a las esperadas (r> 0.98 en el escenario 1, y r>0.79 en el escenario 2) siendo además *plg* muy consistente en ambos escenarios (r=0.98 (0.01) y 0.89 (0.13)). Por ende, el uso de la transformación logró una mejor representación que la estandarización o normalización de los datos originales. Esta característica se debe a que es más importante captar la relación media-varianza que una estandarización o normalización homogénea para todos los datos. En datos de RNA-seq la importancia de modelar correctamente la relación media-varianza ha sido recientemente enfatizada en modelos de expresión diferencial (Law et al. 2014).

Otras disimilaridades que utilizan trasformaciones logaritmos y contemplan la modelación de la relación media-varianza (Law et al. 2014; Love, Huber, and Anders 2014), como en el caso de *rld*, *edg* y *voo*, tuvieron alta concordancia (r>0.9) con la estructura jerárquica esperada y alta consistencia sobre todo en el escenario 1. En el escenario 2, *voo* tuvo un desempeño equivalente al de *pln* ya que aplica logaritmo a los conteos por millón

estandarizando por la profundidad de la librería. La disimilaridad *edg,* no logro captar la estructura en el escenario 2, ya que la influencia de los transcripts no diferenciados determina que los 500 transcripts con mayor varianza no sean los más propicios para representar las muestras en este contexto. En este sentido los autores recomiendas su uso luego de realizar un análisis diferencial (edgeR Manual).

Distancias basadas en correlaciones no lineales como *spe* y *gcc* también obtuvieron una representación altamente correlacionada con la generada por *rld* (r>0.9). Se destaca que *spe* se utilizó incluso sin normalización y que la consistencia fue muy buena (0.99 (0.01) y 0.86 (0.14) en ambos escenarios), posiblemente debido a que esta medida preserva las relaciones entre los rangos relativos y es menos influenciada por asimetría y datos extremos (Kendall and Gobbons 1990). Si bien no se recomienda su uso cuando el objetivo es agrupar genes (Liu and Si 2014) porque se dispone de un número de muestras muy pequeño, la medida sería conveniente cuando el interés reside en agrupar muestras basándose en un gran número de genes. La disimilaridad basada en *gcc* puede resultar de interés porque considera no sólo la relación sino también los valores de los conteos (Ma and Wang 2012), sin embargo tuvo mayor variabilidad en la conformación de los dendrogramas (consistencia de 0.82 (0.24) y 0.88 (0.16) en ambos escenarios).

La distancia Chi-cuadrado, *chi*, genero dendrogramas de estructura similar a los de *rnr* en el escenario 1 y al igual que ésta no representó la estructura jerárquica esperada siendo además muy inconsistente (r=0.44(0.29)). En el escenario 2, *chi* tuvo una mejor correlación con la estructura de disimilaridades como *rld* (r=0.92(0.14)) aunque siguió siendo inconsistente (0.78(0.21)). La doble ponderación de las distancias Euclideas por fila (transcript) y columna (muestra) que efectúa la distancia chi-cuadrado (François Husson, Le, and Pages 2011) al igual que en las distancias donde sólo se efectuó normalización (*rnr*) o estandarización (*pea*) no resultó suficiente para captar la estructura natural de las muestras. En el escenario 2, y en igual sentido que *rnr* y *pea*, aunque alcanzando mayores valores de correlación que por ejemplo con *rld,* el mayor peso que reciben los transcripts no diferenciados

hace que la representación lograda con *chi* se asemeje más a la estructura esperada (r=0.89 (0.14) respecto a *rld*) si bien siguió siendo inconsistente (0.78 (0,21)).

La descomposición en valores singulares Poisson, *psv*, presentó valores satisfactorios en relación a la estructura generada por *rld*, (0.88 (0.19) y 0.85 (0.14) en cada escenario) aunque la consistencia en la obtención de dendrogramas fue escasa (0.79 (0.25) y 0.72 (0.20) en cada escenario). Un incremento en el número de dimensiones que se retuvieron en la descomposición (2 dimensiones) podría haber mejorado el uso de este tipo de filtro, aunque en este análisis no fue posible debido a problemas de convergencia del algoritmo de cálculo provisto por los autores (S. Lee et al. 2013) cuando se incrementó el número de dimensiones.

Por último la disimilaridad Poisson, presentó dendrogramas congruentes con los esperados en ambos escenarios (0.93 (0.13) y 0.96 (0.04) respectivamente) si bien la consistencia fue mayor en el escenario 2 (0.93 (0.09)) respecto a la obtenida en el escenario 1 (0.85 (0.21)). Esta medida también ha sido reportada como adecuada anteriormente (Witten 2011; Reeb, Bramardi, and Steibel 2015).

4.1.3. Resultados del análisis de ordenación a dos vías

Las doce disimilaridades utilizadas en combinación con las dos técnicas de escalamiento multidimensional hacen un total de 24 procedimientos de ordenación. En particular fue de interés analizar qué combinación de disimilaridad y tipo de escalamiento reproducía en dos dimensiones la configuración de las observaciones según los efectos de tratamiento y celda de flujo generados en los plasmodios, y evaluar si, para una misma medida de disimilaridad, el uso de un determinado tipo de escalamiento proporcionaba diferentes configuraciones.

La concordancia entre las procedimientos evaluados en el escenario 1 se presenta en la Figura 12 donde la media de los coeficientes RV de las configuraciones obtenidas en dos dimensiones se presenta en la parte superior de la matriz y el desvío estándar en la parte inferior. En función de los valores medios se destaca que *raw_m* y *raw_N* tienen en su mayoría

valores menores a 0.4 con los demás procedimientos. También se diferencia otro conjunto de procedimientos que incluye a {*chi_m*, *chi_N*, *rnr_m* , *rnr_N*, *pea_m*, *pea_N*} muy correlacionados entre sí pero con correlaciones entre 0.4 y 0.8 con el resto de los procedimientos. Las demás combinaciones de disimilaridades y escalamientos integradas por las medidas *rld*, *plg*, *spe voo*, *poi*, *edg* tanto con EM métrico o EM no métrico, presentaron generalmente valores medios superiores a 0.8 entre sí. Por otro lado en el escenario 2 (Figura 13), se destaca que el conjunto con medias más bajas está formado no sólo por *raw*_m y *raw_N* sino también por *edg_m* y *edg_N* . A su vez, el conjunto con valores intermedios (0.40 a 0.80) formado por {*pea_N*, *pea_m*, *voo_N*, *pln_N*, *rnr_m*, *rnr_N*} tiene generalmente valores más altos y cercanos al a 0.8.

Para todas las disimilaridades y en ambos escenarios, se observó una elevada concordancia (medias superiores a 0.9) entre las configuraciones obtenidas por escalamiento métrico y no métrico de una misma medida.

Al igual que en los resultados observados en el análisis de conglomerados, en ambos escenarios la variabilidad aumentó a medida que la concordancia disminuyó.

Los valores de consistencia se presentan en la Tabla 6. Similarmente a la evaluación de las medidas en el análisis de cluster, se consideró que un procedimiento era consistente si las correlaciones obtenidas por el coeficiente RV entre las configuraciones de todos los plasmodios de una misma disimilaridad tenía una media mayor a 0.80 y un desvío estándar menor a 0.15. Bajo el escenario 1 las medidas *raw*, *rnr*, *pea* y *chi* fueron inconsistentes (media [0.38 a 0.64] y desvíos [0.19 a 0.26]) y las demás presentaron valores con medias superiores a 0.84. Por el contrario, bajo el escenario 2, solamente las disimilaridades *rld*, *voo*, *plg*, *pln y spe* aplicadas al EM métrico, y edg con ambas versiones de EM resultaron consistentes.

En la Tabla 7 se muestran los valores promedio de explicación en el plano principal para los procedimientos de EM métrico y en la Tabla 8 se presentan los valores de stress para los procedimientos de EM no métrico. Estos valores sirven para describir la calidad de los resultados obtenidos con cada procedimiento.

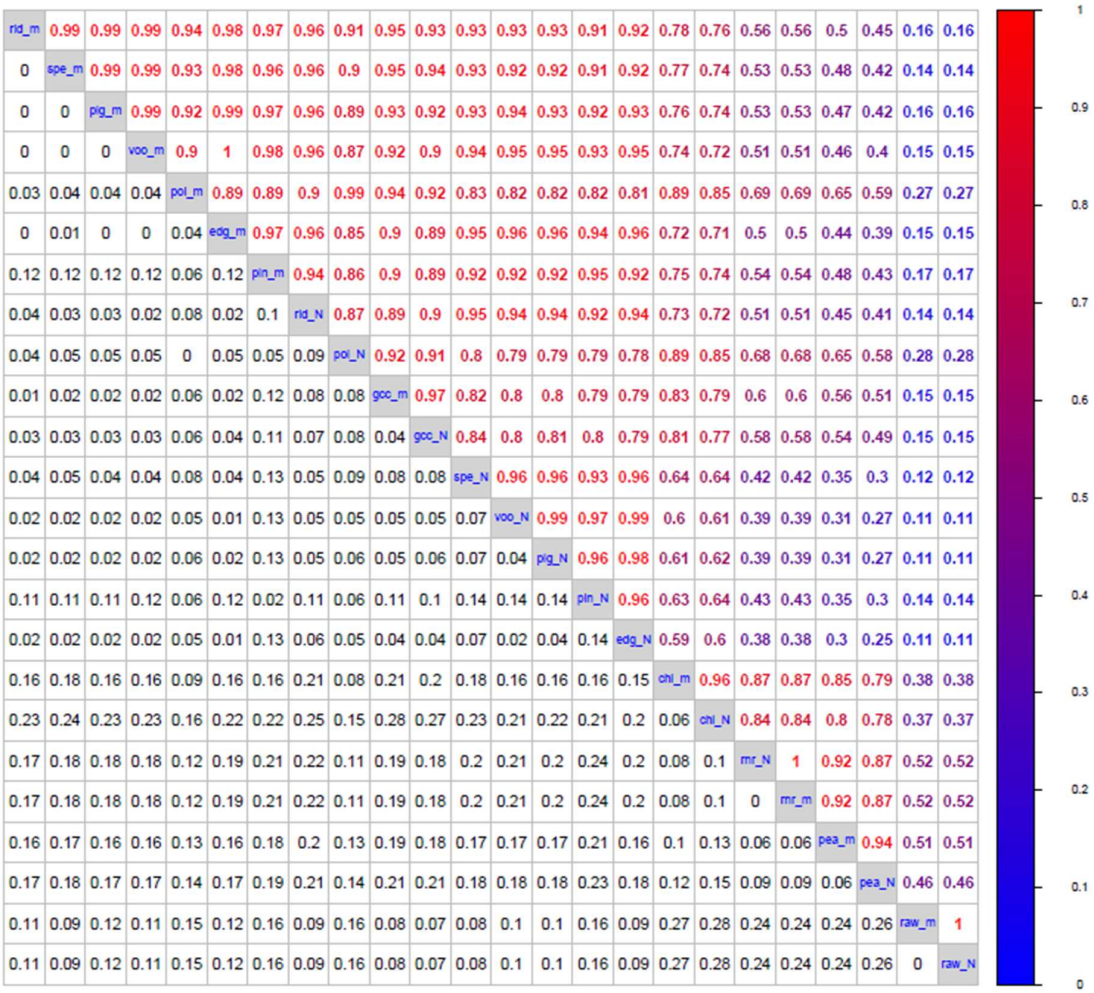

Figura 12. Concordancia entre disimilaridades en análisis de ordenación para el escenario $ED_{[100\%]}$

Matriz de concordancia entre disimilaridades utilizando solo los transcripts con expresión diferencial $ED_{[100\%]}$. La parte superior contiene la media y la parte inferior el desvío estándar de las correlaciones entre matrices de distancias en el plano principal (n=50). *raw= distancia Euclidea sobre datos originales, rnr= distancia Euclidea sobre datos normalizados, rld= regularización logaritmo, voo= transformación logaritmo sobre conteros por millón, edg= distancia según edgeR, pea=correlación de Pearson, plg=correlación de Pearson sobre datos transformados por logaritmo, pln= correlación de Pearson sobre datos transformados por logaritmo y normalizados, spe=correlación de Spearman, gcc=correlación de Gini, chi=distancia chi-cuadrado, poi=disimilaridad Poisson. _m= escalamiento multidimensional métrico, _N escalamiento multidimensional no métrico.*

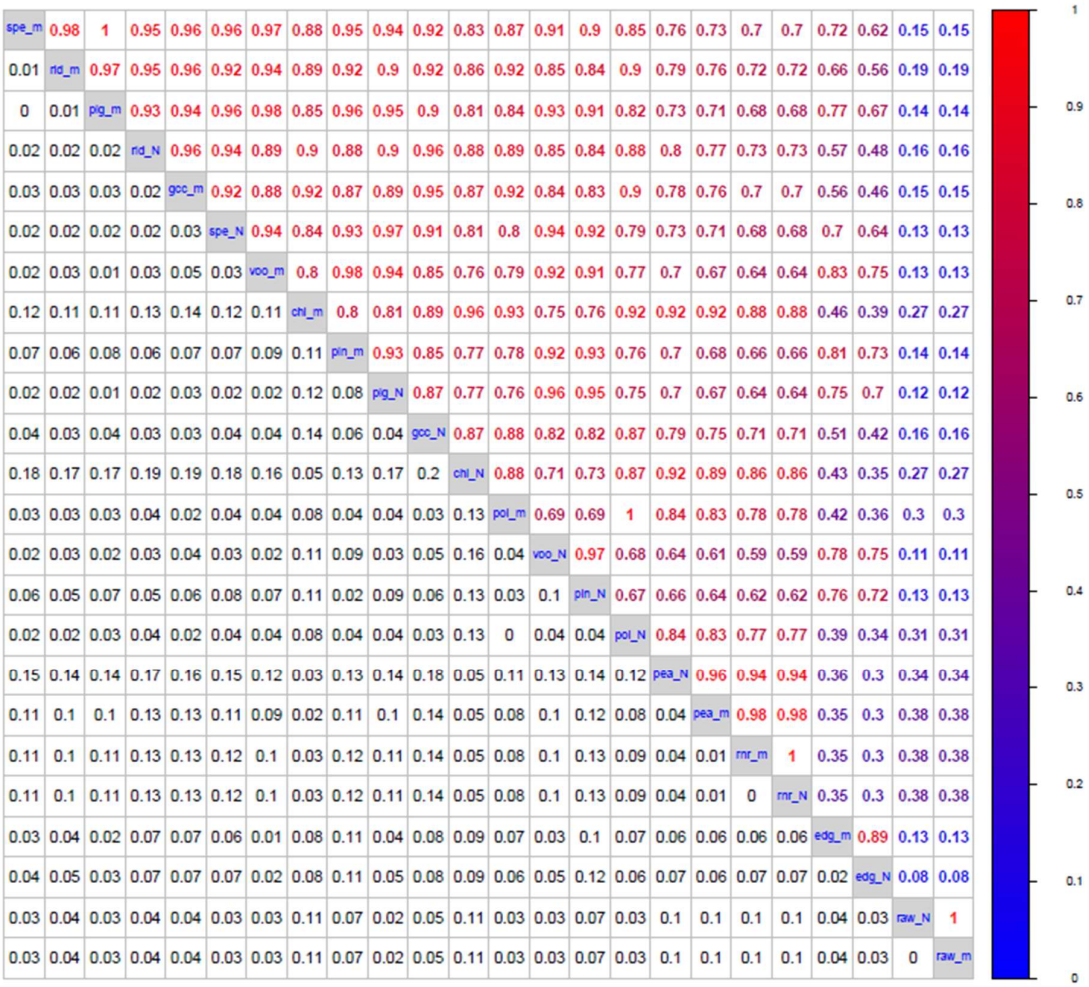

Figura 13. Concordancia entre disimilaridades en análisis de ordenación para el escenario $ED_{[10\%]}+nED_{[90\%]}$

Matriz de concordancia entre disimilaridades utilizando transcripts con y sin expresión diferencial. $ED_{[10\%]}+nED_{[90\%]}$. La parte superior contiene la media y la parte inferior el desvio estándar de las correlaciones entre matrices de distancias en el plano principal (n=50). *raw= distancia Euclidea sobre datos originales, rnr= distancia Euclidea sobre datos normalizados, rld= regularización logaritmo, voo= transformación logaritmo sobre conteros por millón, edg= distancia según edgeR, pea=correlación de Pearson, plg=correlación de Pearson sobre datos transformados por logaritmo, pln= correlación de Pearson sobre datos transformados por logaritmo y normalizados, spe=correlación de Spearman, gcc=correlación de Gini, chi=distancia chi-cuadrado, poi=disimilaridad Poisson. _m= escalamiento multidimensional métrico, _N escalamiento multidimensional no métrico.*

Disimilaridad	Consistencia	
	$ED_{[100\%]}$	$ED_{[10\%]}+nED_{[90\%]}$
raw_m	0.61 (0.19)	0.71 (0.14)
rnr_m	0.43 (0.23)	0.53 (0.18)
rld_m	0.96 (0.02)	0.85 (0.08)
voo_m	0.97 (0.01)	0.88 (0.06)
edg_m	0.97 (0.01)	0.94 (0.02)
pea_m	0.43 (0.22)	0.55 (0.18)
plg_m	0.96 (0.03)	0.84 (0.08)
pln_m	0.92 (0.17)	0.84 (0.12)
spe_m	0.95 (0.03)	0.82 (0.08)
gcc_m	0.90 (0.06)	0.76 (0.10)
chi_m	0.64 (0.22)	0.70 (0.14)
poi_m	0.87 (0.08)	0.77 (0.11)
raw_N	0.61 (0.19)	0.71 (0.14)
rnr_N	0.43 (0.23)	0.53 (0.18)
rld_N	0.93 (0.05)	0.77 (0.11)
voo_N	0.98 (0.04)	0.74 (0.10)
edg_N	0.99 (0.02)	0.80 (0.09)
pea_N	0.38 (0.26)	0.58 (0.21)
plg_N	0.98 (0.05)	0.76 (0.10)
pln_N	0.92 (0.18)	0.72 (0.14)
spe_N	0.92 (0.08)	0.77 (0.10)
gcc_N	0.86 (0.08)	0.76 (0.12)
chi_N	0.59 (0.26)	0.68 (0.20)
poi_N	0.84 (0.11)	0.76 (0.12)

Tabla 6 . Consistencia para disimilaridades en análisis de ordenación

Media y (desvío estándar) de las correlaciones entre matrices de distancia en el plano principal (n=50) para cada disimilaridad. *raw= distancia Euclidea sobre datos originales, rnr= distancia Euclidea sobre datos normalizados, rld= regularización logaritmo, voo= transformación logaritmo sobre conteros por millón, edg= distancia según edgeR, pea=correlación de Pearson, plg=correlación de Pearson sobre datos transformados por logaritmo, pln= correlación de Pearson sobre datos transformados por logaritmo y normalizados, spe=correlación de Spearman, gcc=correlación de Gini, chi=distancia chi-cuadrado, poi=disimilaridad Poisson. _m= escalamiento multidimensional métrico, _N escalamiento multidimensional no métrico.*

En la Figura 14 se observan las configuraciones típicas logradas por tres medidas de disimilaridad, *rld*, *rnr* y *raw,* una por cada uno de los 3 conjuntos encontrados en las Figuras Figura 12 y Figura 13, al aplicárseles EM métrico. Solo para *rld* se presentan además las configuraciones luego de aplicar EM no métrico ya que para las otras dos disimilaridades no hubo diferencias en los resultados entre ambos tipos de escalamiento. Se espera que las muestras se agrupen por su cercanía en el plano en función de los efectos de tratamiento y

celda de flujo. En el escenario 1 las configuraciones obtenidas con *rld_m* y *rld_N* dividen las muestras a lo largo de la primera dimensión en función de los tratamientos (A y B) y en función de las celdas de flujo a lo largo de la segunda dimensión, tal como se espera para este escenario. Se destaca que sobre el segundo eje, la celda de flujo 4 es notoriamente diferente al resto y a su vez las celdas de flujo 6 y 7 son relativamente similares. Por el contrario, para *rnr_m* y *raw_m* las configuraciones no se condicen con ningún patrón esperado. En el caso del escenario 2, las muestras se dividen principalmente por las celdas de flujo a lo largo de la primera dimensión según lo esperado. Solo *rld_m* y *rld_N* separan claramente la celda 4 y en oposición *raw_m* separa marcadamente a la celda 7.

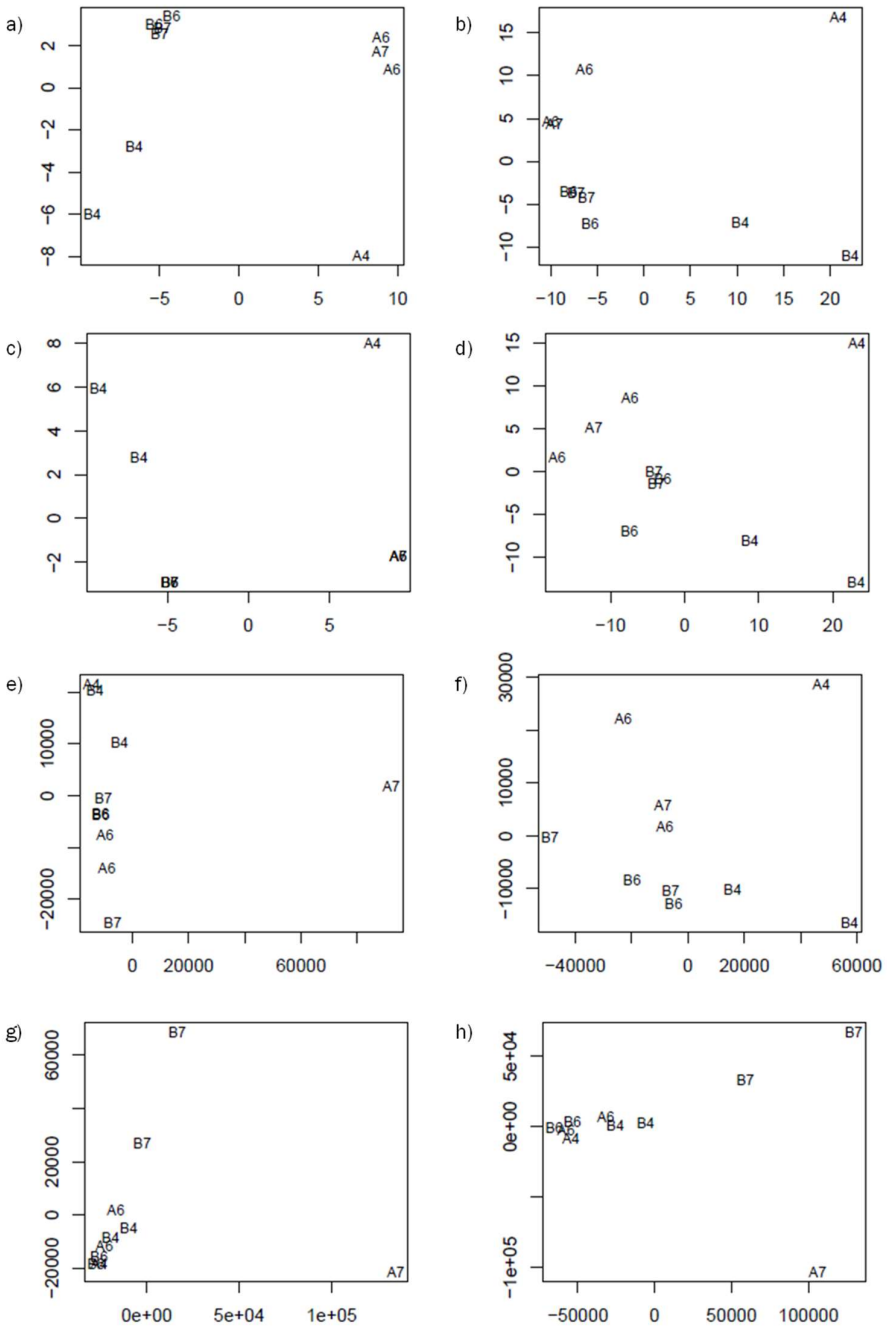

Figura 14. Gráfico de dispersión de observaciones utilizando escalas multidimensionales

Diagramas de dispersión típicos de las muestras en dos dimensiones obtenidos por: a) y b) escalamiento multidimensional métrico de la disimilaridad rld (*rld_m*); c) y d) escalamiento multidimensional no métrico de la disimilaridad rld (*rld_N*); e) y f) escalamiento multidimensional métrico de la disimilaridad rnr (*rnr_m*); g) y h) escalamiento multidimensional métrico de la disimilaridad raw (*raw_m*). En la izquierda están los diagramas obtenidos para el escenario 1, transcripts con expresión diferencial $ED_{[100\%]}$, y en la derecha los obtenidos para el escenario 2, transcripts con y sin expresión diferencial. $ED_{[10\%]}$+$nED_{[90\%]}$. Las etiquetas de la muestras corresponden al tratamiento principal (A o B) y a la celda de flujo (4, 6 o 7)

Disimilaridad	% explicación	
	$ED_{[100\%]}$	$ED_{[10\%]}+nED_{[90\%]}$
raw_m	95.23 (2.53)	92.89 (1.81)
rnr_m	92.74 (3.31)	85.47 (2.60)
rld_m	72.93 (2.86)	53.89 (1.78)
voo_m	65.99 (2.33)	42.21 (1.03)
edg_m	66.46 (2.41)	43.18 (1.61)
pea_m	72.03 (5.83)	70.65 (3.85)
plg_m	75.34 (2.49)	60.40 (3.41)
pln_m	67.44 (6.27)	44.92 (8.47)
spe_m	75.38 (2.15)	63.15 (3.64)
gcc_m	68.41 (4.54)	68.12 (2.41)
chi_m	74.11 (3.91)	73.34 (2.86)
poi_m	72.46 (3.34)	69.59 (2.23)

Tabla 7. Porcentaje de explicación en el plano principal

centaje de explicación obtenido con escalamiento multidimensional métrico en a disimilaridad). *raw= distancia Euclidea sobre datos originales, rnr= distancia s, rld= regularización logaritmo, voo= transformación logaritmo sobre conteros edgeR, pea=correlación de Pearson, plg=correlación de Pearson sobre datos = correlación de Pearson sobre datos transformados por logaritmo y e Spearman, gcc=correlación de Gini, chi=distancia chi-cuadrado, scalamiento multidimensional métrico.*

Disimilaridad	Stress [%]	
	$ED_{[100\%]}$	$ED_{[10\%]}+nED_{[90\%]}$
raw_N	2.35 (1.43)	2.64 (1.68)
rnr_N	2.64 (2.19)	4.20 (1.26)
rld_N	0.22 (0.76)	6.51 (1.11)
voo_N	0.33 (0.99)	8.48 (1.10)
edg_N	0.16 (0.81)	9.55 (1.31)
pea_N	0.68 (1.00)	1.33 (1.13)
plg_N	0.23 (0.98)	7.07 (1.16)
pln_N	0.61 (1.29)	8.06 (1.18)
spe_N	0.27 (0.75)	6.71 (1.19)
gcc_N	1.43 (2.02)	5.82 (1.77)
chi_N	1.11 (1.25)	3.74 (2.05)
poi_N	7.01 (1.96)	9.73 (1.60)

Tabla 8. Porcentaje de Stress

y (desvío estándar) de los valores de stress (%) según la fórmula de Kruskal (1964) obtenidos en 2 siones del escalamiento multidimensional no métrico (n=50 para cada disimilaridad). *raw= distancia ea sobre datos originales, rnr= distancia Euclidea sobre datos normalizados, rld= regularización logaritmo, ransformación logaritmo sobre conteros por millón, edg= distancia según edgeR, pea=correlación de on, plg=correlación de Pearson sobre datos transformados por logaritmo, pln= correlación de Pearson sobre transformados por logaritmo y normalizados, spe=correlación de Spearman, gcc=correlación de Gini, stancia chi-cuadrado, poi=disimilaridad Poisson. _N escalamiento multidimensional no métrico.*

4.1.4. Discusión del análisis de ordenación a dos vías

Los métodos de ordenación son utilizados con menos frecuencia que el análisis de cluster para la representación de resultados de RNA-seq. Una posible causa puede ser que suele ser de mayor interés representar resultados finales de genes o de genes y muestras a la vez por lo que el uso de análisis de clusters en forma de heatmaps constituye la herramienta doble propósito de preferencia. Otra posible causa es que la delimitación de grupos en las configuraciones de los métodos de ordenación es más arbitraria que la estructura determinada por análisis de conglomerados dado que la distancia entre individuos y la inclusión de un individuo en uno u otro grupo depende más de la experiencia y objetividad del investigador. No obstante, las técnicas de ordenación pueden aportar otra visualización complementaria para explorar y validar resultados.

Los procedimientos basados en *raw* no lograron representar ninguna configuración esperada (Figura 14 g y h). Al normalizar los datos, *rnr_m* se logró una configuración que disperso las muestras a lo largo de las dimensiones en función de celda de flujo y tratamientos aunque la consistencia observada fue muy baja (046 (0.23)). Valores similares de concordancia y consistencia se obtuvieron para *pea_m* que correspondería a la aplicación de un análisis de componentes principales sobre una matriz de correlaciones. Al igual que en análisis de conglomerados, la mera normalización o estandarización no logran una adecuada representación. La doble estandarización de la distancia chi-cuadrado, *chi*, tampoco arrojó una configuración adecuada en el escenario 1, aunque resultó adecuada en el escenario 2, donde predomina la influencia de los transcripts no diferenciados.

El procedimiento *edg_m,* corresponde a la propuesta actual por defecto implementada en edgeR, uno de los programas más utilizados para análisis diferencial de datos de RNA-seq. Al considerar solo los 500 transcripts más variables, el procedimiento realiza un filtrado antes de computar las disimilaridades y lograr una representación muy similar a la obtenida por *rld_m* (0.98 (0.01)) en el escenario 1. Sin embargo este mismo filtro aplicado en el escenario 2 no es adecuado para representar ninguna asociación en las muestras (0.52 (0.07)). Cabe mencionar

que también pueden calcularse las disimilaridades en base a comparación de a pares entre muestras y también se puede incluir la información de la matriz de diseño experimental al efectuar la normalización con lo que podrían obtener resultados diferentes. No se consideró la información del diseño experimental en el cálculo de ésta ni de otras disimilaridades puesto que el enfoque principal de la tesis es el análisis exploratorio y que la inclusión de esta información podría alterar la comparación de las disimilaridades.

Las disimilaridades que incluyen transformación (*plg*, *pln, voo)* o regularización logaritmo (*rld*), o consideran correlaciones de rangos (*spe*, *gcc*) generaron configuraciones esperadas y semejantes (concordancia > 0.80) a las obtenidas por *rld_m* (Figura 14a y b). Además, todas fueron consistentes en el escenario 1 (>0.85), pero sólo *rld_m, voo_m* y *edg* lo fueron en el escenario 2. La disimilaridad *poi*, genero una configuraciones esperadas y similares a la presentada por *rld_m* (concordancia >0.88) pero al igual que las anteriores fue consistente (>0.81) solo en el escenario 1.

Las configuraciones obtenidas por escalamiento métrico o no métrico para una misma disimilaridad fueron cercanas para todas las disimilaridades (media [0.89 a 1]) por lo que las distancias entre muestras en el plano se preservaron en forma similar bajo ambos procedimientos y en ambos escenarios. Los valores de stress de las versiones de EM no métrico (Tabla 8) pueden ser considerados como de excelentes a buenos según la escala propuesta por Kruskal (Kruskal 1964) ya que siempre fueron inferiores a 10%, lo que indica que hay una buena relación monótona entre las disimilaridades y las distancias representadas. Los porcentajes de explicación alcanzados por las versiones métricas fueron en general superiores al 65% y se consideran valores elevados dado el gran número de variables involucradas. En consecuencia, ambos métodos pueden ser utilizados satisfactoriamente.

4.2. Análisis multivariado a tres vías

4.2.1. Resultados del análisis multivariado a 3 vías

Las configuraciones consenso del análisis multifactorial múltiple dual se presentan en la Figura 15. Los gráficos de dispersión corresponden a la configuración típica utilizando la disimilaridad *rld* (Figura 15 a y b) o *voo* (Figura 15 c-f). A su vez, los gráficos (Figura 15 a-d) están basados en plasmodios generados en el escenario 1 mientras que los gráficos de la última fila (Figura 15 e y f) fueron obtenidos para plasmodios del escenario 2.

La concordancia entre las configuraciones obtenidas con *rld* y *voo* fueron mayores a 0.99 tanto cuando se utilizó como tercera vía de análisis la celda de flujo o el tratamiento. Las consistencias de las combinaciones entre disimilaridades, escenarios y tercera vía de análisis fueron elevadas (>0.80) (Tabla 9) salvo para *voo* utilizando la celda de flujo como tercera vía en el escenario 2 (0.36 (0.15)).

Disimilaridad	3ra vía	Consistencia	
		$ED_{[100\%]}$	$ED_{[10\%]}+nED_{[90\%]}$
rld	Celda de flujo	0.91 (0.05)	-
voo	Celda de flujo	0.90 (0.05)	0.36 (0.15)
rld	Tratamiento	0.83 (0.00)	-
voo	Tratamiento	0.82 (0.09)	0.82 (0.09)

Tabla 9. Consistencia para disimilaridades en el análisis multifactorial múltiple dual

Media y (desvío estándar) (n=50) de las coeficientes RV entre configuraciones en el plano principal consenso para cada disimilaridad. *rld= regularización logaritmo, voo= transformación logaritmo sobre conteos por millón.*

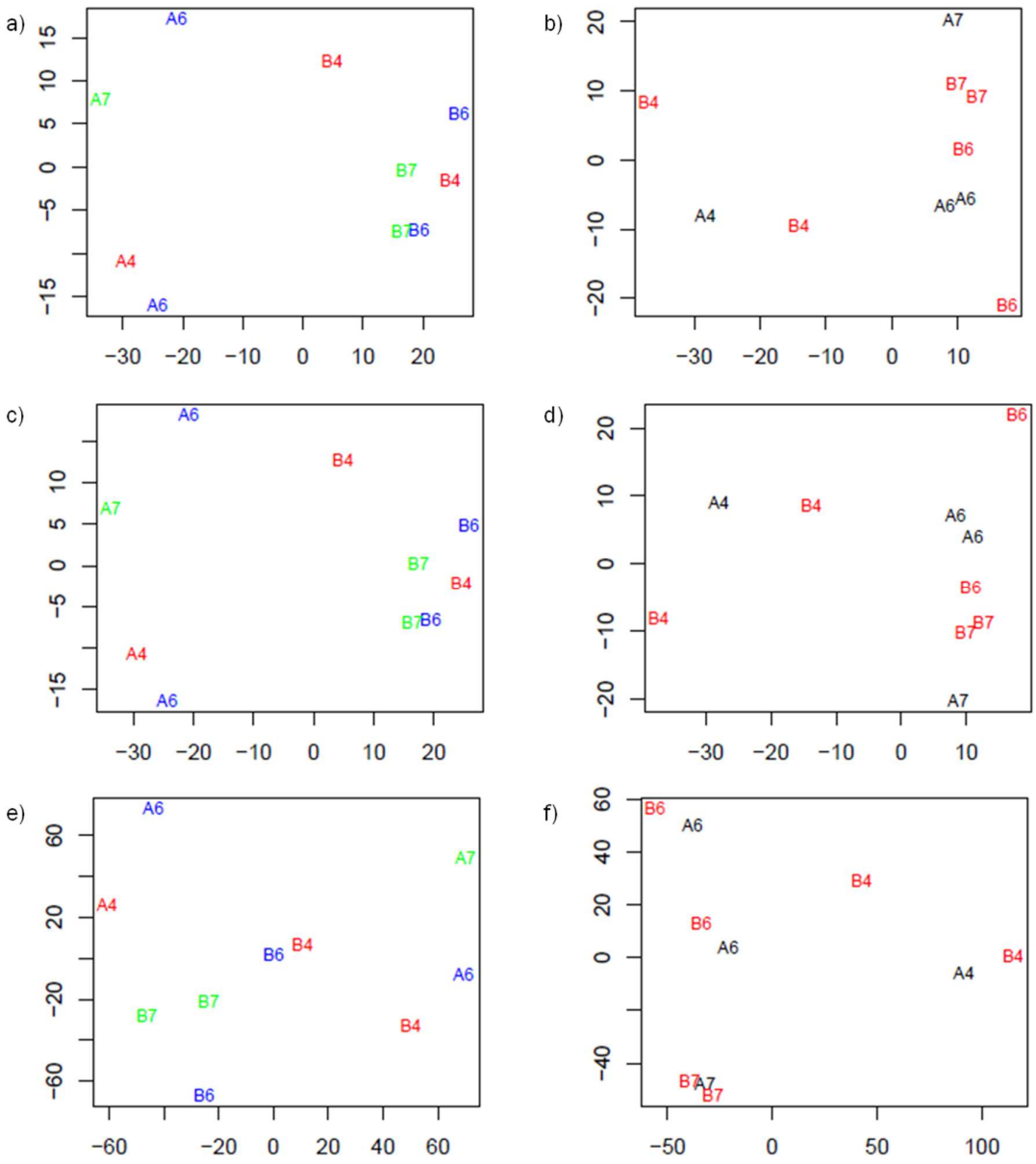

Figura 15. Gráfico de dispersión de observaciones utilizando análisis factorial múltiple dual

Diagramas de dispersión típicos de las muestras en dos dimensiones obtenidos por análisis factorial múltiple dual basados en: a) y b) disimilaridad *rld* para el escenario 1; c) y d) disimilaridad *voo* para el escenario 1; e) y f) disimilaridad *voo* para el escenario 2. En la izquierda están las configuraciones consenso utilizando como tercera vía el agrupamiento de muestras por flowcell y en la derecha por tratamiento. Las etiquetas de la muestras corresponden al tratamiento principal (A o B) y a la celda de flujo (4, 6 o 7) y los colores a los grupos utilizados en la tercera vía.

Disimilaridad	3ra vía	% explicación	
		$ED_{[100\%]}$	$ED_{[10\%]}+nED_{[90\%]}$
rld	Celda de flujo	65.73 (2.96)	-
voo	Celda de flujo	65.73 (3.02)	40.13 (1.03)
rld	Tratamiento	49.70 (1.20)	-
voo	Tratamiento	49.57 (1.20)	49.59 (1.06)

Tabla 10. Porcentaje de explicación en el plano principal (DMFA)

Media y (desvío estándar) (n=50) del porcentaje de explicación obtenido con análisis multifactorial múltiple dual en el plano principal consenso. *rld= regularización logaritmo, voo= transformación logaritmo sobre conteros por millón.*

En la Tabla 10 se muestran los porcentajes de explicación en el plano principal consenso. Se obtuvieron valores mayores (>65%) al utilizar la celda de flujo como tercera vía de clasificación en el escenario 1, mientras que el resto presentó valores cercanos a 50%.

4.2.2. Discusión análisis factorial múltiple dual

El análisis factorial múltiple es una generalización del ACP (Abdi, Williams, and Valentin 2013) que permite dividir el análisis global en grupos y luego comparar las configuraciones y/o variables de los grupos en un espacio consenso. Por lo tanto solo medidas que resulten adecuadas para representar los datos en un ACP pueden utilizarse también para el AFM. En particular, el ACP asume correlaciones lineales entre las variables. Las dos medidas que mejores resultados generaron en el análisis a dos vías, *rld* y *voo*, fueron utilizadas para adicionar una tercera vía de análisis. Las transformaciones utilizadas para su cálculo fueron propuestas con objetivos de mejorar la estimación de variabilidad (Love, Huber, and Anders 2014) y de factores de pesos para modelación lineal (Law et al. 2014) de expresión diferencial en RNA-seq.

Los resultados obtenidos utilizando como tercera vía el agrupamiento por celdas de flujo o tratamiento son naturalmente diferentes ya que los 3 o 2 ACP en los que se particional el análisis respectivamente, estandarizan los grupos de acuerdo a su propia estructura. Ambas estrategias pueden ser útiles al indagar la estructura de

las muestras. Al utilizar como tercera vía la celda de flujo, se observa que en todos las celdas de flujo la tendencia es a separar los tratamientos A y B (Figura 15 a y c), mientras que al utilizar como tercera vía el tratamiento, las muestras correspondientes a las celdas 6 y 7 son más parecidas entre sí, oponiéndose a la celda 4 a lo largo de la primera dimensión.

Una limitante constituye la representación de las variables dado que es imposible representar las 952 o 9500 variables (transcripts) de los escenarios 1 y 2 respectivamente, en un círculo unitario clásico o en un biplot. Sin embargo las matrices de cada grupo podrían ser explorada con otras técnicas de análisis para complementar el análisis como un análisis de cluster o conociendo grupos de variables asociadas, por ejemplo a una ruta metabólica, se puede asociar un análisis multifactorial jerárquico (Abdi, Williams, and Valentin 2013). De esta manera se disminuye la dimensionalidad y se exploran las relaciones entre grupos de variables.

Otra limitante importante es la eficiencia de cálculo y demanda de memoria en la aplicación del algoritmo utilizado. Si bien se pudo ejecutar el análisis para el escenario 1 a ambas transformaciones, sólo la transformación *voo* fue posible *de* emplear en el escenario 2. En este último caso, se debió recurrir a un centro de alta performance, requiriendo el análisis de cada plasmodio 12 Gb de memoria y un tiempo promedio de 12.4 horas para cada plasmodio. El filtrado previo constituye aquí otro punto importante a considerar para mejorar la aplicación del AFM.

4.3. Discusión general

Las medidas de disimilaridad evaluadas presentaron resultados similares en referencia a los agrupamientos esperados en la generación de los plasmodios. Es decir, las disimilaridades que lograron separar las muestras por tratamientos y/o celda de flujo en la estructura de dendrogramas también lo hicieron en las configuraciones de las técnicas de ordenación.

Tanto el análisis de cluster jerárquico como las técnicas de ordenación empleadas pueden ser empleados sin asumir distribución de probabilidades en las variables utilizadas. Sin embargo, la aproximación de las variables a una distribución de tipo Gaussiana ha sido reportada como beneficiosa para obtener resultados adecuados (Liu and Si 2014; Law et al. 2014). Esto se debe a que las características de simetría y relación media-varianza constante facilitan los procesos de estandarización y la utilización de la distancia Euclidea. Cuando las variables son fuertemente asimétricas y con rangos muy amplios como en el caso de los conteos en experimentos de RNA-seq (Simon Anders et al. 2013; Zhang, Pounds, and Tang 2013) el uso de la distancia Euclidea o estandarización por media y varianza no son adecuadas, tal como se observó con el uso de *raw*, *pea* y *rnr* en ambas técnicas. Una alternativa ampliamente utilizada tradicionalmente en datos de microarreglos (Bryan 2004; Jiang, Tang, and Zhang 2004) consiste en el uso de funciones que transforman o regularizan los datos. Todas las propuestas publicadas (*rld*, *voo*, *edg*, *gcc*, *psv*) (Love, Huber, and Anders 2014; Law et al. 2014; Robinson, McCarthy, and Smyth 2010; Ma and Wang 2012; S. Lee et al. 2013) para datos de RNA-seq mejoraron las representaciones logradas. El uso de la correlación de Spearman (*spe*) también logro buenos resultados ya que preserva el orden de rangos y es menos influenciada por outliers (Kendall and Gobbons 1990). Por último, otra estrategia consiste en utilizar medidas propias de conteos. En este sentido, la distancia chi cuadrado (*chi*) es muy usada en datos de frecuencias, pero no resulto adecuada en los datos evaluados. Posiblemente porque se trata esencialmente de una distancia Euclidea estandarizada por marginales fila y columna. Contrariamente, la disimilaridad basada en un modelo Poisson (*poi*) que integra normalización y sobredispersion de los conteos (Witten 2011) resulto de utilidad.

La consideración conjunta de las medidas propuestas de concordancia y consistencia permitieron evaluar la adecuación de las disimilaridades en ambos escenarios. En el análisis de cluster, únicamente las disimilaridades *rld*, *voo*, *plg* y *spe* resultaron concordantes con las estructuras jerárquicas esperadas y fueron consistentes en ambos escenarios. En el análisis de ordenación, esta lista se reduce a *rld_m* y *voo_m*. Nótese que para algunas propuestas como *edg*, *poi* o *gcc* pueden implementarse con otras técnicas de normalización o formas de calcular

las disimilaridades que podrían mejorar las representaciones. Por ejemplo, se podría combinar la disimilaridad Poisson con otra técnica de normalización o estimación de sobredispersión. Cabe mencionar que estrategias de uso conjunto de técnicas de ordenación y clustering pueden mejorar la representación (Chae and Warde 2006).

Según los objetivos de estudio, es posible elegir una u otra disimilaridad. Si el objetivo es representar resultados luego de un análisis de expresión diferencial, existe una mayor cantidad de disimilaridades que se pueden usar convenientemente. Esto se debe a que una vez filtrados los transcripts, las medidas son menos influenciadas por el ruido de factores externos a los principales. Por el contrario cuando, el objetivo es explorar la configuración de muestras sin filtros, solo *rld* y *voo* fueron aptas tanto en análisis de conglomerados como análisis de escalas multidimensionales métrico.

El AFM aplicado a los datos transformados por *rld* o *voo* provee una herramienta para explorar los resultados en función de una tercera vía de clasificación. Si bien la estructura de muestras pueden ser fácilmente analizada, la representación de las variables requiere de filtrado y/o agrupamiento para poder ser analizadas. La implementación en un contexto sin filtrado previo no fue posible o demandó recursos informáticos más allá de los disponibles en una computadora de escritorio.

El uso de plasmodios provee una plataforma de comparación complementaria al uso clásico de simulaciones y proporciona escenarios de variabilidad basados en datos reales de RNA-seq (Reeb and Steibel 2013). Si bien aquí se presentan plasmodios basados en un experimento, resultados similares fueron reportados con este (Reeb, Steibel, and Bramardi 2013) y otro experimento (Reeb, Bramardi, and Steibel 2015) aplicados al análisis de cluster jerárquico. La utilidad de los plasmodios como complemento a otras formas de evaluación (X. Zhou, Lindsay, and Robinson 2014; Vaughan et al. 2009) se incrementará a futuro a medida que estén disponibles públicamente más conjuntos de datos, facilitando la reproducción de la evaluación bajo diferentes condiciones (Reeb and Steibel 2013).

Capítulo 5: CONCLUSIONES

5.1. Conclusiones

La comparación de resultados de técnicas estadísticas requiere la utilización de datos de referencia en los que se conoce la estructura esperada con anticipación. En datos de alta dimensionalidad, como los de experimentos de secuenciación de ARN, es muy difícil que se logre una simulación paramétrica que contemple la complejidad de estos datos. En este trabajo se empleó la técnica de plasmodios para generar datos sintéticos con estructura conocida bajo dos escenarios posibles de análisis que permitieron evaluar el uso de disimilaridades en técnicas de análisis multivariado a dos y tres vías.

La representación con técnicas de análisis multivariado no supervisado de datos generados en RNA-seq requiere la transformación de los conteos originales. Las transformaciones basadas en logaritmo rescatan la estructura natural de las muestras tanto en análisis de cluster como en escalamiento multidimensional. Además, las funciones basadas en logaritmo que regularizan contemplando las diferencias de variabilidad a lo largo del rango de conteo, como *rld*, proveen resultados más consistentes. La disimilaridad basada en correlación de Spearman, *spe*, y la disimilaridad Poisson, también son alternativas válidas para representar muestras ya que en ambas se contempla indirectamente la transformación de los conteos. La utilización de disimilaridades que utilizan estandarización o normalización por sí solas no generan representaciones confiables.

La elección de una medida adecuada debe considerar el escenario a representar en términos de la cantidad de la relación señal-ruido ya que no totas las medidas muestran la configuración natural de las muestras al filtrar o no transcripts por ejemplo por su expresión diferencial.

Existe un número de disimilaridades concordantes y consistentes para ser empleadas y la preferencia por una u otra puede basarse en los paquetes estadísticos de preferencia del usuario.

En el análisis de cluster jerárquico, las disimilaridades *rld*, *voo*, *spe*, *plg* fueron consistentes tanto en el escenario que contempló sólo transcripts diferenciados como en el que contenía además transcripts no diferenciados.

En el escalamiento multidimensional, las configuraciones obtenidas con los métodos métricos y no métricos fueron muy similares, por lo que se recomienda utilizar la versión métrica que resulta de más fácil interpretación a los investigadores. Las disimilaridades *rld*, *voo*, *plg*, *spe* y *pln* fueron consistentes en ambos escenarios.

El análisis multivariado a tres vías utilizado, análisis factorial múltiple dual, fue concordante y consistente utilizando las disimilaridades *rld* y *voo* en el escenario de transcripts diferenciados. Por el contrario, en el escenario de transcripts diferenciados y no diferenciados, surgen problemas de recursos informáticos y de los propios algoritmos de los paquetes utilizados que limitan su uso. Además la consistencia de las configuraciones depende de la variabilidad de los individuos y grupos en la tercera vía de análisis.

Se concluye que en general el uso de las disimilaridades *rld*, *voo*, *plg* y *spe* es eficientes para representar datos de RNA-seq con fines exploratorios como control de calidad o para ilustrar resultados de la docimasia de hipótesis de expresión diferencial.

5.2. Trabajos futuros

Los resultados presentados se focalizaron en la representación de individuos (muestras) pero resulta de igual interés representar las relaciones entre genes. La técnica de plasmodios también puede utilizarse con este fin. Las disimilaridades que resultaron consistentes en esta tesis también pueden evaluarse con este fin utilizando análisis factorial múltiple jerárquico, al agrupar grupos de genes por ejemplo según vías metabólicas. Este tipo

de análisis podría contribuir como una etapa exploratoria y complementaria para continuar el análisis de redes.

La adecuación de las disimilaridades estudiadas, habilita el uso de técnicas de integración de información con otros conjuntos de datos. Esta integración es motivo de una activa investigación en datos “omics” donde resulta esencial adicionar al análisis del transcriptoma resultados de metabolitos como proteínas (datos proteomicos).

Un campo de investigación que deberá continuar para facilitar la integración de resultados es el estudio de medidas de correlación adecuadas a la multidimensionalidad y no linealidad de las relaciones entre atributos, tales como el uso de información mutua (Speed 2011) y proporcionalidad (Lovell et al. 2014).

REFERENCIAS BIBLIOGRAFICAS

Abdi, Hervé. 2007. “RV Coefficient and Congruence Coefficient.” In *Encyclopedia of Measurement and Statistics*, edited by Neil Salkind, 849–53. Thousand Oaks, CA: Sage.

Abdi, Hervé, Lynne J. Williams, and Domininique Valentin. 2013. “Multiple Factor Analysis: Principal Component Analysis for Multitable and Multiblock Data Sets.” *Wiley Interdisciplinary Reviews: Computational Statistics* 5 (2): 149–79. doi:10.1002/wics.1246.

Adiconis, Xian, Diego Borges-Rivera, Rahul Satija, David S Deluca, Michele a Busby, Aaron M Berlin, Andrey Sivachenko, et al. 2013. “Comparative Analysis of RNA Sequencing Methods for Degraded or Low-Input Samples.” *Nature Methods*, no. may (May). doi:10.1038/nmeth.2483.

Allison, David B., Peter M. Visscher, Guilherme J.M. Rosa, and Christopher I. Amos. 2009. “Statistical Genetics & Statistical Genomics: Where Biology, Epistemology, Statistics, and Computation Collide.” *Computational Statistics & Data Analysis* 53 (5). Elsevier B.V.: 1531–34. doi:10.1016/j.csda.2009.01.005.

Anders, S, and W Huber. 2010. “Differential Expression Analysis for Sequence Count Data.” *Genome Biology* 11 (10): R106. doi:10.1186/gb-2010-11-10-r106.

Anders, Simon. 2013. “HTSeq: Analysing High-Throughput Sequencing Data with Python.” http://www-huber.embl.de/users/anders/HTSeq/doc/overview.html.

Anders, Simon, Davis J McCarthy, Yunshun Chen, Michal Okoniewski, Gordon K Smyth, Wolfgang Huber, and Mark D Robinson. 2013. “Count-Based Differential Expression Analysis of RNA Sequencing Data Using R and Bioconductor.” *Nature Protocols* 8 (9): 1765–86. doi:10.1038/nprot.2013.099.

Anders, Simon, Alejandro Reyes, and Wolfgang Huber. 2012. “Detecting Differential Usage of Exons from RNA-Seq Data.” *Genome Research*, June, 2008–17. doi:10.1101/gr.133744.111.

Auer, P L, and R W Doerge. 2010. “Statistical Design and Analysis of RNA Sequencing Data.” *Genetics* 185 (2): 405–16. doi:10.1534/genetics.110.114983.

Bishop, Yvonne M. M., Stephen E. Fienberg, and Paul W. Holland. 2007. *Discrete Multivariate Analysis. Theory and Practice*. New York, New York, USA: Springer.

Blekhman, Ran, John C Marioni, Paul Zumbo, Matthew Stephens, and Yoav Gilad. 2010. “Sex-Specific and Lineage-Specific Alternative Splicing in Primates.” *Genome Research* 20 (2): 180–89. doi:10.1101/gr.099226.109.

Borodina, Tatiana, James Adjaye, and Marc Sultan. 2011. “A Strand-Specific Library Preparation Protocol for RNA Sequencing.” In *Methods in Enzymology*, 500:79–98. doi:10.1016/B978-0-12-385118-5.00005-0.

Bottomly, Daniel, Nicole A R Walter, Jessica Ezzell Hunter, Priscila Darakjian, Sunita Kawane, Kari J Buck, Robert P Searles, Michael Mooney, Shannon K McWeeney, and Robert Hitzemann. 2011. “Evaluating Gene Expression in C57BL/6J and DBA/2J Mouse Striatum Using RNA-Seq and Microarrays.” *PLoS ONE* 6 (3). Public Library of Science: e17820. doi:10.1371/journal.pone.0017820.

Bryan, Jenny. 2004. “Problems in Gene Clustering Based on Gene Expression Data.” *Journal of Multivariate Analysis* 90 (1): 44–66. doi:10.1016/j.jmva.2004.02.011.

Bullard, J H, E Purdom, K D Hansen, and S Dudoit. 2010. “Evaluation of Statistical Methods for Normalization and Differential Expression in mRNA-Seq Experiments.” *BMC Bioinformatics* 11: 94.

Camacho-Sanchez, Miguel, Pablo Burraco, Ivan Gomez-Mestre, and Jennifer a. Leonard. 2013. “Preservation of RNA and DNA from Mammal Samples under Field Conditions.” *Molecular Ecology Resources*, April. doi:10.1111/1755-0998.12108.

Carroll, J D, and P Arabie. 1980. “Multidimensional Scaling.” *Annual Review of Psychology* 31 (January): 607–49. doi:10.1146/annurev.ps.31.020180.003135.

Cattell, Raymond B. 1966. “The Scree Test For The Number Of Factors.” *Multivariate Behavioral Research* 1 (2). Routledge: 245–76. doi:10.1207/s15327906mbr0102_10.

Cattell, Raymond B, and Joseph Jaspers. 1967. “General Plasmode No. 30-10-5-2 for Factor Analytic Exercises and Research.” *Multivariate Behavioral Research*.

Chae, Seong S., and William D. Warde. 2006. “Effect of Using Principal Coordinates and Principal Components on Retrieval of Clusters.” *Computational Statistics & Data Analysis* 50 (6): 1407–17. doi:10.1016/j.csda.2005.01.013.

Costea, Paul I, Georg Zeller, Shinichi Sunagawa, and Peer Bork. 2014. “A Fair Comparison.” *Nature Methods* 11 (4). Nature Publishing Group: 359–359. doi:10.1038/nmeth.2897.

Dalton, Lori, Virginia Ballarin, and Marcel Brun. 2009. “Clustering Algorithms: On Learning, Validation, Performance, and Applications to Genomics.” *Current Genomics* 10 (6): 430–45. doi:10.2174/138920209789177601.

De Tayrac, Marie, Sébastien Lê, Marc Aubry, Jean Mosser, and François Husson. 2009. “Simultaneous Analysis of Distinct Omics Data Sets with Integration of Biological Knowledge: Multiple Factor Analysis Approach.” *BMC Genomics* 10 (ii): 32. doi:10.1186/1471-2164-10-32.

Dillies, M. A., A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, et al. 2012. “A Comprehensive Evaluation of Normalization Methods for Illumina High-Throughput RNA Sequencing Data Analysis.” *Briefings in Bioinformatics*, September. doi:10.1093/bib/bbs046.

Dinsdale, Elizabeth a, Robert a Edwards, Barbara a Bailey, Imre Tuba, Sajia Akhter, Katelyn McNair, Robert Schmieder, et al. 2013. “Multivariate Analysis of Functional Metagenomes.” *Frontiers in Genetics* 4 (April): 41. doi:10.3389/fgene.2013.00041.

Dua, Sumeet, and Pradeep Chowriappa. 2013. *Data Mining for Bioinformatics*. Boca Raton: CRC Press Taylor & Francis Group.

Ekblom, R, and J Galindo. 2011. “Applications of next Generation Sequencing in Molecular Ecology of Non-Model Organisms.” *Heredity* 107 (1). The Genetics Society: 1–15. doi:10.1038/hdy.2010.152.

Escofier, B., and J. Pagès. 1994. “Multiple Factor Analysis (AFMULT Package).” *Computational Statistics & Data Analysis* 18 (1): 121–40. doi:10.1016/0167-9473(94)90135-x.

Escofier, B., and Jérôme Pagès. 1988. “Analyses Factorielles Simples et Multiples: Objectifs, Methodes, Interpretations.” Paris.

Escoufier, Yves. 1973. “Le Traitement Des Variables Vectorielles.” *Biometrics* 29 (4): 751–60.

Ewing, B, and P Green. 1998. “Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities.” *Genome Research* 8 (3): 186–94.

Ewing, B., L. Hillier, M. C. Wendl, and P. Green. 1998. “Base-Calling of Automated Sequencer Traces Using Phred. I. Accuracy Assessment.” *Genome Research* 8 (3): 175–85. doi:10.1101/gr.8.3.175.

Fang, Zhide, Jeffery A Jeffrey Jeffery A Martin, and Zhong Wang. 2012. “Statistical Methods for Identifying Differentially Expressed Genes in RNA-Seq Experiments.” *Cell & Bioscience* 2 (1). BioMed Central Ltd: 26. doi:10.1186/2045-3701-2-26.

Fayyad, Usama, Gregory Piatetsky-shapiro, and Padhraic Smyth. 1996. “From Data Mining to Knowledge Discovery in,” 37–54.

Fontanillas, Pierre, Christian R Landry, Patricia J Wittkopp, Carsten Russ, Jonathan D Gruber, Chad Nusbaum, and Daniel L Hartl. 2010. “Key Considerations for Measuring Allelic Expression on a Genomic Scale Using High-Throughput Sequencing.” *Molecular Ecology* 19 Suppl 1 (March): 212–27. doi:10.1111/j.1365-294X.2010.04472.x.

Frazee, Alyssa, Ben Langmead, and Jeffrey Leek. 2011. “ReCount: A Multi-Experiment Resource of Analysis-Ready RNA-Seq Gene Count Datasets.” *BMC Bioinformatics* 12 (1): 449.

Gadbury, G L, Q Xiang, L Yang, S Barnes, G P Page, and D B Allison. 2008. “Evaluating Statistical Methods Using Plasmode Data Sets in the Age of Massive Public Databases: An Illustration Using False Discovery Rates.” *PLoS Genetics* 4 (6): e1000098. doi:10.1371/journal.pgen.1000098.

Gadbury, GaryL., KarenA. Garrett, and DavidB. Allison. 2009. “Challenges and Approaches to Statistical Design and Inference in High-Dimensional Investigations.” In *Plant Systems Biology SE - 9*, edited by Dmitry A Belostotsky, 553:181–206 LA –

English. Methods in Molecular Biology™. Humana Press. doi:10.1007/978-1-60327-563-7_9.

Garrett, Elizabeth S, Rafael A Irizarry, and Scott L Zeger. 2003. *The Analysis of Gene Expression Data. Methods and Software*. Springer.

Gehlenborg, Nils, Seán I O'Donoghue, Nitin S Baliga, Alexander Goesmann, Matthew a Hibbs, Hiroaki Kitano, Oliver Kohlbacher, et al. 2010. "Visualization of Omics Data for Systems Biology." *Nature Methods* 7 (3 Suppl). Nature Publishing Group: S56–68. doi:10.1038/nmeth.1436.

Givan, Scott A, Christopher A Bottoms, and William G Spollen. 2012. "Computational Analysis of RNA-Seq." In *RNA Abundance Analysis*, edited by Hailing Jin and Walter Gassmann, 883:201–19. Methods in Molecular Biology. Totowa, NJ: Humana Press. doi:10.1007/978-1-61779-839-9.

Gower, J C. 1966. "Some Distance Properties of Latent Root and Vector Used in Multivariate Analysis." *Biometrika* 53 (3): 325–38.

———. 1975. "Generalized Procrustes Analysis." *Psychometrika* 40 (1). Springer: 33–51. doi:10.1007/BF02291478.

Grabherr, Manfred G, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn a Thompson, Ido Amit, Xian Adiconis, et al. 2011. "Full-Length Transcriptome Assembly from RNA-Seq Data without a Reference Genome." *Nature Biotechnology* 29 (7): 644–52. doi:10.1038/nbt.1883.

Greenacre, Michael. 2010. *Biplots in Practice*. Barcelona: Fundacion BBVA.

Halkidi, Maria, Yannis Batistakis, and Michalis Vazirgiannis. 2001. "On Clustering Validation Techniques." *Journal of Intelligent Information Systems* 17 (2-3): 107–45.

Handl, Julia, Joshua Knowles, and Douglas B Kell. 2005. "Computational Cluster Validation in Post-Genomic Data Analysis." *Bioinformatics (Oxford, England)* 21 (15): 3201–12. doi:10.1093/bioinformatics/bti517.

Hassan, Musa a, and Jeroen P J Saeij. 2014. "Incorporating Alternative Splicing and mRNA Editing into the Genetic Analysis of Complex Traits." *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology*, August, 1–9. doi:10.1002/bies.201400079.

Hastie, Trevor, Robert Tibshirani, and Jerome H Friedman. 2009. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Second Edi. New York, New York, USA: Springer.

Hong, Shengjun, Xiangning Chen, Li Jin, and Momiao Xiong. 2013. "Canonical Correlation Analysis for RNA-Seq Co-Expression Networks." *Nucleic Acids Research*, March, 1–15. doi:10.1093/nar/gkt145.

Hornik, Kurt. 2005. "A CLUE for CLUster Ensembles." *Journal of Statistical Software* 14 (12).

Hotelling, H. 1933. “Analysis of a Complex of Statistical Variables into Principal Components.” *Journal of Educational Psychology* 24 (6). Warwick & York: 417–41. doi:10.1037/h0071325.

Husson, F., and J. Pagès. 2006. “INDSCAL Model: Geometrical Interpretation and Methodology.” *Computational Statistics & Data Analysis* 50 (2): 358–78. doi:10.1016/j.csda.2004.08.005.

Husson, François, Sebastien Le, and Jerome Pages. 2011. *Explorative Multivariate Analysis by Example Using R*. Boca Raton: CRC Press Taylor & Francis Group.

Illumina. 2010. *Illumina Sequencing Technology*. San Diego, Ca.

Izenman, A. J. 2008. *Modern Multivariate Statistical Techniques. Regression, Classification, and Manifold Learning*. New York, New York, USA: Springer.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning with Applications in R*. New York: Springer. doi:10.1007/978-1-4614-7138-7.

Jiang, Daxin, Chun Tang, and Aidong Zhang. 2004. “Cluster Analysis for Gene Expression Data: A Survey.” *IEEE Transactions on Knowledge and Data Engineering* 16 (11): 1370–86. doi:10.1109/TKDE.2004.68.

Johnson, Richard Arnold, and Dean W Wichern. 2002. *Applied Multivariate Statistical Analysis*. 5th ed. Upper Saddle River, N.J.: Prentice Hall.

Kendall, M G, and J D Gobbons. 1990. *Rank Correlation Methods*. *Science Forum*. 5th ed. Vol. 3. USA: Oxford University Press.

Kiers, Henk a. L. 1991. “Hierarchical Relations among Three-Way Methods.” *Psychometrika* 56 (3): 449–70. doi:10.1007/BF02294485.

———. 2000. “Towards a Standardized Notation and Terminology in Multiway Analysis.” *Journal of Chemometrics* 14 (3): 105–22. doi:10.1002/1099-128X(200005/06)14:3<105::AID-CEM582>3.0.CO;2-I.

Kroonenberg, P. M. 2008. *Applied Multiway Data Analysis*. New Jersey, USA: John Wiley & Sons, Inc.

Kruskal, J B. 1964. “MULTIDIMENSIONAL SCALING BY OPTIMIZING GOODNESS.” *Psychometrika* 29 (1).

Kumar, Ravi, Yasunori Ichihashi, Seisuke Kimura, Daniel H Chitwood, Lauren R Headland, Jie Peng, Julin N Maloof, and Neelima R Sinha. 2012. “A High-Throughput Method for Illumina RNA-Seq Library Preparation.” *Frontiers in Plant Genetics and Genomics* 3: 202. doi:10.3389/fpls.2012.00202.

Langmead, Ben, Kasper D Hansen, and Jeffrey T Leek. 2010. “Cloud-Scale RNA-Sequencing Differential Expression Analysis with Myrna.” *Genome Biology* 11 (8): R83. doi:10.1186/gb-2010-11-8-r83.

Langmead, Ben, and Steven L Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4). Nature Publishing Group: 357–60. doi:10.1038/nmeth.1923.

Law, Charity W, Yunshun Chen, Wei Shi, and Gordon K Smyth. 2014. “Voom: Precision Weights Unlock Linear Model Analysis Tools for RNA-Seq Read Counts.” *Genome Biology* 15 (2): R29. doi:10.1186/gb-2014-15-2-r29.

Le, Sebastien, Julie Josse, and François Husson. 2008. “FactoMineR: An R Package for Multivariate Analysis.” *Journal of Statistical Software* 25 (1): 1–18.

Lê, Sébastien, and Jérôme Pagès. 2010. “DMFA: Dual Multiple Factor Analysis.” *Communications in Statistics - Theory and Methods* 39 (3): 483–92. doi:10.1080/03610920903140114.

Lee, Jae-hyung, Jason K Ang, and Xinshu Xiao. 2013. “Analysis and Design of RNA Sequencing Experiments for Identifying RNA Editing and Other Single-Nucleotide Variants,” no. Strategy 1: 725–32. doi:10.1261/rna.037903.112.Park.

Lee, Seonjoo, Pauline E Chugh, Haipeng Shen, R Eberle, and Dirk P Dittmer. 2013. “Poisson Factor Models with Applications to Non-Normalized microRNA Profiling.” *Bioinformatics (Oxford, England)* 29 (9): 1105–11. doi:10.1093/bioinformatics/btt091.

Leng, Ning, John a Dawson, James a Thomson, Victor Ruotti, Anna I Rissman, Bart M G Smits, Jill D Haag, Michael N Gould, Ron M Stewart, and Christina Kendziorski. 2013. “EBSeq: An Empirical Bayes Hierarchical Model for Inference in RNA-Seq Experiments.” *Bioinformatics (Oxford, England)*, February, 1–9. doi:10.1093/bioinformatics/btt087.

Levin, Joshua Z, Moran Yassour, Xian Adiconis, Chad Nusbaum, Dawn Anne Thompson, Nir Friedman, Andreas Gnirke, and Aviv Regev. 2010. “Comprehensive Comparative Analysis of Strand-Specific RNA Sequencing Methods.” *Nature Methods* 7 (9). Nature Publishing Group: 709–15.

Li, B, V Ruotti, R M Stewart, J A Thomson, and C N Dewey. 2010. “RNA-Seq Gene Expression Estimation with Read Mapping Uncertainty.” *Bioinformatics* 26 (4): 493–500. doi:10.1093/bioinformatics/btp692.

Li, Jun, Daniela M Witten, Iain M Johnstone, and Robert Tibshirani. 2012. “Normalization, Testing, and False Discovery Rate Estimation for RNA-Sequencing Data.” *Biostatistics (Oxford, England)* 13 (3): 523–38. doi:10.1093/biostatistics/kxr031.

Liu, Peng, and Yaqing Si. 2014. “Cluster Analysis of RNA-Sequencing Data.” In *Statistical Analysis of Next Generation Sequencing Data SE - 10*, edited by Somnath Datta and Dan Nettleton, 191–217. Frontiers in Probability and the Statistical Sciences. Springer International Publishing. doi:10.1007/978-3-319-07212-8_10.

Love, Michael I., Simon Anders, and Wolfgang Huber. 2014. *Differential Analysis of Count Data - the DESeq2 Package*.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” *bioRxiv*, February. doi:10.1101/002832.

Lovell, David, Vera Pawlowsky-glahn, Juan José Egozcue, and Juan Jos. 2014. "Proportionality : A Valid Alternative to Correlation for Relative Data Proportionality." *bioRxiv*.

Ma, Chuang, and Xiangfeng Wang. 2012. "Application of the Gini Correlation Coefficient to Infer Regulatory Relationships in Transcriptome Analysis." *Plant Physiology* 160 (1): 192–203. doi:10.1104/pp.112.201962.

Marioni, J C, C E Mason, S M Mane, M Stephens, and Y Gilad. 2008. "RNA-Seq: An Assessment of Technical Reproducibility and Comparison with Gene Expression Arrays." *Genome Research* 18: 1509–17.

Martin, Jeffrey a, and Zhong Wang. 2011. "Next-Generation Transcriptome Assembly." *Nature Reviews. Genetics* 12 (10). Nature Publishing Group: 671–82. doi:10.1038/nrg3068.

Martín-Rodríguez, Jesús, Ma Purificación Galindo-Villardón, and José L. Vicente-Villardón. 2002. "Comparison and Integration of Subspaces from a Biplot Perspective." *Journal of Statistical Planning and Inference* 102 (2): 411–23. doi:10.1016/S0378-3758(01)00101-X.

Mayer, Claus-dieter, Julie Lorent, and Graham W Horgan. 2011. "Exploratory Analysis of Multiple Omics Datasets Using the Adjusted RV Coefficient Exploratory Analysis of Multiple Omics Datasets Using the Adjusted RV Coefficient ∗" 10 (1).

McGee, Victor E. 1968. "Multidimensional Scaling Of N Sets Of Similarity Measures: A Nonmetric Individual Differences Approach." *Multivariate Behavioral Research* 3 (2). Routledge: 233–48. doi:10.1207/s15327906mbr0302_7.

Mehta, T, M Tanik, and D B Allison. 2004. "Towards Sound Epistemological Foundations of Statistical Methods for High-Dimensional Biology." *Nature Genetics* 36 (9): 943–47. doi:10.1038/ng1422.

Mehta, Tapan S, Stanislav O Zakharkin, Gary L Gadbury, and David B Allison. 2006. "Epistemological Issues in Omics and High-Dimensional Biology: Give the People What They Want." *Physiological Genomics* 28 (1): 24–32. doi:10.1152/physiolgenomics.00095.2006.

Metzker, Michael L. 2010. "Sequencing Technologies - the next Generation." *Nature Reviews. Genetics* 11 (1). Nature Publishing Group: 31–46. doi:10.1038/nrg2626.

Morozova, O, and M A Marra. 2008. "Applications of next-Generation Sequencing Technologies in Functional Genomics." *Genomics* 92 (5): 255–64. doi:10.1016/j.ygeno.2008.07.001.

Morozova, Olena, Martin Hirst, and Marco A Marra. 2009. "Applications of New Sequencing Technologies for Transcriptome Analysis." *Annual Review of Genomics and Human Genetics* 10 (1). Annual Reviews: 135–51. doi:10.1146/annurev-genom-082908-145957.

Mortazavi, Ali, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. 2008. "Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq." *Nature*

Methods 5 (7). Nature Publishing Group: 621–28. doi:http://www.nature.com/nmeth/journal/v5/n7/suppinfo/nmeth.1226_S1.html.

Mutz, Kai-Oliver, Alexandra Heilkenbrinker, Maren Lönne, Johanna-Gabriela Walter, and Frank Stahl. 2012. "Transcriptome Analysis Using next-Generation Sequencing." *Current Opinion in Biotechnology*, September. doi:10.1016/j.copbio.2012.09.004.

Neafsey, Daniel E, and Brian J Haas. 2011. "'Next-Generation' Sequencing Becomes 'Now-Generation'." *Genome Biology* 12 (3): 303. doi:10.1186/gb-2011-12-3-303.

Nicolae, Marius, Serghei Mangul, Ion I Măndoiu, and Alex Zelikovsky. 2011. "Estimation of Alternative Splicing Isoform Frequencies from RNA-Seq Data." *Algorithms for Molecular Biology : AMB* 6 (1). BioMed Central Ltd: 9. doi:10.1186/1748-7188-6-9.

Nie, Lei, Gang Wu, David E. Culley, Johannes C. M. Scholten, and Weiwen Zhang. 2008. "Integrative Analysis of Transcriptomic and Proteomic Data: Challenges, Solutions and Applications," October. Informa UK Ltd UK.

Oshlack, A, and M J Wakefield. 2009. "Transcript Length Bias in RNA-Seq Data Confounds Systems Biology." *Biology Direct* 4: 14. doi:10.1186/1745-6150-4-14.

Oshlack, Alicia, Mark Robinson, and Matthew Young. 2010. "From RNA-Seq Reads to Differential Expression Results." *Genome Biology* 11 (12): 220.

Ozsolak, F, and P M Milos. 2011. "RNA Sequencing: Advances, Challenges and Opportunities." *Nature Reviews. Genetics* 12 (2): 87–98. doi:10.1038/nrg2934.

Pachter, Lior. 2011. "Models for Transcript Quantification from RNA-Seq." Genomics; Methodology. *ArXiv* 1104.3889 (April): 1–28.

Parkhomchuk, Dmitri, Tatiana Borodina, Vyacheslav Amstislavskiy, Maria Banaru, Linda Hallen, Sylvia Krobitsch, Hans Lehrach, and Alexey Soldatov. 2009. "Transcriptome Analysis by Strand-Specific Sequencing of Complementary DNA." *Nucleic Acids Research* 37 (18). Oxford University Press: e123.

Paulson, Joseph N, O Colin Stine, Héctor Corrada Bravo, and Mihai Pop. 2013. "Differential Abundance Analysis for Microbial Marker-Gene Surveys." *Nature Methods* 10 (12): 1200–1202. doi:10.1038/nmeth.2658.

R Development Core Team. 2014. "R: A Language and Environment for Statistical Computing." Vienna, Austria: R Foundation for Statistical Computing.

Rand, William M. 1971. "Objective Criteria for the Evaluation of Clustering Methods." *Journal of the American Statistical Association* 66 (336): 846–50. doi:10.1080/01621459.1971.10482356.

Rau, Andrea, Gilles Celeux, and Cathy Maugis-rabusseau. 2011. *Clustering High-Throughput Sequencing Data with Poisson Mixture Models*.

Raz, Tal, Philipp Kapranov, Doron Lipson, Stan Letovsky, Patrice M Milos, and John F Thompson. 2011. "Protocol Dependence of Sequencing-Based Gene Expression Measurements." *PLoS ONE* 6 (5). Public Library of Science: e19287.

Reeb, Pablo D, Sergio J Bramardi, and Juan P Steibel. 2015. “Assessing Dissimilarity Measures for Sample- Based Hierarchical Clustering of RNA Sequencing Data Using Plasmode Datasets.” *PloS One* 10 (7): 1–18. doi:10.1371/journal.pone.0132310.

Reeb, Pablo D., and Juan P. Steibel. 2013. “Evaluating Statistical Analysis Models for RNA Sequencing Experiments.” *Frontiers in Genetics* 4 (September): 1–9. doi:10.3389/fgene.2013.00178.

Reeb, Pablo D., Juan P. Steibel, and Sergio J. Bramardi. 2013. “Analisis de Agrupamiento Jerarquico Para Muestras de Secuenciacion de ARN.” In *IV Ecuentro Iberoamericano de Biometria*, 1–3. Mar del Plata, Argentina: Region Argentina de la International Biometric Society.

Reshef, David N, Yakir a Reshef, Hilary K Finucane, Sharon R Grossman, Gilean McVean, Peter J Turnbaugh, Eric S Lander, Michael Mitzenmacher, and Pardis C Sabeti. 2011. “Detecting Novel Associations in Large Data Sets.” *Science (New York, N.Y.)* 334 (6062): 1518–24. doi:10.1126/science.1205438.

Risso, Davide, Katja Schwartz, Gavin Sherlock, and Sandrine Dudoit. 2011. “GC-Content Normalization for RNA-Seq Data.” *BMC Bioinformatics* 12 (1): 480.

Robert, P, and Yves Escoufier. 1976. “A Unifying Tool for Linear Multivariate Statistical Methods : The RV-Coefficient.” *Appl. Statist.* 25 (3): 257–65.

Robinson, M D, D J McCarthy, and G K Smyth. 2010. “edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data.” *Bioinformatics* 26 (1): 139–40. doi:10.1093/bioinformatics/btp616.

Robinson, M D, and A Oshlack. 2010. “A Scaling Normalization Method for Differential Expression Analysis of RNA-Seq Data.” *Genome Biology* 11 (3): R25. doi:10.1186/gb-2010-11-3-r25.

Robinson, M D, and G K Smyth. 2007. “Moderated Statistical Tests for Assessing Differences in Tag Abundance.” *Bioinformatics* 23 (21): 2881–87. doi:10.1093/bioinformatics/btm453.

Romero, Maria del Carmen. 2009. “Seleccion de Atributos En Contextos de Alta Dimensionalidad.” Universidad Nacional de Cordoba.

Scholkopf, B., A. J. Smola, and K. R. Muller. 1998. “Nonlinear Component Analysis as a Kernel Eigenvalue Problem.” *Neural Computation* 10: 1299–1319.

Schwochow, Doreen, Laurel E K Serieys, Robert K Wayne, and Olaf Thalmann. 2012. “Efficient Recovery of Whole Blood RNA--a Comparison of Commercial RNA Extraction Protocols for High-Throughput Applications in Wildlife Species.” *BMC Biotechnology* 12 (1): 33. doi:10.1186/1472-6750-12-33.

Si, Yaqing. 2011. “Model-Based Clustering for RNA-Seq Data,” 1–26.

Skelly, Daniel A, Marnie Johansson, Jennifer Madeoy, Jon Wakefield, and Joshua M Akey. 2011. “A Powerful and Flexible Statistical Framework for Testing Hypotheses

of Allele-Specific Gene Expression from RNA-Seq Data." *Genome Research* 21 (10): 1728–37. doi:10.1101/gr.119784.110.

Sloutsky, Roman, Nicolas Jimenez, S Joshua Swamidass, and Kristen M Naegle. 2013. "Accounting for Noise When Clustering Biological Data." *Briefings in Bioinformatics* 14 (4): 423–36. doi:10.1093/bib/bbs057.

Speed, Terry. 2011. "Mathematics. A Correlation for the 21st Century." *Science (New York, N.Y.)* 334 (6062): 1502–3. doi:10.1126/science.1215894.

Steibel, Juan Pedro, Rosangela Poletto, Paul M Coussens, and Guilherme J M Rosa. 2009. "A Powerful and Flexible Linear Mixed Model Framework for the Analysis of Relative Quantification RT-PCR Data." *Genomics* 94 (2). Elsevier Inc.: 146–52. doi:10.1016/j.ygeno.2009.04.008.

Touw, Wouter G, Jumamurat R Bayjanov, Lex Overmars, Lennart Backus, Jos Boekhorst, Michiel Wels, and Sacha a F T van Hijum. 2012. "Data Mining in the Life Sciences with Random Forest: A Walk in the Park or Lost in the Jungle?" *Briefings in Bioinformatics*, July. doi:10.1093/bib/bbs034.

Trapnell, C, L Pachter, and S L Salzberg. 2009. "TopHat: Discovering Splice Junctions with RNA-Seq." *Bioinformatics* 25: 1105–11.

Trapnell, Cole, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, and Lior Pachter. 2012. "Differential Gene and Transcript Expression Analysis of RNA-Seq Experiments with TopHat and Cufflinks." *Nature Protocols* 7 (3). Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.: 562–78. doi:10.1038/nprot.2012.016.

Trapnell, Cole, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. 2010. "Transcript Assembly and Quantification by RNA-Seq Reveals Unannotated Transcripts and Isoform Switching during Cell Differentiation." *Nature Biotechnology* 28 (5). Nature Publishing Group: 511–15. doi:10.1038/nbt.1621.

Tucker, Ledyard R. 1966. "Some Mathematical Notes on Three-Mode Factor Analysis." *Psychometrika* 31 (3): 279–311.

Tukey, John Wilder. 1977. *Exploratory Data Analysis*. Addison-Wesley.

Twyman, Richard. 2002. "What Are 'Model Organisms'?" *Wellcome Trust*. http://genome.wellcome.ac.uk/doc_WTD020803.html.

Van De Wiel, Mark a, Gwenaël G R Leday, Luba Pardo, Håvard Rue, Aad W Van Der Vaart, and Wessel N Van Wieringen. 2013. "Bayesian Analysis of RNA Sequencing Data by Estimating Multiple Shrinkage Priors." *Biostatistics (Oxford, England)*, September, 113–28. doi:10.1093/biostatistics/kxs031.

Van Verk, Marcel C., Richard Hickman, Corné M.J. Pieterse, and Saskia C.M. Van Wees. 2013. "RNA-Seq: Revelation of the Messengers." *Trends in Plant Science*, March, 175–79. doi:10.1016/j.tplants.2013.02.001.

Vapnik, Vladimir Naumovich. 1998. *Statistical Learning Theory*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. New York: Wiley.

Vaughan, L K, J Divers, M Padilla, D T Redden, H K Tiwari, D Pomp, and D B Allison. 2009. “The Use of Plasmodes as a Supplement to Simulations: A Simple Example Evaluating Individual Admixture Estimation Methodologies.” *Computational Statistics & Data Analysis* 53 (5): 1755–66. doi:10.1016/j.csda.2008.02.032.

Waller, Niels G, J Michael Underhill, and A Heather. 2010. “Multivariate Behavioral A Method for Generating Simulated Plasmodes and Artificial Test Clusters with User-Defined Shape , Size , and,” no. June 2013: 37–41.

Wang, N., Y. Wang, H. Hao, L. Wang, Z. Wang, J. Wang, and R. Wu. 2013. “A Bi-Poisson Model for Clustering Gene Expression Profiles by RNA-Seq.” *Briefings in Bioinformatics*, May. doi:10.1093/bib/bbt029.

Wang, Zhong, Mark Gerstein, and Michael Snyder. 2009. “RNA-Seq: A Revolutionary Tool for Transcriptomics.” *Nat Rev Genet* 10 (1). Nature Publishing Group: 57–63. doi:10.1038/nrg2484.

Wei, Dan, Qingshan Jiang, Yanjie Wei, and Shengrui Wang. 2012. “A Novel Hierarchical Clustering Algorithm for Gene Sequences.” *BMC Bioinformatics* 13 (1): 174. doi:10.1186/1471-2105-13-174.

Williams, C. K. I. 2001. “On a Connection between Kernel PCA and Metric Multidimensional Scaling.” In *Advances in Neural Information Processing Systems 13*, edited by T. K. Leen, T. G. Dietterich, and V. Treps. Cambridge, MA: MIT Press.

Witten, Daniela M. 2011. “Classification and Clustering of Sequencing Data Using a Poisson Model.” *The Annals of Applied Statistics* 5 (4). The Institute of Mathematical Statistics: 2493–2518. doi:10.1214/11-AOAS493.

Xiong, Hui, and Zhongmou Li. 2013. “Clustering Validation Measures.” In *Data Clustering*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. Chapman and Hall/CRC. doi:doi:10.1201/b15410-24.

Xu, Yanxun, Juhee Lee, Yuan Yuan, Riten Mitra, Shoudan Liang, and M Peter. 2013. “Nonparametric Bayesian Bi-Clustering for Next,” no. 2: 1–22.

Yitzhaki, Shlomo, and Edna Schechtman. 2013. *The Gini Methodology: A Primer on a Statistical Methodology*. *Springer Series in Statistics. New York and Heidelberg: Springer, 2013, Pp. Xvi, 548*. Springer Series in Statistics. New York and Heidelberg: Springer.

Zapala, Matthew a., and Nicholas J. Schork. 2012. “Statistical Properties of Multivariate Distance Matrix Regression for High-Dimensional Data Analysis.” *Frontiers in Genetics* 3 (September): 1–10. doi:10.3389/fgene.2012.00190.

Zhang, Hui, Stanley B Pounds, and Li Tang. 2013. “Statistical Methods for Overdispersion in mRNA-Seq Count Data,” 34–40.

Zhong, Silin, Je-Gun Joung, Yi Zheng, Yun-ru Chen, Bao Liu, Ying Shao, Jenny Z Xiang, Zhangjun Fei, and James J Giovannoni. 2011. “High-Throughput Illumina Strand-

Specific RNA Sequencing Library Preparation." *Cold Spring Harbor Protocols* 2011 (8): 940–49.

Zhou, Xiaobei, Helen Lindsay, and Mark D Robinson. 2014. "Robustly Detecting Differential Expression in RNA Sequencing Data Using Observation Weights." *Nucleic Acids Research*, April, 1–10. doi:10.1093/nar/gku310.

Zhou, Zheng Alan, and Huan Liu. 2011. "Data of High Dimensionality and Challenges." In *Spectral Feature Selection for Data Mining*, 1–19. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. Chapman and Hall/CRC. doi:doi:10.1201/b11426-2.

ABREVIACIONES

ACP:	Análisis de Componentes Principales
NGS:	Secuenciación de nueva generación. (del inglés, Next Generation Sequencing)
RNA-seq:	secuenciación de ARN. (del inglés, RNA sequencing)
HTS:	Tecnología de secuenciación de alto rendimiento (del inglés, High Throughput Sequencing)
cDNA:	ADN complementario (del inglés complementary DNA)
PCR:	Polimerase Chain Reaction
mRNA:	ARN mensajero (del inglés Messenger RNA)
AM:	Análisis multivariado
AC:	Análisis de conglomerados o análisis de cluster
AFM:	Análisis factorial múltiple (en inglés puede verse citado como Multiple Factor Análisis o Múltiple Factorial Analysis)
EM:	Escalamiento multidimensional (en inglés MDS MultiDimensional Scaling)
AFMD	Análisis factorial múltiple dual (en inglés DMFA Dual Multi Factor Analysis)

INDICE DE TEMAS

APENDICE 1: Códigos R

Generación de plasmodios para el experimento de Bottomly

```
###---------------------------------
### Function: partitionBlk2
### Description: genera la partición de un conjunto de datos real en un conjunto de 3
### bloques y s repeticiones en cada bloque
### Input:
###     n1= numero de librerías en el bloque 1 de los datos originales
###     n2= numero de librerías en el bloque 2 de los datos originales
###     n3= numero de librerías en el bloque 2 de los datos originales
###     s=número de elementos a muestrear
###     Condition s  => n
###---------------------------------


partitionBlk2<-function(n1,n2,n3,s1,s2,s3) {
b1<-sample(n1,size=s1,replace=F)
b2<-sample(n2,size=s2,replace=F)
b3<-sample(n3,size=s3,replace=F)
part<-cbind(b1,b2,b3)
return(list(part=part))
}


###Generación de particiones

##50 particiones con  3 bloques y 2 repeticiones por bloque a partir de 10 muestras de
Bottomly (Solamente strain B6)

##    phenodata del archive original
               #sample.id num.tech.reps strain experiment.number lane.number
               #SRX033480 1 C57BL/6J 6 1
               #SRX033488 1 C57BL/6J 7 1
               #SRX033481 1 C57BL/6J 6 2
               #SRX033489 1 C57BL/6J 7 2
               #SRX033482 1 C57BL/6J 6 3
               #SRX033490 1 C57BL/6J 7 3
               #SRX033483 1 C57BL/6J 6 5
               #RX033476 1 C57BL/6J 4 6
               #SRX033478 1 C57BL/6J 4 7
               #SRX033479 1 C57BL/6J 4 8

lib6<-c("SRX033480","SRX033481","SRX033482","SRX033483")  # Librerías del experimento 6
lib7<-c("SRX033488","SRX033489","SRX033490")              # Librerías del experimento 7
lib4<-c("SRX033476","SRX033478","SRX033479")              # Librerías del experimento 4

p<-replicate(100,partitionBlk2(lib6,lib7,lib4,4,3,3))
save(p, file="50part_B6.RData")



### Creación de un plasmodio
       ## Muestreo de efectos del archivo original Bottomly
       ## Creación de un conjunto nulo a partir de una partición
       ## Adición de efectos al conjunto nulo dataset=plasmode
       ## Generación de archivos con conteos y efectos


cts<-read.table("path/bottomly.txt",row.names="gene")
cts<-cts[,-1]  #elimina la primera columna con numero de gen
cnt<-cts[rowSums(cts)>0,] #elimina genes con conteos = 0 en todas las librerías

library(edgeR)
library(qvalue)

load("/50part_B6.RData")    #conjunto de particiones
id<-colnames(cnt)
```

```
idB6<-id[1:10]


partnum <- as.numeric(Sys.getenv("PARTNUM"))  #lee variable ambiental con numero de
particion


### Efectos

strain<-factor(c(rep("B6",10),rep("D2",11)))
exprmt<-factor(c(6,7,6,7,6,7,6,4,4,4,4,4,4,7,6,7,6,7,6,7))
design<-model.matrix(~exprmt+strain)

dge<-DGEList(counts=cnt,group=strain,remove.zeros=T)
cpm.dge<-cpm(dge)
dge<-dge[rowSums(cpm.dge>1)>=9,]   #Filtra transcripts con menos de 1 cpm en 10 o mas
muestras
dim(dge)
dge<-calcNormFactors(dge)
dge<-estimateGLMCommonDisp(dge,design)
dge$common.dispersion
dge<-estimateGLMTagwiseDisp(dge,design)
fit<-glmFit(dge,design)
de.tagwiseGLM<-glmLRT(dge,fit,coef=4)       #test D2 vs B6
tags<-rownames(de.tagwiseGLM$table)         #extrae nombres
q<-qvalue(de.tagwiseGLM$table$PValue)       #calcula qvalues
r<-cbind(de.tagwiseGLM$table$logFC,q$qvalues)
rownames(r)<-tags
colnames(r)<-c("logFC","qvalues")

## Muestrea una proporción de transcripts a partir de todos los que resultaron
diferencialmente expresados con qvalue <0.05

prop<-0.1  #proporción de transcripts diferenciados a muestrear y adicionar al conjunto
nulo
rde<-r[q$qvalues<0.05,]          #selecciona transcripts con qvalues <0.05
s<-sample(rownames(rde),round(prop*length(tags)))      #muestreo sobre los diferenciados
i<-tags%in%s #index con transcript seleccionados

#logFC<-de.tagwiseGLM$table$logFC[i]       #logFc de los transcripts seleccionados

##Filtrado de logFC de los transcripts seleccionados
f<-cbind(r,i)
i_logFC<-r[,"logFC"]*i   #logFC de los transc seleccionados
f<-cbind(r,i,i_logFC)
ineglogFC<-f[,"i_logFC"]<0      #index de los logFC con signo negativosof
iposlogFC<-f[,"i_logFC"]>0      #index de los logFC con signo positivo
ineg_logFC<-f[,"logFC"]*ineglogFC      #logFC  negativos
ipos_logFC<-f[,"logFC"]*iposlogFC      #logFC positives
logFC_A<-ineg_logFC*-1     #logFC a adicinar al trt A
logFC_B<-ipos_logFC        #logFC a adicionar al trt B
eff<-cbind(r,i,logFC_A,logFC_B)



## Conjunto nulo

lib<-
c(p[partnum]$part[1,1],p[partnum]$part[2,1],p[partnum]$part[3,1],p[partnum]$part[4,1],p[par
tnum]$part[1,2],p[partnum]$part[2,2],p[partnum]$part[3,2],p[partnum]$part[1,3],p[partnum]$p
art[2,3],p[partnum]$part[3,3]) # samples selected at random
tags_cnt<-rownames(cnt)
filter<-tags_cnt%in%tags


p1<-cnt[filter,lib]    #particion con las librerías seleccionadas y solo los transcripts
communes a las 21 librerías

logC<-log(p1+1,2) # transformacion a logFC


## Adición de efectos al conjunto nulo

     #adición de  log2 counts + logFC a cada librería

r6A1<-logC[,1]+eff[,"logFC_A"]
```

```
r6B1<-logC[,2]+eff[,"logFC_B"]
r6A2<-logC[,3]+eff[,"logFC_A"]
r6B2<-logC[,4]+eff[,"logFC_B"]

r7A1<-logC[,5]+eff[,"logFC_A"]
r7B1<-logC[,6]+eff[,"logFC_B"]
r7B2<-logC[,7]+eff[,"logFC_B"]

r4A1<-logC[,8]+eff[,"logFC_A"]
r4B1<-logC[,9]+eff[,"logFC_B"]
r4B2<-logC[,10]+eff[,"logFC_B"]


pl<-cbind(r6A1,r6B1,r6A2,r6B2,r7A1,r7B1,r7B2,r4A1,r4B1,r4B2)
plasmode<-round(2^pl-1)           # Trasformación exponencial  y sustracción de 1 para
regenerar los conteos

## Guarda archivos con efectos y conteos ambos dge and de.tagwiseGLM
save(eff, file=paste("path/eff_plasmode_",partnum,".RData",sep=""))      # para recuperar
i= indice de transcript seleccionado y efectos
save(plasmode, file=paste("path/plasmode_",partnum,".RData",sep=""))     #plasmodio con
conteos
```

Calculo de disimilaridades

```
### Función para calcular disimilaridades  basadas en correlaciones pea, plg, spe

diss_corr<-function(cnt)
{
  ## Pearson Correlation
  corr.p<-cor(cnt,method="pearson")
  dist.p <- as.dist(1-corr.p,diag=TRUE)

  ## Pearson Correlation on log2(cnt+1)
  log2cnt1<-log2(cnt+1)
  corr.plog<-cor(log2cnt1,method="pearson")
  dist.plog <- as.dist(1-corr.plog,diag=TRUE)

  ## Spearman Correlation
  corr.sp<-cor(cnt,method="spearman")
  dist.sp <- as.dist(1-corr.sp,diag=TRUE)

  distances<-list(dist.p,dist.plog,dist.sp)
  distances_names<-(c("pea","plg","spe"))
  names(distances)<-distances_names
  return(distances)

}


### Función para calcular las disimilaridades raw, rnr, rld, pln

library(DESeq2)
diss_deseq2<-function(cnt)
{
  ##Data structure
  colData<-
data.frame(cbind(colnames(cnt),c("A","B","A","B","A","B","B","A","B","B"),c(6,6,6,6,7,7,7,4
,4,4)))
  strainXexperiment<-paste(colData[,2],colData[,3],sep=":")
  colData<-cbind(colData,strainXexperiment)
  colnames(colData)<-c("sample","strain","experiment","strainXexperiment")

  ## Create a DESeqDataSet object, from count matrix input
  dds <- DESeqDataSetFromMatrix(countData = cnt,
                                colData = colData,
                                design = ~ experiment+strain)

  ## Perform analysis with DESeq2
  dds<-DESeq(dds)
```

```
  ## compute distance matrices
  ddsBlind <- dds
  design(ddsBlind) <- formula(~ 1)
  ddsBlind <- estimateDispersions(ddsBlind)
  rld <- rlogTransformation(ddsBlind)
  vsd <- varianceStabilizingTransformation(ddsBlind)
  distsRL <- dist(t(assay(rld)))
  distsVSD <- dist(t(assay(vsd)))
  ##raw data
  distsRAW <- dist(t(assay(dds)))
  ## raw data normalized
  a<-counts(dds,normalized=TRUE)
  distsRAWNOR <- dist(t(a))
  ## raw data normalized
  loga<-log2(a+1)
  distsPEALOGNOR <- dist(t(loga))


  distances<-list(distsRAW,distsRAWNOR,distsRLD,distsVSD,distsPEALOGNOR)
  distances_names<-(c("raw","rnr","rld","vsd","pln"))
  names(distances)<-distances_names
  return(distances)

}



### Función para calcular disimilaridad voo

library(edgeR)

diss_voom<-function(cnt)
{

y <- DGEList(counts=cnt)
y <- calcNormFactors(y) # apply TMM normalization as proposed by edgeR
v <- voom(y) #not desing because I want to explore a priori

distsVOO <- dist(t(v$E)) #distances of log-cpm normalized data
pdf()
mdsres<-plotMDS(v,top=500,gene.selection="common") # distances using plotMDS as proposed by
edgeR, ## limma, it takes by default 500 top genes with lasrgest standar deviations.
option=commom takes the top ## genes between samples. pairwise consider pairs of samples (I
think is like comparing using different ## variables
dev.off()

distsedgeR<-as.dist(mdsres$distance.matrix)

  distances<-list(distsVOO,distsedgeR,mdsres)
  distances_names<-(c("voo","edg","mdsres"))
  names(distances)<-distances_names
  return(distances)

}



### Función para calcular la disimilaridad poi

library(PoiClaClu)
diss_poi<-function(cnt)
{
  dd <- PoissonDistance(t(cnt),type="deseq")    #type= normalization method, deseq as
implemented in such package
  dist.poi<-dd$dd

  distances<-list(dist.poi)
  distances_names<-(c("poi"))
  names(distances)<-distances_names
  return(distances)

}


### Function para calcular la disimilaridad gcc
```

```
library(rsgcc)
diss_gcc<-function(cnt)
{
d<-gcc.dist(t(cnt), method = "GCC", distancemethod = "Raw")
distsGCC <- d$dist

  distances<-list(distsGCC)
  distances_names<-(c("gcc"))
  names(distances)<-distances_names
  return(distances)

}


### Función para calcular la disimilaridad chi

###
### FUNCTION: ChisqDist
###           Computes Chi squares distances between columns of a matrix
###
###           Input: a matrix
###           Output: a distance matrix with Chi squares distances between columns


ChisqDist<-function(m){
  n<-ncol(m)
  D<-matrix(NA,rep(n*n))
  dim(D)<-c(n,n)
  p<-prop.table(m,2)

  for(i in 2:n){
    for (j in 2:i-1){
      #print(paste(i,j,sep="_")) #control
      sqdiff<-function(e)
      {
        diff<-(e[i]-e[j])**2
      }
      sqd<-apply(p,1,FUN=sqdiff)
      idx<-rowSums(m)!=0 # filter row with all zeros
      w<-1/rowSums(m[idx,])
      D[i,j]<-sum(w*sqd[idx])

    }
  }

  chisqD<-as.dist(D)
  attr(chisqD,"Labels")<-colnames(m)
 return(chisqD)
}



### Función para calcular la disimilaridad psv

## PSVD algorithm
source(paste(path,"/Scripts/PSVD_offsetv4.R",sep=""))

PSVDOS<-function(cnt)
{
  y<-cnt
  psvdos = pSVD.offsetv4(y,K=2,verbose=1,err = 0.0001,niter = 300)       #psvdos algorithm
K=2 dimensions
  ynew<-psvdos$mu + psvdos$rowoffset + psvdos$B%*%t(psvdos$F)  #observations according to 2
factors extracted by psvdos algorithm
  psvddist<-dist(t(ynew))
  attr(psvddist,"Labels")<-colnames(y)
  res<-list(psvddist,ynew,psvdos)
  res_names<-(c("psv","psvcounts","psvres"))
  names(res)<-res_names
  return(res)
}
```

Análisis de cluster

Comparacion entre distancias (concordancia)

```
## Compara las jerarquías obtenidas con las distancias para cada uno de los 50 plasmodios

library(clue)

ds="plde" #archivo con resultados para el escenario 1, solo 10% de los transcripts
diferenciados

## FUNCTION: extractdist
## Input: partnum : partion number / dataset: plde, plasmode / dist: distance matrix: raw,
rnr, rld, vsd, pln, pea, plg, spe, poi, gcc, voo ,chi, psv
##


extractdist<-function(partnum,dataset="plde")
{
  load(paste("path/dist_",partnum,".RData",sep=""))
  d<-alldist[[dataset]]

}

dd<-lapply(1:50,extractdist,dataset="plde")


exphier<-function(partnum)
{
  hclust_results<-lapply(dd[[partnum]],function(d) hclust(d))
  ens<-cl_ensemble(list=hclust_results)
  pdf(file=paste("path/plasmode_",partnum,".pdf",sep=""))
  plot(hclust(cl_dissimilarity(ens,method="cophenetic")))
  dev.off()
  hc<-hclust(cl_dissimilarity(ens,method="cophenetic"))
  coph<-cl_agreement(ens,method="cophenetic")
  s<-summary(c(coph))
  m<-mean(c(coph))
  sd<-sd(c(coph))
  return(list(coph=coph,summary=s,mean=m,sd=sd,ens=ens,hc=hc))

}
```

Comparacion entre distancias (consistencia)

```
## Compara las jerarquias obtenidas por los 50 plasmodios para cada distancias

library(clue)

## FUNCTION: extractdist
## Input: partnum : partion number / dataset: plde, plde50add, plasmode / dist: distance
matrix raw, rnr, rld, vsd, pln, pea, plg, spe, poi, gcc, voo, edg

extractdist<-function(partnum,dataset="plde",dist="raw")
{
  load(paste(path,"/Outputs/Distances/dist_",partnum,".RData",sep=""))
  d<-alldist[[dataset]][[dist]]
}

ds="plde"

dd1<-lapply(1:50,extractdist,dataset="plde",dist="raw")
dd2<-lapply(1:50,extractdist,dataset="plde",dist="rnr")
dd3<-lapply(1:50,extractdist,dataset="plde",dist="rld")
dd4<-lapply(1:50,extractdist,dataset="plde",dist="vsd")
dd5<-lapply(1:50,extractdist,dataset="plde",dist="pln")
dd6<-lapply(1:50,extractdist,dataset="plde",dist="pea")
dd7<-lapply(1:50,extractdist,dataset="plde",dist="plg")
dd8<-lapply(1:50,extractdist,dataset="plde",dist="spe")
dd9<-lapply(1:50,extractdist,dataset="plde",dist="poi")
dd10<-lapply(1:50,extractdist,dataset="plde",dist="gcc")
```

```
dd11<-lapply(1:50,extractdist,dataset="plde",dist="voo")
dd12<-lapply(1:50,extractdist,dataset="plde",dist="edg")
dd13<-lapply(1:50,extractdist,dataset="plde",dist="chi")
dd14<-lapply(1:50,extractdist,dataset="plde",dist="psv")


dd<-list(dd1,dd2,dd3,dd4,dd5,dd6,dd7,dd8,dd9,dd10,dd11,dd12,dd13,dd14)
distances<-c("raw", "rnr", "rld", "vsd","pln","pea", "plg", "spe", "poi","gcc", "voo",
"edg","chi","psv")
names(dd)<-distances

exphier<-function(dist)
{
  hclust_results<-lapply(dd[[dist]],function(d) hclust(d))
  ens<-cl_ensemble(list=hclust_results)
  pdf(file=paste(path,"/Outputs/Results/plde/dist_",dist,".pdf",sep=""))
  plot(hclust(cl_dissimilarity(ens,method="cophenetic")))
  dev.off()
  coph<-cl_agreement(ens,method="cophenetic")
  s<-summary(c(coph))
  m<-mean(c(coph))
  sd<-sd(c(coph))
  return(list(summary=s,mean=m,sd=sd))
}

res<-lapply(distances, exphier)
names(res)<-distances

res2<-lapply(distances, exphier2<-function(d) res[[d]][["mean"]])
m<-unlist(res2)

res3<-lapply(distances, exphier2<-function(d) res[[d]][["sd"]])
sd<-unlist(res3)

res<-cbind(m,sd)
rownames(res)<-distances
```

Análisis de escalamiento multidimensional

```
## distancias calculadas en la configuracion de 2 dimensiones, escalamiento
multidimensional métrico

mds_dist<-function(d){
  loc <- cmdscale(d,k=2,eig=TRUE,add=TRUE)
  X_eigen <- loc$eig
  dd<-dist(loc$point)
  res<-list(dist=dd)
  return(res)
  }

## resultados de cmdscale y criterios de  Mardia para % explicación, escalamiento métrico

mds_res<-function(d){
  loc <- cmdscale(d,k=2,eig=TRUE,add=TRUE)
  X_eigen <- loc$eig
  mardia1<-cumsum(abs(X_eigen)) / sum(abs(X_eigen))
  mardia2<-cumsum(X_eigen^2) / sum(X_eigen^2)
  res<-list(mdsres=list(cmdscale=loc,mardia1=mardia1,mardia2=mardia2))
  return(res)
  }


## distancias calculadas en la configuracion de 2 dimensiones, escalamiento
multidimensional métrico
library(MASS)

nonmetric_mds_dist<-function(d){
  loc <- isoMDS(d,k=2,maxit=200)
  dd<-dist(loc$points)
  res<-list(dist=dd)
```

```
  return(res)
  }


## resultados del escalamiento multidimensional no métrico, isoMDS

nonmetric_mds_res<-function(d){
  loc <- isoMDS(d,k=2,maxit=200)
  res<-list(nonmetricmdsres=list(isoMDS=loc))
  return(res)
  }



### Comparación de técnicas de ordenación, resultados obtenidos entre distancias
(concordancia)


## MDS
extractcoord_MDS<-function(diss,partnum,dataset){
  load(paste(path,"/Outputs/MDS/MDSres.RData",sep=""))
  d<-MDSres[[dataset]][[partnum]][[diss]]$mdsres$cmdscale$points
  }

## nonmetric MDS
extractcoord_nmMDS<-function(diss,partnum,dataset){
  load(paste(path,"/Outputs/nonmetric_MDS/nonmetric_MDSres.RData",sep=""))
  d<-nonmetric_MDSres[[dataset]][[partnum]][[diss]]$nonmetricmdsres$isoMDS$points
  }


diss_names<-
c("raw","rnr","rld","vsd","pln","pea","plg","spe","poi","gcc","voo","edg","chi")
proc<-c("raw_m", "rnr_m", "rld_m", "vsd_m","pln_m","pea_m", "plg_m", "spe_m",
"poi_m","gcc_m", "voo_m", "edg_m","chi_m", "raw_N", "rnr_N", "rld_N",
"vsd_N","pln_N","pea_N", "plg_N", "spe_N", "poi_N","gcc_N", "voo_N", "edg_N","chi_N")


extract_all<-function(p,ds){
  c1<-lapply(diss_names,extractcoord_MDS,partnum=p,dataset=ds)
  c2<-lapply(diss_names,extractcoord_nmMDS,partnum=p,dataset=ds)
  c<-c(c1,c2)
  proc<-c("raw_m", "rnr_m", "rld_m", "vsd_m","pln_m","pea_m", "plg_m", "spe_m",
"poi_m","gcc_m", "voo_m", "edg_m","chi_m", "raw_N", "rnr_N", "rld_N",
"vsd_N","pln_N","pea_N", "plg_N", "spe_N", "poi_N","gcc_N", "voo_N", "edg_N","chi_N")
  names(c)<-proc
  return(c)
}


## cálculo del coeficiente  RV coefficien para todas las matrices
library(FactoMineR)

calculate_RV<-function(configs){
  n<-length(configs)
  size<-n*n
  rvs<-matrix(data=NA,n,n)
  for (i in 1:length(configs)){
    for (j in 1:length(configs)){
      res<-coeffRV(configs[[i]],configs[[j]])
      rv<-res$rv
      #print(i)
      #print(j)
      #print(rv)
      rvs[i,j]<-rv
    }
  }
  return(rvs)
}

rv_xpartnum<-function(p){
  calculate_RV(coord_all[[p]])
}


ds<-"plde"
```

```
coord_all<-lapply(1:50,extract_all,ds=ds)
rv_all<-lapply(1:50,rv_xpartnum)
ordmean<-round(apply(simplify2array(rv_all), 1:2, mean),3)
ordsd<-round(apply(simplify2array(rv_all), 1:2, sd),3)

write.csv(ordmean,file=paste(path,"/AgreementRV_Ord_",ds,"_mean.csv",sep=""))
write.csv(ordsd,file=paste(path,"/AgreementRV_Ord_",ds,"_sd.csv",sep=""))



### Comparación de técnicas de ordenación, resultados obtenidos para los 50 plasmodios
obtenidos para cada distancia (consistencia)

## MDS
extractcoord_MDS<-function(partnum,diss,dataset)
{
  load(paste(path,"path/MDSres.RData",sep=""))
  d<-MDSres[[dataset]][[partnum]][[diss]]$mdsres$cmdscale$points
}

## nonmetric MDS
extractcoord_nmMDS<-function(partnum,diss,dataset)
{
  load(paste(path,"/nonmetric_MDSres.RData",sep=""))
  d<-nonmetric_MDSres[[dataset]][[partnum]][[diss]]$nonmetricmdsres$isoMDS$points
}

diss_names<-
c("raw","rnr","rld","vsd","pln","pea","plg","spe","poi","gcc","voo","edg","chi")
proc<-c("raw_m", "rnr_m", "rld_m", "vsd_m","pln_m","pea_m", "plg_m", "spe_m",
"poi_m","gcc_m", "voo_m", "edg_m","chi_m", "raw_N", "rnr_N", "rld_N",
"vsd_N","pln_N","pea_N", "plg_N", "spe_N", "poi_N","gcc_N", "voo_N", "edg_N","chi_N")


extract_all_MDS<-function(diss,ds){
  c<-lapply(1:50,extractcoord_MDS,diss=diss,dataset=ds)
  return(c)
  }

extract_all_nmMDS<-function(diss,ds){
  c<-lapply(1:50,extractcoord_nmMDS,diss=diss,dataset=ds)
  return(c)
  }


## calculo del coeficiente RV

library(FactoMineR)

calculate_RV<-function(configs){
  n<-length(configs)
  size<-n*n
  rvs<-matrix(data=NA,n,n)
  for (i in 1:length(configs)){
    for (j in 1:length(configs)){
      res<-coeffRV(configs[[i]],configs[[j]])
      rv<-res$rv
      #print(i)
      #print(j)
      #print(rv)
      rvs[i,j]<-rv
    }
  }
  return(rvs)
}

rv_xdiss<-function(p){
  calculate_RV(coord_all[[p]])
}


ds<-"plde"


all_MDS<-lapply(diss_names,extract_all_MDS,ds=ds)
all_nmMDS<-lapply(diss_names,extract_all_nmMDS,ds=ds)
coord_all<-c(all_MDS, all_nmMDS)
```

```
rv_all<-lapply(1:26,rv_xdiss)

stats_rv<-function(m,results) {
  m.dist<-as.dist(results[[m]])
  m<-mean(m.dist)
  sd<-sd(m.dist)
return(list(mean=m,sd=sd))
}

res_rv<-sapply(1:26,stats_rv,results=rv_all)
colnames(res_rv)<-proc


write.csv(t(res_rv),file=paste(path," /ConsistencyRV_Ord_",ds,"_.csv",sep=""))



## calculo de porcentaje de explicación y stress
ext_res<-function(partnum,method,dataset,dist) {
  if(method=="MDS"){
    load(paste(path,"/Outputs/MDS/MDSres.RData",sep=""))
    res<-MDSres[[dataset]][[partnum]][[dist]]$mdsres$mardia1[[2]]
    }
  if(method=="nMDS"){
    load(paste(path,"/Outputs/nonmetric_MDS/nonmetric_MDSres.RData",sep=""))
    res<-nonmetric_MDSres[[dataset]][[partnum]][[dist]]$nonmetricmdsres$isoMDS$stress
    }
return(res)
}


explanation<-function(d,dataset){
  r<-unlist(lapply(1:50,ext_res,method="MDS",dataset,dist=d))
  return(c(mean(r), sd(r)))
}

stress<-function(d,dataset){
  r<-unlist(lapply(1:50,ext_res,method="nMDS",dataset,dist=d))
  return(c(mean(r), sd(r)))
}


distances<-c("raw", "rnr", "rld", "vsd","pln","pea", "plg", "spe", "poi","gcc", "voo",
"edg","chi")

exp_plde<-sapply(distances,explanation,dataset="plde")
```

Análisis Factorial Múltiple

```
### Obtención de datos originales transformados

### Función para calcular transformaciones rld y log2+1 a los conteos originales

library(DESeq2)
library(edgeR)

transf_counts<-function(cnt)
{
  ##Data structure
  colData<-
data.frame(cbind(colnames(cnt),c("A","B","A","B","A","B","B","A","B","B"),c(6,6,6,6,7,7,7,4
,4,4)))
  strainXexperiment<-paste(colData[,2],colData[,3],sep=":")
  colData<-cbind(colData,strainXexperiment)
  colnames(colData)<-c("sample","strain","experiment","strainXexperiment")

  ## Create a DESeqDataSet object, from count matrix input
  dds <- DESeqDataSetFromMatrix(countData = cnt,
                                colData = colData,
```

```
                                design = ~ experiment+strain)

  ## Análisis con DESeq2
  dds<-DESeq(dds)

  ## rld
  ddsBlind <- dds
  design(ddsBlind) <- formula(~ 1)
  ddsBlind <- estimateDispersions(ddsBlind)
  rld <- rlogTransformation(ddsBlind)
  rld_data<-assay(rld)
  colnames(rld_data)<-colnames(cnt)

  ## log2 + 1 de los conteos originales
  lg2_data<-log2(cnt+1)

  ## limma voom transformation = log cpm normalized data by edgeR
  y <- DGEList(counts=cnt)
  y <- calcNormFactors(y) # aplica la normalizacion  TMM
  v <- voom(y) # sin diseno
  voo_data<-v$E

  ## output list con datos transformados
  t_counts<-list(lg2=lg2_data,rld=rld_data,voo=voo_data)
  return(t_counts)

}



### Análisis Factorial Multiple Dual

library(FactoMineR)

calculate_DMFA<-function(partnum,dataset,transf){
        load(paste(path,"/transf_counts_",partnum,".RData",sep="")) # file with transformed
        counts

        cnt<-alltransf[[dataset]][[transf]]

        dt<-t(cnt)
        rownames(dt)
        trat<-c("A","B","A","B","A","B","B","A","B","B")
        flowcell<-as.factor(c(6,6,6,6,7,7,7,4,4,4))

        ##El DMFA por defecto centra los datos de cada grupo por su respectiva media y
        desvio y luego hace un ACP para cada grupo.

        ## Análisis agrupando por tratamiento
        trat.DMFA<-data.frame(trat,dt)
        res<-DMFA(trat.DMFA, num.fact=1, ncp=5, scale.unit = TRUE, graph=FALSE)
        dd<-dist(res$ind$coord[,1:2])
        res_trat<-list(eig=res$eig,coord=res$ind$coord,dist=dd)


        ## Análisis agrupando por flowcell
        fc.DMFA<-data.frame(flowcell,dt)
        res<-DMFA(fc.DMFA, num.fact=1, ncp=5, scale.unit = TRUE, graph=FALSE)
        dd<-dist(res$ind$coord[,1:2])
        res_fc<-list(eig=res$eig,coord=res$ind$coord,dist=dd)

        DMFA_res<-list(xtrat=res_trat,xflowcell=res_fc)
        #return(DMFA_res)

        save(DMFA_res,file=paste(path,"/DMFA_res_",dataset,"_",transf,"_",partnum,".RData",
        sep=""))
        }
```