UNIVERSITAT POLITÈCNICA DE VALÈNCIA

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA AGRONÓMICA Y DEL MEDIO NATURAL

Análisis mediante RNAseq de la respuesta transcripcional de *Saccharomyces cerevisiae* en presencia de ácido acético en una cepa silvestre y en otra con pérdida de función del gen *SSD1*

TRABAJO FINAL DE GRADO EN BIOTECNOLOGÍA

AUTOR: Enrique Blanco Carmona
TUTOR: Prof. Dr. D. Ramón Serrano Salom
TUTOR EXPERIMENTAL: Dr. D. Alessandro Rienzo

Curso 2017-2018
Valencia, julio 2018

Datos personales

Nombre y apellidos: Enrique Blanco Carmona

Datos del trabajo de fin de grado

Título: Análisis mediante RNAseq de la respuesta transcripcional de *Saccharomyces cerevisiae* en presencia de ácido acético en una cepa silvestre y en otra con pérdida de función del gen *SSD1*.
Titulación: Grado en Biotecnología
Tutor: Prof. Dr. D. Ramón Serrano Salom
Tutor experimental: Dr. D. Alesandro Rienzo
Lugar y fecha de lectura: Valencia, julio 2018

Resumen

Una de las características distintivas (*hallmark*) de la mayoría de células cancerígenas es la relación pH_i/pH_o inversa respecto al resto de células somáticas, siendo más alcalino el medio intracelular que el extracelular. Este hecho es relevante debido a su aplicabilidad biomédica, puesto que esto deriva en una diferente regulación de eventos celulares como, por ejemplo, el ciclo celular y la tasa de crecimiento, entre otros. De igual manera, se ha demostrado que el suministro de ácidos débiles puede influir de forma negativa en dichos procesos.

En este marco, *SSD1* es un gen implicado en el progreso del ciclo celular, longevidad, patogenicidad, tolerancia a temperatura y morfogénesis celular en *Saccharomyces cerevisiae*. Muchas de sus funciones son aún desconocidas, entre ellas su implicación en la homeostasis del pH. Ssd1 es una proteína de unión al RNA y se cree responsable de la regulación transcripcional y post-transcripcional de genes específicos, entre ellos ciclinas del ciclo celular como *CLN2*. Esta interacción estabiliza mRNAs, aumentando así su vida media en la célula. *SSD1* presenta dos alelos: *SSD1-V* y *ssd1-d*. El primero es silvestre y el otro presenta una mutación con pérdida de función. El alelo *SSD1-V* confiere resistencia a ácido acético en medio mínimo en contraposición al alelo *ssd1-d*.

En el presente TFG se quiere ver la expresión génica diferencial de estos dos alelos en presencia y ausencia de ácido acético (45 mM). Esto se analiza mediante RNAseq, con el fin de individualizar genes responsables del crecimiento diferencial entre ambos alelos. Además, se caracteriza el perfil de crecimiento de ambos alelos mediante medición continua de absorbancia (*Bioscreen*).

Los resultados del *Bioscreen* muestran lo evidenciado en experimentos anteriores, un perfil de crecimiento similar entre ambos alelos, tanto a pH 6 como a pH 4, pero un menor crecimiento de la cepa *ssd1-d* frente a la cepa *SSD1-V* en estrés por ácido acético. Asimismo, a la luz de los resultados preliminares del análisis de expresión diferencial, destaca el gen *PCL2*, que codifica para una ciclina del ciclo celular con función análoga a la ciclina *CLN2*. Estando ambas ciclinas presentes en la fase G_1 y siendo responsables del progreso hacia el ciclo mitótico, *PCL2* es el gen candidato para futuros experimentos de interacción entre *SSD1-CLN2*.

Palabras clave

Saccharomyces cerevisiae, crecimiento, regulación, *SSD1*, RNAseq, pH intracelular.

Abstract

One of the hallmarks of most cancer cells is the inverse pH_i/pH_o relation to the rest of somatic cells, the intracellular medium being more alkaline than the extracellular one. This fact is relevant due to its biomedical applicability, since this leads to a different regulation of cellular events such as the cell cycle and the growth rate, among others. Likewise, it has been demonstrated that the supply of weak acids can negatively influence said processes.

In this framework, SSD1 is a gene involved in cell cycle progress, longevity, pathogenicity, temperature tolerance and cellular morphogenesis in *Saccharomyces cerevisiae*. Many of its functions are still unknown, including its relation in pH homeostasis. Ssd1 is an RNA binding protein. It is believed for Ssd1 to be responsible of the transcriptional and post-transcriptional regulation of specific genes, including cyclins such as *CLN2*. This interaction stabilizes mRNAs, thus increasing their half-life in the cell. *SSD1* has two alleles: *SSD1-V* and *ssd1-d*. The first one is the wild type and the other one has a mutation with loss of function. The *SSD1-V* allele confers resistance to acetic acid in minimal medium.

In the present dissertation we want to see the differential gene expression of these two alleles in the presence and absence of acetic acid (45 mM). This is analyzed by RNAseq, in order to identify genes responsible for differential growth between both alleles. In addition, the growth profile of both alleles is characterized by continuous absorbance measurement (Bioscreen).

The results of the Bioscreen show what was evidenced in previous experiments, a similar growth profile between both alleles, both at pH 6 and at pH 4, but a lower growth of the strain *ssd1-d* compared to the strain *SSD1-V* in stress by acetic acid. Likewise, in light of the preliminary results of the differential expression analysis, the *PCL2* gene, which codes for a cyclin of the cell cycle with analogous function to the cyclin *CLN2*, stands out. With both cyclines present in the G_1 phase and being responsible for the progress towards the mitotic cycle, *PCL2* is the candidate gene for future interaction experiments between *SSD1-CLN2*.

Key words

Saccharomyces cerevisiae, growth, regulation, *SSD1*, RNAseq, intracellular pH.

AGRADECIMIENTOS

Me gustaría realizar una especial mención:

- Al profesor Ramón Serrano, por abrirme las puertas de su laboratorio y ser fuente de consejo en las decisiones que han marcado mi futura trayectoria profesional.

- Al doctor Alessandro Rienzo, por tutorizar y supervisar los experimentos realizados tanto en mis prácticas curriculares como en el presente TFG.

- Al profesor Javier Forment, por su disposición a resolver cualquier duda bioinformática y académica surgida.

- A todo el equipo del laboratorio, por su ayuda desinteresada.

- A Emilio Francés y Marc Muniesa, por recordarme que no estaba solo en este barco.

- A mi familia y amigos, por ayudar a canalizar los picos de estrés derivados de la realización del presente TFG.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ÍNDICE DE ABREVIATURAS

AcH	Ácido acético
ADN/DNA	Ácido desoxirribonucleico
ADNc/cDNA	ADN circular
Agua DEPC	Agua milliQ libre de RNAsas
AMPc	Adenosín monofosfato cíclico
ARN/RNA	Ácido ribonucleico
ARNm/mRNA	Ácido ribonucleico mensajero
ARNnc/ncRNA	ARN no codificante
ARNr	ARN ribosómico
atm	Atmósfera
DEG	Genes diferencialmente expresados
EDTA	Ácido etilendiaminotetraacético
FC	*Fold Change*
FDR	Frecuencia de falsos positivos
FPKM	*Fragments Per Kilobase of transcript per Million fragments sequenced*
FWER	*Family Wise Error Rate*
GFP	Proteína fluorescente verde
HCl	Ácido clorhídrico
his	Histidina
INDEL	Inserción-Deleción
leu	Leucina
MAPK	Proteín quinasa activada po mitógenos
MATα	Tipo de levadura haploide que produce junto con otra levadura tipo MATa una célula diploide
MCS	Sitio de clonación múltiple
met	Metionina
miARN/miRNA	Micro ácido ribonucleico
miliQ	Ultrapuro
MPSS	*Massive Parallel Signature Sequencing*
NaAc	Acetato sódico
NGS	Tecnologías de secuenciación masiva
NLS	Dominio de localización nuclear
OD	Absorbancia
OligoT	Oligonucleótidos de timina
PB	Cuerpos P
pb/bp	Pares de bases
PCA	Análisis de componentes principales
PCI	Fenol Cloroformo Isoamilalcohol
PCR	Reacción en cadena de la polimerasa
pKa	Constante de disociación ácida
qPCR	PCR cuantitativa
RAM	Regulación de Ace2 y morfogénesis polarizada
RBD	Dominio de unión a ARN
RIN	Número de integridad del ARN
RPKM	*Reads Per Kilobase of exon model per Million reads*
rpm	Revoluciones por minuto
SAGE	Análisis seriado de la expresión génica

SD	Medio mínimo
SDS	Dodecilsulfato sódico
SG	Gránulos de estrés
SNP	Polimorfismo de nucleótido único
TOR	Diana de la rapamicina
TPM	Transcritos por millón de lecturas
ura	Uracilo
UTR	Región sin traducción delante de la trama de lectura

1. INTRODUCCIÓN

1.1. PH INTRACELULAR Y SU RELACIÓN CON EL CRECIMIENTO CELULAR

El cáncer es considerado un grupo heterogéneo de enfermedades raras, donde cada una de ellas es capaz de presentar diferentes subtipos dependiendo del desarrollo del mismo en el paciente. La necesidad de encontrar características comunes entre todos los tipos de cáncer fue acrecentándose. De ella surge el término *Hallmark* (Hanahan y Weinberg, 2000), comprendiendo inicialmente 6 características, posteriormente revisadas y ampliadas a 10 (Hanahan y Weinberg, 2011). Dicha clasificación no es fija. Por lo tanto, nuevas características son candidatas a ser un *Hallmark* del cáncer, entre ellas el pH intracelular (Sharma *et al.*, 2015).

Las células humanas normales poseen un pH intracelular comprendido entre 7,1 y 7,2, siendo este más ácido que el medio extracelular (pH 7,4). Esto es debido a la homeostasis del pH, conseguida mediante el uso de bombas de protones, junto con otros tipos de transportadores con actividad regulada por el pH (Damaghi *et al.*, 2013). En el caso del microambiente tumoral, esta relación se invierte como consecuencia de una hipoxia local, una actividad vacuolar afectada, y un aumento del metabolismo fermentativo (Hashim *et al.*, 2011).

La acidificación del medio externo ocurre en las primeras etapas de desarrollo del cáncer (Barathova *et al.*, 2008), donde las células se alejan de los vasos sanguíneos, encontrándose en un ambiente anóxico. Esto genera que, por efecto Pasteur (Barker *et al.*, 1964) su metabolismo derive a uno de tipo fermentativo (Gatenby y Gillies, 2004). Los subproductos de este metabolismo, protones y ácido láctico, afectan a la homeostasis del pH intracelular, pudiendo producir citotoxicidad e inducción de la apoptosis (Gottlieb *et al.*, 1996; Lagadic-Gossmann *et al.*, 2004). En respuesta, se activan transportadores de monocarboxilatos y bombas de sodio (Gillies, 2002; Gallagher *et al.*, 2008), a fin de secretar el lactato y los protones, respectivamente. De esta manera, se genera la inversión en el gradiente de pH mencionado anteriormente. Ello promueve la migración e invasión tumoral (Bradley *et al.*, 2011; Hanahan y Weinberg, 2011), favoreciendo la progresión del tumor.

1.2. LEVADURA COMO SISTEMA MODELO

Saccharomyces cerevisiae es un organismo ampliamente utilizado como sistema modelo, entre otras aplicaciones (Botstein y Fink, 2011). Un gran número de características son las responsables de su popularidad, siendo una de las más remarcables su rápido crecimiento y facilidad de manipulación, tanto en su forma haploide como diploide. En consecuencia, existe un amplio abanico de métodos para su manipulación.

Siendo el primer organismo cuyo genoma fue completamente secuenciado y publicado en la red (Goffeau *et al.*, 1996), su contribución en el desarrollo de las técnicas moleculares fue determinante. En este marco, colecciones de cepas de deleción (*knockouts*) fueron generadas en levaduras (Giaever *et al.*, 2002; Winzeler *et al.*, 1999). Asimismo, librerías para cribados (*screenings*) de sobreexpresión, importantes para el estudio de rutas genéticas, fueron a su vez generadas (Jones *et al.*, 2008). Todo esto, junto con la creación de *S. cerevisiae* recombinante con etiquetas de proteína fluorescente verde (*Green Fluorescent* Protein, GFP) (Ghaemmaghami *et al.*, 2003; Huh *et al.*, 2003), ha supuesto la consolidación de una madurez técnica y experimental en torno a *S. cerevisiae*.

En consecuencia, con el desarrollo de las técnicas ómicas, se opta por elegir la levadura como el organismo donde se efectúan los ensayos. Esto es aplicable a estudios transcriptómicos (Lashkari *et al.*, 1997; Dang *et al.*, 2014), proteómicos (Zhu *et al.*, 2001) y metabolómicos (Villas-Bôas *et al.*, 2005; Jewett *et al.*, 2006). Es común encontrar estudios en levadura focalizados en la determinación de las interacciones celulares, tanto a nivel de proteína-proteína (Ito *et al.*, 2001; Krogan *et al.*, 2006), proteína-ADN (Iyer *et al.*, 2001; Lieb *et al.*, 2001), y genéticas (Tong *et al.*, 2001; Costanzo *et al.*, 2010), conformando el interactoma.

En adición, las levaduras y el ser humano poseen una fracción significativa de rutas metabólicas y funcionales en común. Dichas rutas controlan aspectos cruciales, tales como el ciclo celular (Hoose *et al.*, 2012; McInerny, 2016; Mahmoud *et al.*, 2017), metabolismo (Petranovic *et al.*, 2010), muerte celular programada (Munoz *et al.*, 2012), y rutas de señalización como las de las proteín kinasas activadas por mitógenos (*mitogen-activated protein kinase,* MAPK) (Widmann *et al.*, 1999), junto con la ruta de la diana de la rapamicina (*target of rapamycin*, (TOR)) (De Virgilio y Loewith, 2006), entre otras.

1.3. FUNCIÓN DE *SSD1* EN LA REGULACIÓN DEL CRECIMIENTO

El gen *SSD1* fue descubierto como el supresor de la letalidad causada por la deleción del gen *SIT4* (Sutton *et al.*, 1991). Esta letalidad fue rebatida posteriormente, sustituyéndose por un fenotipo de sensibilidad a temperatura y defectos en el crecimiento. *SSD1* presenta dos alelos: *SSD1-V* (*viable*) y *ssd1-d* (*dead*). *SSD1-V* es considerado el alelo silvestre, mientras que *ssd1-d* presenta una mutación con pérdida de función por truncación de *SSD1*, siendo recesivo e inviable en combinación con la mutación de *SIT4*. *SSD1-V* es utilizado como supresor de la activación de la ruta RAS/AMPc (Wilson *et al.*, 1991).

SSD1 es un gen de 4,9 kb, que codifica una proteína que está presente tanto en el citoplasma como en el núcleo celular. Posee un amplio abanico de funciones, siendo muchas de ellas desconocidas. Ssd1 presenta un dominio de localización nuclear (*Nuclear localization signal,* NLS), y un dominio de unión a ARN (*RNA-binding domain*, RBD). El RBD de *SSD1* es un dominio con similitud a la RNAsa II carente de capacidad catalítica, pero que preserva la capacidad de interacción con el ARN (Wanless *et al.*, 2014). Se une preferentemente a ARNm destinados a una localización y transcripción polar (Kurischko, Kuravi, *et al.*, 2011), entre ellos mensajeros que codifician para proteínas del ciclo celular (Jansen *et al.*, 2009). Una de ellas es la ciclina Cln2, una de las responsables de la transición G_1-S. Ssd1 se une a la región 5' UTR (*Untranslated Region*) del mensajero de *CLN2 in vivo* en el momento de su transcripción en el núcleo, ayudando a su estabilización y a su transcripción continua bajo condiciones específicas (Ohyama *et al.*, 2010)

En condiciones de estrés, Ssd1 junto con los ARNm unidos a él migran hacia gránulos de estrés (*Stress Granules,* SGs) y hacia cuerpos P (*P-bodies*, PBs). SG y PB son estructuras sin membrana que comprenden distintos complejos citoplasmáticos de proteínas y ARNm transcripcionalmente parados. Dependiendo del ARNm, Ssd1 lo puede transportar fuera del núcleo hacia sitios de crecimiento polar (Kurischko, Kim, *et al.*, 2011; Uesono *et al.*, 1997) e interacciona directamente con SGs y PBs, formando parte de ellos. (Tarassov *et al.*, 2008; Richardson *et al.*, 2012). Pese a ello, esta función no está determinada, puesto que se desconoce el momento de la unión de Ssd1 con los mensajeros, y si es el mismo para todos ellos.

Ssd1 es sustrato esencial de Cbk1, proteín kinasa que regula Ssd1 por retroalimentación negativa (*feedback* negativo) (Wanless *et al.*, 2014) mediante la fosforilación de su extremo N-terminal. Ante la falta de los sitios de fosforilación de Ssd1 o una deleción de CBK1, la unión de Ssd1 a los PBs se hace permanente, presentando un fenotipo letal (Jansen *et al.*, 2009; Kurischko, Kim, *et al.*, 2011).

De igual modo, se ha relacionado a Ssd1 con la ruta de señalización RAM (*regulation of Ace2 and polarized morphogenesis*), que regula el crecimiento polar, la expresión diferencial de genes y el mantenimiento de la integridad celular (Racki *et al.*, 2000; Bidlingmaier *et al.*, 2001). Esta relación viene dada por unión de Ssd1 a la kinasa de la ruta RAM, ortóloga de *CBK1*, y su consecuente fosforilación (Racki *et al.*, 2000; Kurischko *et al.*, 2011). Por otro lado, las células con mutaciones en la ruta RAM presentan un fenotipo letal si se da expresión de Ssd1 (Jorgensen *et al.*, 2002; Du y Novick, 2002).

1.4. ÁCIDOS DÉBILES Y SU PAPEL EN EL CRECIMIENTO CELULAR

Se define como ácido débil aquella molécula que en solución acuosa no está totalmente disociada. Posee tanto la capacidad de donar como de adquirir protones. De esta forma, en una disolución acuosa, se genera un equilibrio ácido-base entre los protones y las bases conjugadas con la forma sin disociar. Esta cualidad es aprovechable con el fin de producir una acidificación intracelular en levadura.

Para este fin es necesario elegir aquel ácido débil cuya constante de disociación ácida (pKa) sea similar a las condiciones extracelulares, con el fin de conseguir la correcta incorporación del mismo al citosol por difusión simple. Es el caso del ácido acético, que presenta una pKa de 4,76. Ante el fenómeno de represión por glucosa, el ácido acético no es metabolizado y difunde hacia el interior celular (Ludovico *et al.*, 2001). En el citosol, en el caso de que éste sea más alcalino que el medio extracelular, el ácido acético produce una acidificación celular y una acumulación aniónica (Pampulha y Loureiro, 1989). Al producirse una bajada de pH, se compromete la viabilidad celular, inhibiéndose el metabolismo fermentativo y disminuyendo el crecimiento (Palmqvist y Hahn-Hägerdal, 2000).

1.5. CONCEPTOS BÁSICOS DE BIOINFORMÁTICA

1.5.1. RIN

La medición de la integridad del ARN es un parámetro relevante en experimentos de expresión génica. Debido a ello, su correcta determinación influye a la hora de realizar una secuenciación masiva. Previo a la aparición del valor RIN (*RNA Integrity Number*), esto se llevaba a cabo mediante el cálculo del cociente de ARNr 28S:18S, estando este método sujeto a variaciones debidas al factor humano. Por lo tanto, se carecía de reproducibilidad.

El valor RIN es, pues, una medida automática y reproducible (Schroeder *et al.*, 2006) de determinación de la integridad del ARN, calculado mediante un algoritmo. En él se tienen en cuenta los diferentes parámetros observados en el electroferograma obtenido tras la electroforesis del ARN total, que aporten información relevante sobre el estado de degradación del ARN. De esta manera, el algoritmo devuelve un valor RIN, que será más elevado conforme mayor sea la calidad del ARN.

1.5.2. FC y log_2FC

Se entiende por *Fold Change* (FC) un cociente entre dos condiciones, generalmente una muestra y un control (Dembélé y Kastner, 2014). Para RNAseq, es una comparación de los niveles de expresión de un gen en dos condiciones. Es calculado con la media de las medidas normalizadas (*Reads Per Kilobase Million* o RPKM) de las lecturas mapeadas para cada gen en cada réplica y condición, dando lugar a un valor equivalente a cómo de expresado está la condición frente al control. Por lo tanto, un FC de 1 implica que ambas condiciones tienen la misma expresión, mientras que un FC positivo o negativo indica mayor o menor expresión de la condición frente al control.

Es común transformar el FC a escala logarítmica en base 2 debido a los beneficios que ello supone, tales como presentar una distribución de error simétrica. Por lo tanto, un valor de 0 indica igual expresión, mientras que valores positivos indican mayor expresión y negativos menor. Además, esta representación ayuda a una mejor interpretación de los resultados.

1.5.3. P-VALUE Y LA NECESIDAD DE APLICAR CORRECCIÓN POR TEST MÚLTIPLE

Suponiendo un caso estadístico, se quiere comprobar si dos muestras presentan diferencias clínicas frente a dos tratamientos. Por lo tanto, se plantea un test de hipótesis donde la hipótesis nula (H_0) se define como: *no hay diferencias clínicas entre los dos tratamientos*. La hipótesis alternativa (H_1) es: *hay diferencias clínicas entre los dos tratamientos*.

El valor p (*p-value*) es una probabilidad condicional calculada asumiendo que H_0 es cierta. Es la probabilidad de que, debido al azar, los resultados pudieran dar una diferencia igual o mayor a la observada (Dorey, 2010). Un *p-value* pequeño, hace que el factor azar no sea relevante, siendo H_1 la hipótesis más probable. Es, por lo tanto, una medida de la fuerza de la evidencia contra H_0.

De la misma forma, un *p-value* alto no significa que H_0 sea cierta, dado que puede ser resultado de una población de datos pequeña. En un test de hipótesis, se utiliza el *p-value* junto con un umbral arbitrario, llamado nivel de significancia. Este umbral suele ser 0,05. Si el *p-value* es inferior a 0,05, entonces se rechaza H_0. Es equivalente a decir que un 5% de las veces que H_0 sea rechazada, será un error, un falso positivo. Sin embargo, siendo el *p-value* dependiente del tamaño poblacional, donde un test de hipótesis no se descarte H_0, aumentando el número poblacional el resultado es supceptible de variar, debido al aumento del poder estadístico.

Esto, aplicado a análisis transcriptómicos como RNAseq, supone que para cada gen se realice un test de hipótesis. En él, H_0 es: *el gen no está diferencialmente expresado entre las dos condiciones*. Y, por lo tanto, H_1 es: *el gen está diferencialmente expresado entre las dos condiciones*. Al ser el *p-value* intrínseco de cada gen, no se puede utilizar como medida comparativa entre genes dentro de la misma muestra. Es decir, se precisa de una corrección por test múltiple para ajustar la confianza estadística de las medidas al número de test independientes realizados.

1.5.4. AJUSTE DE BONFERRONI, FDR, P-VALUE AJUSTADO POR FDR Y Q-VALUE.

Para poder comparar datos procedentes de múltiples test de hipótesis independientes se desarrolló el ajuste de Bonferroni (Bland y Altman, 1995). Es uno de los ajustes más utilizados, donde se define α como el umbral de error, generalmente 0,05, y n como el número total de tests realizados. Por lo tanto, si el *p-value* es inferior o igual al cociente entre α y n, se consideran las diferencias como significativas, rechazando H_0. Este resultado es comparable entre los demás tests. De esta forma, el ajuste de Bonferroni controla el *Family Wise Error Rate* (FWER), que es la probabilidad de tener falsos positivos tras la realización de múltiples tests de hipótesis.

Por lo tanto, el ajuste de Bonferroni asegura que, para un α = 0,05, tras su aplicación haya un 95% de confianza de que ninguno de los test de hipótesis donde H_0 sea rechazada haya sido debido al azar. Utilizarlo, en el contexto de RNAseq, puede causar la exclusión de datos de potencial relevancia, dado que es una corrección por test múltiple muy estricta. En su lugar se prefiere aplicar el *False Discovery Rate* (FDR). En lugar de que se posea un 95% de confianza de que ninguna H_0 haya sido rechazada al azar, se trabaja con un porcentaje arbitrario de que, por azar, sí haya falsos positivos.

Hay distintos métodos para trabajar con FDR. El más usado calcula un *p-value* ajustado por FDR (*FDR-adjusted p-value*) a partir del *p-value* obtenido del test de hipótesis. Este es el método de Benjamini-Hochberg (Benjamini y Hochberg, 1995). Se basa en ordenar todos los *p-value* por valor ascendente, y darles un rango, siendo el *p-value* con valor inferior el rango 1, y así sucesivamente. Entonces, el *p-value* ajustado por FDR es el producto del *p-value* por el cociente entre el número total de tests y el rango. Si el resultado es inferior al FDR se rechaza la H_0, siendo un % de esos tests falsos positivos de acorde al FDR.

El método de Benjamini-Hochberg presenta también sus limitaciones. Se puede dar el caso en el que dos tests, siendo el *p-value* del primero inferior al del segundo, tras aplicar el método el primero presente un *p-value* ajustado por FDR mayor que el segundo. Ello puede complicar la interpretación de los resultados. Frente a esto, surgió el valor q (*q-value*) (Storey, 2002). El *q-value* es el mínimo *p-value* ajustado por FDR alcanzado, bien siendo el obtenido por el método de Benjamini-Hochberg, o bien siendo el mismo *p-value* ajustado por FDR calculado para el *p-value* de un rango superior, solventando el problema mencionado anteriormente.

La elección del método de corrección por test múltiple depende de la relevancia que posea un falso positivo en los análisis posteriores. En el contexto de RNAseq, en casos en que sólo se hagan experimentos posteriores con un único gen, es recomendable usar la corrección de Bonferroni, ya que es la más fiable. Asimismo, existe una alternativa, llamada FDR local (Efron *et al.*, 2001). Es la probabilidad de que un test de hipótesis particular rinda un falso positivo. No es ampliamente utilizado debido a que su cálculo es complejo (Noble, 2009).

1.6. RNAseq Y SU CONTEXTO COMO TÉCNICA TRANSCRIPTÓMICA

El estudio de todos los transcritos presentes en una célula en un momento determinado (transcriptoma) ha sido abarcado desde diversas técnicas. Las técnicas transcriptómicas analizan el transcriptoma con tecnologías de secuenciación masiva (*Next-Generation Sequencing,* NGS). Los objetivos prioritarios de estas técnicas son la catalogación de los diferentes tipos de transcritos. Esto incluye ARNm, ARN no codificantes (ncRNA) y pequeños ARN (*small RNAs*).

Además, se busca determinar la estructura génica, incluyendo sitios de iniciación, extremos 5' y 3', isoformas y modificaciones post-transcripcionales, junto con la determinación de la expresión génica y su cambio en diferentes condiciones (Wang *et al.*, 2009).

Frente a esta demanda se desarrollaron distintos tipos de tecnologías. La más antigua es la basada en hibridación. Surge en la década de los 90, y se basa en la incubación de un ADNc marcado fluorescentemente en chips (*microarrays*), tanto comerciales como no comerciales. De esta manera, se cuantifica un set de genes predefinidos mediante la hibridación de sus transcritos a sus secuencias complementarias (Schena *et al.*, 1995).

El *microarray*, dependiendo del número de secuencias presentes, puede ser de alta o baja densidad génica. Al examinar un número elevado de genes, se abaratan los costes (Heller, 2002). Sin embargo, los *microarrays* presentan el inconveniente de poder analizar únicamente genes conocidos anteriormente, debido a la necesidad de diseñarlos previamente como sondas.

El segundo tipo de tecnologías se basan en la secuenciación. Sus inicios radican en la técnica *Serial Analysis of Gene Expression* (SAGE), en la cual se secuencian fragmentos de transcritos aleatorios de forma concatenada (Velculescu *et al.*, 1995). Estos fragmentos se cuantificaban por su mapeo a genes conocidos. Esta técnica fue, sin embargo, dejada de lado con el surgimiento de nuevas técnicas de alto rendimiento (*High Throughput*) de secuenciación de transcritos.

En este marco destaca RNAseq. Esta técnica fue influenciada por predecesoras como *Massive Parallel Signature Sequencing* (MPSS), basada en la generación de secuencias de 16-20 pb mediante una serie de hibridaciones (Brenner *et al.*, 2000). MPSS presenta un rendimiento suficiente para cuantificar la expresión de genes de *Arabidopsis Thaliana* (Meyers *et al.*, 2004).

Con el surgimiento de RNAseq, el rendimiento inicial era de 10^5 lecturas utilizando la NGS 454 (Bainbridge *et al.*, 2006), siendo suficiente para cuantificar la expresión del transcriptoma. Finalmente, la tecnología fue popularizándose y es usada a día de hoy como técnica modelo para estudios transcriptómicos con el surgimiento de Illumina, tecnología de secuenciación masiva capaz de generar 10^9 lecturas (Wilhelm *et al.*, 2008). Este rendimiento permite la correcta cuantificación del transcriptoma de una especie, el ensamblaje *de novo* de transcriptomas, junto con la detección e identificación de nuevos genes con transcritos poco abundantes, entre otros.

1.7. TIPOS DE RNAseq

Dentro del análisis transcriptómico por RNAseq, el tipo de RNAseq viene determinado por el objetivo final de lo secuenciado. Puede darse el caso de querer generar un transcriptoma *de novo* (Grabherr *et al.*, 2011), mediante el ensamblaje de las lecturas. Una vez ensamblado el transcriptoma, se anota para obtener la función de los genes identificados.

Otra opción es la detección de genes diferencialmente expresados (*Differentially Expressed Genes* o DEG), mediante la medición de las lecturas de los ARNm que mapean frente a un genoma o transcriptoma de referencia. De igual manera, se pueden identificar y cuantificar distintas isoformas del mismo gen (Hartley y Mullikin, 2016). Asimismo, se puede utilizar esta técnica para la cuantificación de ARN poco frecuentes, como miARN (Friedländer *et al.*, 2008). Todas estas funciones pueden ser combinadas, si en el diseño del experimental se ha tenido en cuenta las necesidades de cada una de ellas.

1.8. FLUJO DE TRABAJO EN UN RNAseq ENFOCADO A LA DETECCIÓN DE EXPRESIÓN DIFERENCIAL

Un experimento de RNAseq cuyo fin es la determinación de DEGs cuenta con la presente serie pasos.

1.8.1. DISEÑO EXPERIMENTAL

1.8.1.1. RÉPLICAS BIOLÓGICAS

Para experimentos de expresión diferencial, un mayor número de réplicas ayuda a la obtención de datos más robustos estadísticamente (Auer y Doerge, 2010). Por el contrario, experimentos de RNAseq enfocados al ensamblaje de novo del transcriptoma o a la identificación de genes o isoformas nuevas, no precisarán de un número elevado de réplicas biológicas, siendo un mínimo de tres para poder realizar análisis con inferencia en la población (Conesa *et al.*, 2016).

1.8.1.2. PURIFICACIÓN DEL ARNr

Una vez diseñado el experimento en las condiciones y réplicas deseadas, se procede a la extracción del ARN total. Es necesario purificar el ARNm. Para ello, existen varias alternativas para descartar el ARNr, que constituye alrededor del 90% del ARN total. Esto se puede conseguir seleccionando los ARNm por la cola Poli-A, o bien eliminando el ARNr. El primero requiere de ARN de alto valor de RIN, y produce un mayor rendimiento de lecturas mapeadas. Para muestras con calidad inferior o bacteriológicas, se requiere del segundo método (Conesa *et al.*, 2016).

1.8.1.3. TIPO DE LIBRERÍA

Existen diversos protocolos para la generación de librerías de RNAseq. Una opción es la retrotranscripción de los ARNm mediante *random hexamer priming*. Este se considera un método *unstranded*, al no conservar información específica de cadena (Mortazavi *et al.*, 2008). La otra opción es la generación de librerías que sí retengan esta información, llamadas *stranded*. Esto se realiza mediante protocolos como el método de dUTP (Levin *et al.*, 2010), donde son incorporados en la síntesis de la segunda cadena del ADNc (Parkhomchuk *et al.*, 2009).

1.8.1.4. TIPO Y TAMAÑO DE LECTURAS

Las lecturas *paired-end* son las más adecuadas a la hora del descubrimiento de nuevos transcritos o del análisis de expresión de diferentes isoformas de los genes (Katz *et al.*, 2010; Garber *et al.*, 2011). De igual manera, lecturas más largas ayudan a una mejor identificación de los transcritos y mapeo (Garber *et al.*, 2011; Łabaj *et al.*, 2011). Por lo tanto, se decanta por *single-end* para los análisis de expresión diferencial en organismos con un genoma de referencia de calidad, dado que es más económico (Conesa *et al.*, 2016).

1.8.1.5. PROFUNDIDAD DE SECUENCIACIÓN

Un número mayor de transcritos serán detectados y cuantificados con mayor precisión cuanto mayor sea la profundidad de secuenciación (Mortazavi *et al.*, 2008). La cantidad exacta varía entre autores, siendo de 5 millones de lecturas para la detección de transcritos altamente expresados a 100 millones para los de menor expresión (Sims *et al.*, 2014). Para experimentos de RNAseq de célula única (*Single-cell RNAseq*), se utilizan un millón lecturas, siendo 50.000 suficientes para los genes mayor expresados (Pollen *et al.*, 2014), y 20.000 para la diferenciación de tipos celulares dentro del mismo tejido (Jaitin *et al.*, 2014). Por el contrario, una elevada profundidad de secuenciación puede conllevar un aumento del ruido a la hora del análisis de datos (Tarazona *et al.*, 2011).

Por lo tanto, las recomendaciones son de 10-25 millones para la detección de expresión diferencial, 50-100 millones para la detección e identificación de *splicing* alternativo y expresión específica de alelo, y más de 100 millones para un ensamblaje *de novo* de transcriptoma (Liu *et al.*, 2013; Liu *et al.*, 2014).

1.8.2. SECUENCIACIÓN

Existen diversas NGS, clasificadas en generaciones. Entre las más utilizadas se incluyen *Illumina*, *Ion Torrent*, y *PacBio*. Recientemente se está popularizando la NGS de tercera generación *Nanopore* en un dispositivo capaz de medir ADNc comercializado por *Oxford Nanopore Technologies* (ONT) (Byrne *et al.*, 2017).

El fundamento de cada NGS varia, pudiendo ser la detección de los nucleótidos fluorescentes incorporados (*Illumina*); de nucleótidos fluorescentes ligados a fósforo (*PacBio*). También pueden estar basadas en *linkers* fluorescentes (*SOLiD*); en la liberación de subproductos derivados de la incorporación de nucleótidos (454); en la detección de fluorescencia o de cambios de pH (*Ion Torrent*) (Buermans y den Dunnen, 2014; SEQC/MAQC-III Consortium, 2014).

Cada una de ellas tiene sus ventajas e inconvenientes. 454, *PacBio* e *Ion Torrent* se caracterizan por lecturas más largas pero menor rendimiento en lecturas. Lo contrario se aplica a *Illumina* y *SOLiD*, siendo estas tecnologías más económicas. De entre todas ellas, *Illumina* es la más utilizada y la que presenta mayor cantidad de *software* analítico disponible a día de hoy (Jazayeri *et al.*, 2015).

1.8.3. TRATAMIENTO DE DATOS

1.8.3.1. LIMPIEZA DE SECUENCIAS Y CONTROL DE CALIDAD

Normalmente, la calidad de las secuencias decrece conforme se avanza hacia el extremo 3'. Este fenómeno ya era presente en la secuenciación por Sanger. Cuando las bases poseen una mala calidad, o la calidad conjunta de la secuencia es mala, es necesario eliminarla para el correcto mapeo de las lecturas. Lo mismo ocurre con los adaptadores utilizados durante la secuenciación. Para ello existen programas como *FASTX-Toolkit* (FASTX-TOOLKIT, 2018), *Trimmomatic* (Bolger *et al.*, 2014) y *CUTADAPT* (Martin, 2011). De esta manera se generan lecturas limpias.

Por otro lado, estas lecturas han de ser sometidas a distintos controles de calidad. Entre ellos destacan el de calidad de secuencia, contenido en GC, presencia de adaptadores, *k-mers* (todos los posibles fragmentos de longitud k resultantes de una secuencia) sobrerrepresentados y lecturas duplicadas para detectar errores de secuenciación, artefactos de PCR y contaminaciones (Conesa *et al.*, 2016). Cada plataforma de secuenciación requiere de controles similares pero personalizados. Por ello, *FastQC* (Andrews, 2014) es el programa más utilizado para el control de calidad de las secuencias provenientes de *Illumina*. Existen también otros más universales, como el *NGSQC* (Dai *et al.*, 2010), aplicable a cualquier plataforma NGS.

1.8.3.2. MAPEO DE SECUENCIAS

En este punto difieren los distintos protocolos de RNAseq. Con las lecturas limpias que han pasado el control de calidad se pueden identificar nuevos genes mediante la identificación de sus transcritos. Asimismo, se puede realizar una reconstrucción del transcriptoma *de novo* a partir de las lecturas de *Illumina*, gracias su elevada profundidad de secuenciación. De igual manera, ello provoca que, debido a la longitud reducida de las lecturas, el ensamblaje se detenga a la altura de un número elevado de *contigs*, siendo necesaria información complementaria provista por otras de las NGS mencionadas anteriormente.

En el caso del mapeo sobre genomas de referencia anotados correctamente, se pueden utilizar distintos tipos de mapeadores. Estos han de tener en cuenta la posibilidad de que exista *splicing* alternativo. Asimismo, dentro de este fenómeno se presentan también las uniones no canónicas (*non-canonical junctions*) (Ding *et al.*, 2017) y ARN quiméricos, transcritos de genes de fusión (Kumar *et al.*, 2016). Uno de los mapeadores más conocidos es *TopHat*/*TopHat2* (Trapnell *et al.*, 2009; Kim *et al.*, 2013). *TopHat* realiza dos pasos secuenciales. El primero es el mapeo de las secuencias que no presentan *splicing* para localizar los exones. Después, las lecturas restantes se dividen y son mapeadas independientemente con el fin de delimitar el sitio de *splicing* (Trapnell *et al.*, 2009; Kim *et al.*, 2013). Actualmente, la popularidad de *TopHat2* ha disminuido en beneficio de *HISAT*/*HISAT2* (Kim *et al.*, 2015).

A su vez, existen mapeadores más específicos capaces de identificar polimorfismos de nucleótido único (SNPs) e inserciones-deleciones (INDELs), como *GSNAP* (Wu y Nacu, 2010), *PALMapper* (Jean *et al.*, 2010) o *MapSplice* (Wang *et al.*, 2010). Otros destacan por detectar *splicing* no canónicos, como *STAR* (Dobin *et al.*, 2013) o *MapSplice*; o por su rapidez, como *STAR* o *GEM* (Marco-Sola *et al.*, 2012); o por mapear lecturas largas, como *STAR*.

En todos ellos, parámetros como si las lecturas provienen de una librería *stranded* o *unstranded*, el número de *mismatches* a tolerar, la longitud de las lecturas y si estas son *single-end* o *paired-end* han de ser tenidos en cuenta para el correcto mapeo de los datos limpios. De igual manera, los mapeadores pueden ser asistidos por modelos génicos, cuando estén presentes, los cuales aportan ficheros de anotación con coordenadas sobre los sitios de *splicing* y los exones, pudiendo tener un impacto en los resultados finales (Zhao y Zhang, 2015). Estas anotaciones pueden ser encontradas en *ensembl* (ENSEMBL GENOME BROWSER, 2018), *RefSeq* (REFSEQ: NCBI REFERENCE SEQUENCE DATABASE, 2018), *UCSC* (UCSC GENOME BROWSER HOME, 2018) y, para levaduras, *SGD* (SACCHAROMYCES GENOME DATABASE | SGD, 2018), entre otros.

1.8.3.3. CONTEO DE LAS LECTURAS

El conteo de las secuencias es un paso clave para determinar la expresión de los genes. Existen diversas formas de hacer este paso. La primera es hacer uso del contaje de *k-mers* en las lecturas sin necesidad de recurrir a un mapeo previo de las secuencias. Esto lo llevan a cabo programas como *Sailfish* (Patro *et al.*, 2014). La alternativa a esto es realizada por programas como *HTSeq-count* (Anders *et al.*, 2015) o *featureCounts* (Liao *et al.*, 2014) que se basan en el mapeo de las secuencias a un genoma de referencia, cuantificando las lecturas a nivel de gen, generándose un fichero en formato *Gene Transfer Format* (GTF). Este fichero contiene las posiciones de los exones y genes.

1.8.3.4. CONFIRMACIÓN DE LA REPRODUCIBILIDAD

De igual manera, es aconsejable realizar un análisis para confirmar que las réplicas se comportan como tal. De modo que, si se poseen diversas réplicas biológicas de cada condición, se espera que éstas se comporten de manera similar entre ellas. Por ello, se realiza un análisis de componentes principales o *Principal Component Analysis* (PCA).

Un PCA es un análisis multidimensional donde cada gen conforma una dimensión. Se intenta reducir el número de dimensiones, buscando aquellas direcciones del espacio que expliquen el mayor porcentaje de la varianza entre las muestras. Estas direcciones se llaman componentes principales o *Principal Component* (PC). Lo esperado es ver cómo las muestras se agrupan por condiciones, pudiendo detectar de esta manera aquellas que no lo hagan, y desestimarlas de cara a análisis posteriores (*outliers*). Se consideran como buenas aquellas muestras con un coeficiente de Spearman $R^2 > 0,9$ (Mortazavi *et al.*, 2008).

1.8.3.5. ANÁLISIS DE LA EXPRESIÓN DIFERENCIAL

Para el análisis de la expresión diferencial, se utilizan programas diseñados específicamente para el tipo de RNAseq llevado a cabo. La estadística tras cada uno de ellos es distinta, incluso para aquellos dirigidos al mismo tipo.

1.8.3.5.1. Sesgos en los datos de secuenciación

Hay una serie de sesgos inherentes a RNAseq, que actúan en mayor o menor medida desvirtuando los datos y pudiendo provocar una incorrecta determinación de genes diferencialmente expresados.

La profundidad de secuenciación es un sesgo fácil de identificar y comprender, debido a que es posible que diferentes condiciones tengan diferente profundidad de secuenciación, distinto número de lecturas. Esto puede provocar que, atendiendo únicamente al número de lecturas mapeadas, si una condición tiene el doble de profundidad de secuenciación que otra, un gen que no tenga diferencias de expresión entre ellas aparezca diferencialmente expresado con un FC de 2. Por ello, es uno de los sesgos más tenidos en cuenta y aplicados a todos los algoritmos de normalización de datos de RNAseq.

Otro es el sesgo por la composición del ARN. Si un grupo de genes presentan expresión en una de las dos muestras (debido a, por ejemplo, que sea una expresión intrínseca de un tejido concreto) o están altamente expresados en una condición, debido a que la profundidad de lectura es fija para todas las condiciones, la cantidad real de lecturas a repartir entre los genes restantes es menor. De no ser normalizado, se generan falsos positivos debido a que se consideran genes como diferencialmente expresados en una condición respecto a otra, cuando en verdad presentan el mismo nivel de expresión. Para corregirlo, se precisan de métodos de normalización entre muestras que partan de los datos crudos (Robinson y Oshlack, 2010).

Asimismo, la secuencia del gen puede ser fruto de sesgos. La relación entre la secuencia de los genes y su expresión es un objeto importante de estudio. En mamíferos, ésta se puede correlacionar con metilación del ADN (Jabbari y Bernardi, 1998), recombinación del ADN (Kong *et al.*, 2002) y densidad génica (Lander *et al.*, 2001), entre otras. Por el contrario, no hay una relación clara entre la secuencia del gen y el patrón de expresión génica. Existe una correlación entre el contenido en GC de la secuencia y el producto obtenido en la amplificación por PCR en el proceso de secuenciación masiva (Dohm *et al.*, 2008). A nivel de estudio transcriptómico, el uso de cebadores (*primers*) hexaméricos aleatorios (*random hexamer priming*), muy usado para la creación de las librerías de RNAseq, se ha demostrado el causante de este sesgo (Hansen *et al.*, 2010). Al ser un sesgo debido al método y no al material biológico en sí, se pudo desarrollar un procedimiento para corregirlo (Zheng *et al.*, 2011).

La longitud del transcrito es uno de los sesgos más importantes de RNAseq (Wang *et al.*, 2012). Esto es debido a que es común utilizar métodos de fragmentación del ARNm a fin de obtener una mayor cobertura de la secuencia de todo el transcrito. Por ello, a mayor sea la longitud del gen, mayor será el número de transcritos. Por lo tanto, el número total de lecturas para un transcrito es proporcional al nivel de expresión del transcrito y a la longitud del mismo, siguiendo una distribución de Poisson (Oshlack y Wakefield, 2009). En cambio, en *microarrays* este sesgo no aparece, debido a la relación proporcional de la intensidad de las medidas con el nivel de expresión del gen y con factores intrínsecos del experimento, como el contenido en GC (Dunning *et al.*, 2008; Wu y Irizarry, 2005).

Otro sesgo es la incertidumbre en el mapeo de las lecturas. Es debido a que, en la actualidad, siendo *Illumina* una de las plataformas de secuenciación masiva más utilizadas (Reuter *et al.*, 2015), la longitud de las lecturas no abarca la longitud total del transcrito. Esto puede dar lugar a que lecturas puedan mapear en genes distintos. Esto es común para genes parálogos, familias génicas y distintas isoformas del mismo gen derivadas por splicing alternativo (Li *et al.*, 2010). Frente a estas lecturas ambiguas, se pueden descartar (Marioni *et al.*, 2008; Morin *et al.*, 2008), o bien se pueden reubicar estas lecturas ambiguas en los genes acorde a la proporción de cobertura otorgado por las lecturas con único mapeo (Faulkner *et al.*, 2008; Mortazavi *et al.*, 2008).

1.8.3.5.2. Normalización de los datos crudos

1.8.3.5.2.a. RPKM, FPKM y TPM

Los datos crudos derivados de la secuenciación masiva están sujetos a los sesgos mencionados en el apartado anterior. Por lo tanto, estos datos no pueden ser usados para el análisis de expresión diferencial de genes y la generación de un modelo de normalización preciso para todos los sesgos es difícil (Tuerk *et al.*, 2017).

Los sesgos más importantes son la longitud del transcrito y la profundidad de secuenciación (Wang *et al.*, 2012). Ante la necesidad de normalizar los datos crudos de la secuenciación surgieron distintos métodos.

RPKM (*Reads Per Kilobase of exon model per Million reads*) es una normalización por longitud del gen seguido de por profundidad de secuenciación. Estas medidas reflejan la concentración molar de los transcritos, facilitando la comparación entre genes y entre distintas muestras (Mortazavi *et al.*, 2008).

FPKM (*Fragments Per Kilobase of transcript per Million fragments sequenced*) es una normalización análoga a RPKM, teniendo en cuenta los mismos parámetros. La diferencia entre los dos modelos radica en que FPKM está diseñado para tener en cuenta que varias lecturas pueden provenir de la misma molécula. Es decir, se usa para lecturas *paired-end* y podría servir más allá de ellas, dando la posibilidad de su uso en futuras técnicas de secuenciación masiva que requieran de otro tipo de lecturas (Trapnell *et al.*, 2010). Por lo tanto, para lecturas *single-end*, RPKM y FPKM son idénticos (Conesa *et al.*, 2016). De hecho, actualmente se está proponiendo un método alternativo, TPM (*Transcripts per Million*), diseñado para eliminar los errores derivados del cálculo tanto de RPKM como de FPKM (Wagner *et al.*, 2012).

1.8.3.5.2.b. Métodos a partir de datos crudos

Pese a lo descrito anteriormente, los tres métodos anteriores fallan en muestras con alta variabilidad entre transcritos, donde haya genes con una elevada expresión o un cambio en su expresión que desvirtúen el conteo de las lecturas para los demás genes (Bullard *et al.*, 2010). Existen métodos de normalización alternativos, que excluyen dichos genes (Conesa *et al.*, 2016). Entre estos métodos destacan: TMM (*Trimmed Mean of M-values*) (Robinson y Oshlack, 2010) implementado en el programa *edgeR* (Robinson *et al.*, 2010), RLE (*Relative Log Expression*) implementado en *DESeq*/*DESeq2* (Anders y Huber, 2010; Love *et al.*, 2014); y MRN (*Median Ration Normalization*) (Maza *et al.*, 2013) . Pese a ser diferentes métodos de normalización, se ha comprobado tanto con datos reales como generados artificialmente que TMM y RLE rinden resultados similares (Dillies *et al.*, 2013; Maza *et al.*, 2013; Reddy, 2015; Rapaport *et al.*, 2013; Li *et al.*, 2015).

Asimismo, existen paquetes de R como *NOISeq R* (Tarazona *et al.*, 2011), cuya función es la determinación de los distintos tipos de sesgos presentes en la muestra y, de acorde con ello, realizar las normalizaciones oportunas.

1.8.3.5.3. Resultados

Tras dejar correr el programa elegido, se genera un fichero de datos separado por tabulaciones (.tsv) en el cual, para cada gen, se muestran distintos valores, entre ellas el *p-value*, FDR o p-value ajustado a test múltiple y el $\log_2$FC.

1.9. INTERPRETACIÓN DE LOS RESULTADOS

1.9.1. VISUALIZACIÓN DEL MAPEO

El proceso de visualización del mapeo es común a otras técnicas ómicas, pudiéndose ver el mapeo a nivel de lecturas usando *ReadXplorer* (Hilker *et al.*, 2014) o bien a nivel de cobertura total, mediante el navegador *Integrative Genomics Viewer* (IGV) (Thorvaldsdóttir *et al.*, 2013) o el *UCSC* (Kent *et al.*, 2002). Otros visualizadores usualmente utilizados son *Savant*/*Savant2* (Fiume *et al.*, 2010; Fiume *et al.*, 2012) y *Genome Maps* (Medina *et al.*, 2013). Existen aquellos especialmente diseñados para análisis transcriptómicos. Es el caso de *RNASeqViewer* (Rogé y Zhang, 2014), el cual permite ver las lecturas mapeadas a nivel de exones, de isoformas y visualizar aquellos transcritos provenientes de genes de fusión. Pese a esto, es más lento que *IGV*, lo que resulta en su menor uso.

1.9.2. ELECCIÓN DE LOS GENES DIFERENCIALMENTE EXPRESADOS

Una vez realizado el proceso de detección de expresión diferencial, es criterio de laboratorio la elección de los umbrales a partir de los cuales seleccionar los genes diferencialmente expresados. La tendencia es la elección de un umbral de *p-value* ajustado por FDR menor o igual a 0,05 (Noble, 2009). Otros autores se decantan por un umbral equivalente de *q-value* (He *et al.*, 2018).

La elección del umbral de FC es más libre. Siendo por elección popular un $\log_2$FC mayor o menor a 2 (Costa-Silva *et al.*, 2017). Existen casos en los que no se busca una expresión diferencial elevada. Entonces no se fija un umbral para este valor, y se analizan todos aquellos genes estadísticamente diferenciados.

1.9.3. DETECCIÓN DE FUNCIONES DIFERENCIALMENTE EXPRESADAS

La detección de funciones diferencialmente expresadas, o *functional profiling*, es generalmente el paso final del análisis transcriptómico. Para ello es necesario un genoma correctamente anotado. Información sobre la anotación de genomas modelo puede encontrarse en *Gene Ontology* (Ashburner *et al.*, 2000), *Bioconductor* (Huber *et al.*, 2015), *DAVID* (Huang *et al.*, 2009) y *Babelomics* (Medina *et al.*, 2010). Se ha generado software específico para RNAseq capaz de identificar las funciones correspondientes a cada gen diferencialmente expresado y analizar si hay alguna función diferencialmente expresada entre ellos. Ejemplos de estos programas son *GOseq* (Young *et al.*, 2010), *Gene Set Variation Analysis* (GSVA) (Hänzelmann *et al.*, 2013) y *SeqGSEA* (Wang y Cairns, 2013).

2. OBJETIVOS

Con el presente trabajo, se pretende alcanzar los siguientes objetivos:

- La disertación de la función de *SSD1* mediante el análisis de expresión diferencial por RNAseq entre una cepa silvestre y otra con pérdida de función de *SSD1* en presencia y ausencia de ácido acético.

- La caracterización de los perfiles de crecimiento de las cepas de *Saccharomyces cerevisiae* utilizadas en el experimento de análisis de expresión diferencial mediante *Bioscreen*.

3. MATERIAL Y MÉTODOS

3.1. CEPAS DE *SACCHAROMYCES CEREVISIAE* Y CONDICIONES DE CRECIMIENTO

En los experimentos llevados a cabo se utilizan diversas cepas. La cepa de origen proviene de la colección del profesor Ramón Serrano, con identificador RS132. Es una cepa derivada de la cepa silvestre S288C. Fue transformada con el plásmido pRS987, descrito en el apartado 3.2, a fin de obtener las cepas RS637 y RS639. De ahora en adelante, dichas cepas serán referidas como cepa *SSD1-V* y *ssd1-d*, respectivamente.

Tabla 1: Cepas y plásmidos de *Saccharomyces cerevisiae* utilizados.

Cepas y plásmidos	Genotipo
RS132	*MATα, ade1-100, his4-519, ura3-52, leu2-33, 112::LEU2*
RS637 (*SSD1-V*)	RS132 + pRS987-*SSD1-V*
RS639 (*ssd1-d*)	RS132 + pRS987-*ssd1-d*
pRS987	URA3

Las levaduras se crecen en diferentes medios líquidos según las necesidades del experimento, a 28ºC y agitación continua de 200 rpm.

El medio utilizado es el siguiente:

- Medio mínimo o SD: Se formula con 0,67 % (p/v) de *Yeast Nitrogen Base*, 2 % (p/v) de glucosa, 1 % (v/v) de His 100X (3 mg/mL) y 1 % de Ade 100X (3 mg/mL), disueltos en agua *miliQ*.

El medio SD se lleva a pH 4 o pH 6 con un 10% (v/v) de ácido succínico 0,05M corregido con Tris-base hasta pH 4 o pH 6, dependiendo de las condiciones del experimento. Posteriormente, se autoclavan los medios durante 20 minutos a 120 ºC y 1 atm para su esterilización. Tras esto, se complementa el medio SD adicionando 1 % (v/v) de His 100X (3 mg/mL) y 1 % (v/v) de Ade 100X (3 mg/mL).

3.2. PLÁSMIDO PRS987

El plásmido pRS416, incluido como pRS987 en la colección del profesor Ramón Serrano, no es comercializado actualmente. Es un plásmido centromérico, con un MCS derivado del plásmido pBlueScript II, con un tamaño de 4,9 Kb, diseñado con el marcador URA3 para la selección de células transformantes. Se utiliza con el inserto del gen *SSD1-V* en la cepa RS637 y con el del gen *ssd1-d* en la cepa RS639 (Tabla 1).

3.3. OBTENCIÓN DE PERFILES DE CRECIMIENTO MEDIANTE *BIOSCREEN*

El *Bioscreen C* (Oy Growth Curves AB Ltd., Finlandia) es un dispositivo diseñado para la medición automática de la absorbancia en placas multipocillo (*Honeycomb Microplate*, 100 pocillos) a intervalos de tiempo regulares y definibles. La agitación es ajustable.

En el caso del experimento, se establecen medidas cada 20 minutos durante un intervalo de 72 h. Para iniciar el ensayo, se realiza un precultivo con las cepas a estudiar y, al alcanzar la fase estacionaria, se mide la absorbancia.

Con este dato se calcula el volumen necesario de inóculo para que en todas las muestras la absorbancia inicial sea 0,05 (OD_0). Cada condición se lleva a cabo en triplicado, administrando 350 µL de medio en las condiciones deseadas y el volumen calculado de inóculo.

Tras transcurrir 72 h, se genera un fichero Excel con las absorbancias medidas. Se realiza la media de cada triplicado y se corrige su error con la siguiente fórmula:

$$OD_{Corregida} = OD_{Promedio} + 0{,}449 * OD^2_{Promedio} + 0{,}191 * OD^3_{Promedio}$$

Al diferir el primer valor con respecto a OD = 0,05, se corrige su desviación para hacer coincidir ambos valores, aplicándolo de la misma forma al resto de valores. Se realiza el logaritmo en base 10 de las medidas a fin de poder definir parámetros de análisis de la curva de crecimiento. Dicha curva está caracterizada por una fase de latencia, una fase exponencial ajustable a una recta, y una fase estacionaria. La pendiente de la ecuación de la recta de la fase exponencial se define como la velocidad de crecimiento (µ, h^{-1}).

Se determina de manera gráfica la fase exponencial de cada condición y se obtiene la pendiente. Se elige una condición como control, cuyo coeficiente de la recta generada equivale al 100 % de velocidad de crecimiento y se realiza una comparación uno a uno de las pendientes, obteniendo así el resto. De manera análoga, se toma como rendimiento el valor máximo de OD de la condición anteriormente elegida como control, y se obtiene el rendimiento relativo en porcentaje de las demás condiciones en función del control por comparación entre dicho valor y el de la condición a ese mismo tiempo.

3.4. RNAseq

3.4.1. EXTRACCIÓN DE ÁCIDOS NUCLEICOS TOTALES

Para la extracción y posterior purificación de ARN se utiliza un protocolo adaptado (Li et al., 2009). Se dejan crecer las células hasta una absorbancia entre 0,4 y 0,6 en 50 mL del medio requerido. Se coleccionan las células centrifugando 5 minutos a 2.000 rpm en frío, a 4 ºC.

Se elimina el sobrenadante y se añaden 400 µL de *Isolation Buffer* (5 % (p/v) SDS, 10 mM EDTA, 50 mM Tris-HCl pH6). Mezclar. Se incuban las muestras 5 minutos a 65ºC y se pasan a hielo rápidamente para dar un shock térmico. Se añaden 200 µL de KCl 0,3 M pH6 y se centrifuga 5 minutos a 12.000 rpm a 4 ºC. Se recupera el sobrenadante. Se adicionan 500 µL de PCl (Fenol:Cloroformo:Isoamilalcohol (25:24:1) pH 3,8). Se centrifuga 5 minutos a 12.000 rpm a 4 ºC. Se retira el sobrenadante y se mide el volumen recogido. Se precipitan los ácidos nucleicos adicionando en este orden: 1/10 parte de NaAc 3M y 25/10 partes de etanol absoluto en un tubo de 2 mL. Se guardan las muestras desde media hora hasta tiempo indefinido a -20ºC mientras se realiza la precipitación.

A continuación, se centrifuga la muestra precipitada 10 minutos a 12.000 rpm y 4 ºC. Se quita el sobrenadante y se lava con 1 mL de etanol 70%. Se vuelve a centrifugar 5 minutos a 12.000 rpm y 4ºC. Se elimina el sobrenadante. Se seca el *pellet* y se centrifuga al vacío 5 minutos.

El tiempo puede depender del volumen residual de etanol que haya quedado en el tubo. Finalmente, se resuspende el pellet en 100 µL de agua *milliQ* libre de RNAsas (*DEPC Water*).

3.4.2. PURIFICACIÓN DEL ARN TOTAL

Para la purificación del ARN total, se utiliza un protocolo modificado del ofrecido por el kit *E.Z.N.A PLANT RNA KIT* de la casa comercial *OMEGA bio-tek*. A partir de la muestra extraída de ácidos nucleicos totales, se mide la concentración de las muestras en un espectrofotómetro *Nanodrop*. Se prepara una mezcla de 75 µL por cada columna utilizada (73,5 µL de buffer DNAsa I y 1,5 µL de DNAsa I). Se prepara otra mezcla con 100 µL de ácidos nucleicos totales, 300 µL de tampón RB y 300 µL de etanol 70 % y se adiciona a la columna. Se centrifuga a 12.000 rpm durante 1 minuto y se descarta el filtrado. De esta forma, los ácidos nucleicos se adhieren a la columna. Se añaden los 75 µL de la mezcla de DNAsa I para eliminar el ADN y se deja incubar durante 15 minutos a temperatura ambiente.

Se añaden 250 µL de *RNA Wash 1*, se espera 2 minutos y se centrifuga a 10.000 rpm durante 1 minuto, para descartar posteriormente el filtrado. Se añaden 700 µL de *RNA Wash 2* y se centrifuga durante 30 segundos a 10.000 rpm. Se repite este paso dos veces. Se centrifuga la columna a máxima velocidad durante 2 minutos y se transfiere ésta a un tubo de 1,5 mL. Para eluir el ARN, se añaden 50 µL de agua libre de RNAsas y se espera 1 minuto antes de centrifugar a máxima velocidad durante 1 minuto. De estos 50 µL, se puede hacer una alícuota de 15 µL para medir su concentración en un espectofotómetro *Nanodrop* y hacer un posterior gel de electroforesis. El resto se guarda a -70ºC.

3.4.3. SERVICIO DE SECUENCIACIÓN DE LA UV

El servicio contratado para la secuenciación del transcriptoma es el ofertado por la sección de genómica del *Servei Central de Suport a la Investigació Experimental* (SCSIE), ubicado en la *Universitat de València* (UV). Comprende una serie de ensayos.

- Comprobación de la calidad y cuantificación del ARN con *Bioanalyzer 2100*, aceptando como ARN de buena calidad aquellas muestras con un RIN superior a 7.
- Enriquecimiento de la muestra en ARNm usando oligoT.
- Preparación de librerías que no retienen información específica de cadena (*unstranded*) con el kit *TruSeq RNA Library Prep Kit v2*.
- Acorde a la plataforma de secuenciación *Illumina*, se realiza una PCR en emulsión para la formación de una agrupación de secuencias idénticas (*cluster*).
- Secuenciación utilizando la máquina *NextSeq 500*, generando un fichero de datos crudos que es tratado a fin de generar un fichero FastQ.

3.4.4. TRATAMIENTO DE DATOS

El tratamiento de los datos recibidos del servicio de secuenciación se lleva a cabo en el Servicio de Bioinformática del IBMCP a cargo del profesor Javier Forment. Se trata de una serie de pasos secuenciales:

- Análisis de la calidad de las secuencias utilizando el programa *FastQC*.
- Limpieza de adaptadores y secuencias de baja calidad utilizando el programa *CUTADAPT*.
- Mapeo de las lecturas en el genoma de referencia utilizando *TopHat2/HISAT2*.

- Generación de la matriz de conteo para cada gen en cada condición utilizando *HTSeq-count*. Es un fichero en formato .tsv, un archivo con datos separados por tabulaciones, donde se muestran los valores brutos del número de veces que las secuencias han mapeado en los genes para cada condición. A dicho valor se le realiza una normalización para obtener el valor RPKM y poderse comparar entre sí en el paso siguiente.
- Comprobación en R de la reproducibilidad de las muestras mediante análisis PCA.
- Análisis de la expresión diferencial utilizando *edgeR*/*DESeq2*. Estos programas generan un fichero .tsv para cada comparación realizada entre condiciones, rindiendo aquellos genes diferencialmente expresados y los que no lo están por los parámetros *p-value* ajustado (*DESeq2*) y FDR (*edgeR*).

3.4.5. ANÁLISIS DE LOS RESULTADOS DE LA EXPRESIÓN DIFERENCIAL

3.4.5.1. GENERACIÓN DE SCRIPTS

A partir de los ficheros de expresión diferencial en formato .tsv, se generan en lenguaje *Python* programas (*scripts*) que son ejecutados en *Linux* mediante la orden "*python [nombre del script] [nombre del primer fichero de datos] [nombre del segundo fichero de datos] (para comparaciones dobles)*". Cada *script* es diseñado con el fin de extraer unos datos concretos del fichero de expresión diferencial. Para cada programa de análisis de expresión diferencial utilizado, se precisan scripts diferentes, que pueden ser encontrados en Anexos, el apartado 7.1.

3.4.5.1.1. Comparaciones simples

Se obtienen aquellos genes diferencialmente expresados, seleccionándose aquellos con un *p-value* ajustado por FDR (llamado FDR en *edgeR* o *p-value* ajustado en *DESeq2*) menor a 0,05. Para estos genes, se obtienen sus identificadores de levadura, y los valores correspondientes al *p-value* y al $\log_2$FC. El $\log_2$FC se transforma, por propiedad fundamental de los logaritmos, en FC. Cada gen diferencialmente expresado es transferido a un nuevo fichero .tsv, manteniendo el orden de columnas: "*ID/Gene/Function/p-value/X/log2FC/FC/Interpretation*", siendo X *Adjusted p-value* o FDR dependiendo del programa utilizado.

De manera análoga, se obtienen aquellos genes no diferencialmente expresados. El fichero obtenido presenta el orden de columnas siguiente: "*ID/Gene/Function/p-value/X*".

3.4.5.1.2. Comparaciones dobles

Comparación doble entre dos ficheros para la obtención de aquellos genes diferencialmente expresados en una condición, pero no diferencialmente expresados en otra. Para ello, y de manera análoga al apartado anterior, se lee cada fichero para obtener para cada uno de ellos un diccionario donde la *key* sea el identificador de levadura de cada gen con un *p-value* ajustado o FDR menor a 0,05 en una condición y mayor a 0,05 en la otra. El correspondiente *value* para cada *key* es una lista con los datos, por orden, de *p-value*, *p-value* ajustado o FDR, $\log_2$FC, FC y su interpretación. Posteriormente, se comparan ambos diccionarios y, para aquellos identificadores que estén presentes en ambos diciconarios, se escribe en un fichero .tsv con el siguiente orden de columnas: *ID/Gene/Function/p-value (Condición 1)/X (Condición 1)/Log2FC*

(Condición 1)/FC (Condición 1)/Interpretation/p-value (Condición 2)/X (Condición 2)", siendo X *Adjusted p-value* o FDR dependiendo del programa utilizado.

Comparación doble entre dos ficheros par la obtención de aquellos genes que estén diferencialmente expresados en ambos. Para ello, se procede de manera análoga al caso anterior, pero seleccionando por *p-value* ajustado o FDR menor a 0,05 en ambos casos. El fichero generado sigue el siguiente patrón de columnas: "*ID/Gene/Function/p-value (Condición 1)/X (Condición 1)/Log2FC (Condición 1)/FC (Condición 1)/Interpretation/p-value (Condición 2)/X (Condición 2)/Log2FC (Condición 2)/FC (Condición 2)/Interpretation*" siendo X *Adjusted p-value* o FDR dependiendo del programa utilizado.

Comparación doble entre dos ficheros para la obtención de aquellos genes diferencialmente reprimidos en uno, que estén diferencialmente inducidos en el segundo fichero. De manera análoga al párrafo anterior, se hace una doble selección para incorporar aquellos genes con *p-value* ajustado o FDR menor a 0,05 y $\log_2$FC negativo para los reprimidos y *p-value* ajustado o FDR menor a 0,05 y $\log_2$FC positivo para los inducidos. Se genera un fichero con el siguiente orden de columnas: "*ID/Gene/Function/p-value (Condición 1)/X (Condición 1)/Log2FC (Condición 1)/FC (Condición 1)/Interpretation/p-value (Condición 2)/X (Condición 2)/Log2FC (Condición 2)/FC (Condición 2)/Interpretation"* siendo X *Adjusted p-value* o FDR dependiendo del programa utilizado.

3.4.5.2. BÚSQUEDA DE TÉRMINOS GO Y FUNCIONES

Partiendo de los ficheros resultantes de los scripts, se hace una búsqueda en *Saccharomyces Genome Database* (SGD) de las funciones correspondientes a cada gen. De igual manera, se obtienen los términos GO asociados a los mismos.

4. RESULTADOS Y DISCUSIÓN

4.1. CARACTERIZACIÓN DEL PATRÓN DE CRECIMIENTO DE *SSD1-V* Y *ssd1-d* MEDIANTE *BIOSCREEN*

La correcta caracterización del perfil de crecimiento de las cepas *SSD1-V* y *ssd1-d* otorga información altamente relevante para el diseño de futuros experimentos donde dicha cepa sea utilizada. Esto se realiza de acorde al apartado 3.3 en las condiciones: SD pH 6, SD pH 4 y SD pH 4 + 45 mM AcH. Los perfiles de crecimiento pueden ser consultados en la Figura 1.

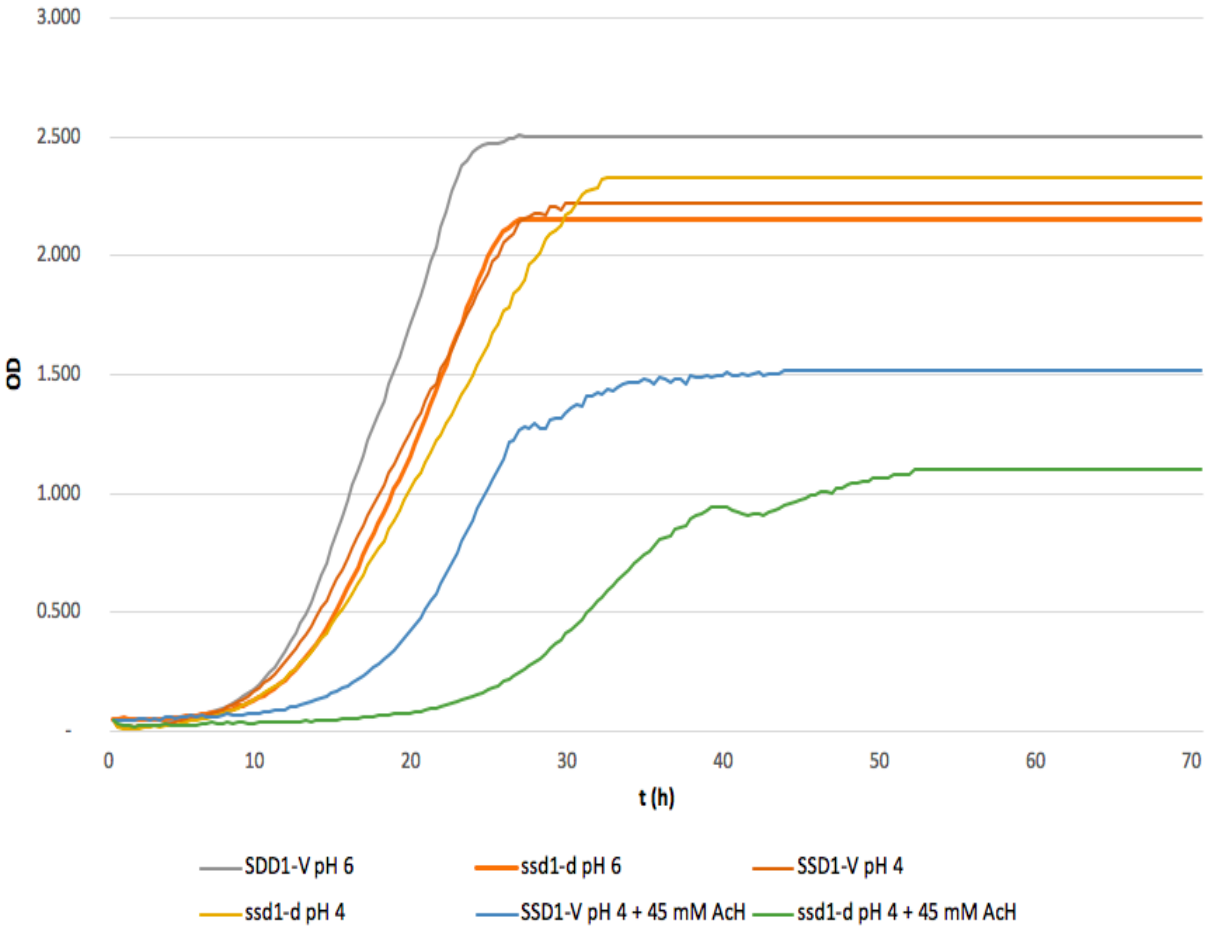

Figura 1: Perfiles de crecimiento de las cepas *SSD1-V* y *ssd1-d* en condiciones de pH 6, pH 4 y pH 4 + 45 mM AcH obtenidas en *Bioscreen*. Se muestran las medidas de absorbancia en función del tiempo.

En la Figura 1 se puede observar el perfil de crecimiento para cada una de las cepas y condiciones detalladas anteriormente. Observando dichos perfiles, se puede visualizar 3 fases: fase de latencia, fase exponencial y fase estacionaria. Para cada una de ellas, se puede calcular un parámetro: el tiempo de latencia para la fase de latencia, la tasa de crecimiento para la fase exponencial, y el rendimeinto total para la fase estacionaria. Los datos numéricos de la duración de la fase de latencia, tasa de crecimiento y rendimiento total pueden ser encontrados en la Tabla 2.

Tabla 2: Resumen de los parámetros de crecimiento obtenidos para cada cepa y condición en *Bioscreen*. Las columnas muestran los valores del tiempo de latencia, tasa de crecimiento, y rendimiento para cada cepa y condición.

Cepas	Condición	Tiempo de latencia (h)	Tasa de crecimiento	Rendimiento (OD Máxima)
	pH 6	4	1	2,5
SSD1-V	pH 4	5	0,6824	2,2
	pH 4 + 45 mM AcH	9	0,6808	1,5
	pH 6	4	0,7863	2,15
ssd1-d	pH 4	5,6	0,6216	2,3
	pH 4 + 45 mM AcH	18	0,3461	1,1

Analizando el rendimiento máximo, se observa que la cepa con mayor rendimiento es *SSD1-V* a pH 6 con una OD = 2,5. De igual manera, se ve cómo las demás cepas, independientemente del pH, presentan rendimientos similares, siendo menor el rendimiendo de *SSD1-V* pH 4 frente a *SSD1-V* pH 6. Esto refleja el comportamiento de las cepas en medio rico, donde no hay diferencias visibles en el patrón de crecimiento en el rango de pH óptimo de levadura (pH 6-4). Al ser medio mínimo, las diferencias en el crecimiento se acrecentan. Esta relación se invierte para la cepa *ssd1-d*. Por el contrario, en las cepas con tratamiento de ácido acético, su rendimiento máximo se ve disminuido, siendo más afectada la cepa *ssd1-d*.

En cuanto a la tasa de crecimiento, ésta es superior en las cepas a pH 6 que en las cepas a pH 4. Comparando entre cepas, *SSD1-V* presenta una mayor tasa de crecimiento frente a *ssd1-d* en todas las condiciones. En condiciones de estrés por ácido acético, *SSD1-V* no ve reducida su tasa de crecimiento, mientras que la de *ssd1-d* es reducida a la mitad. Normalizando la tasa de crecimiento de todas las cepas en función de la de *SSD1-V* pH 6, se obtiene la Figura 2.

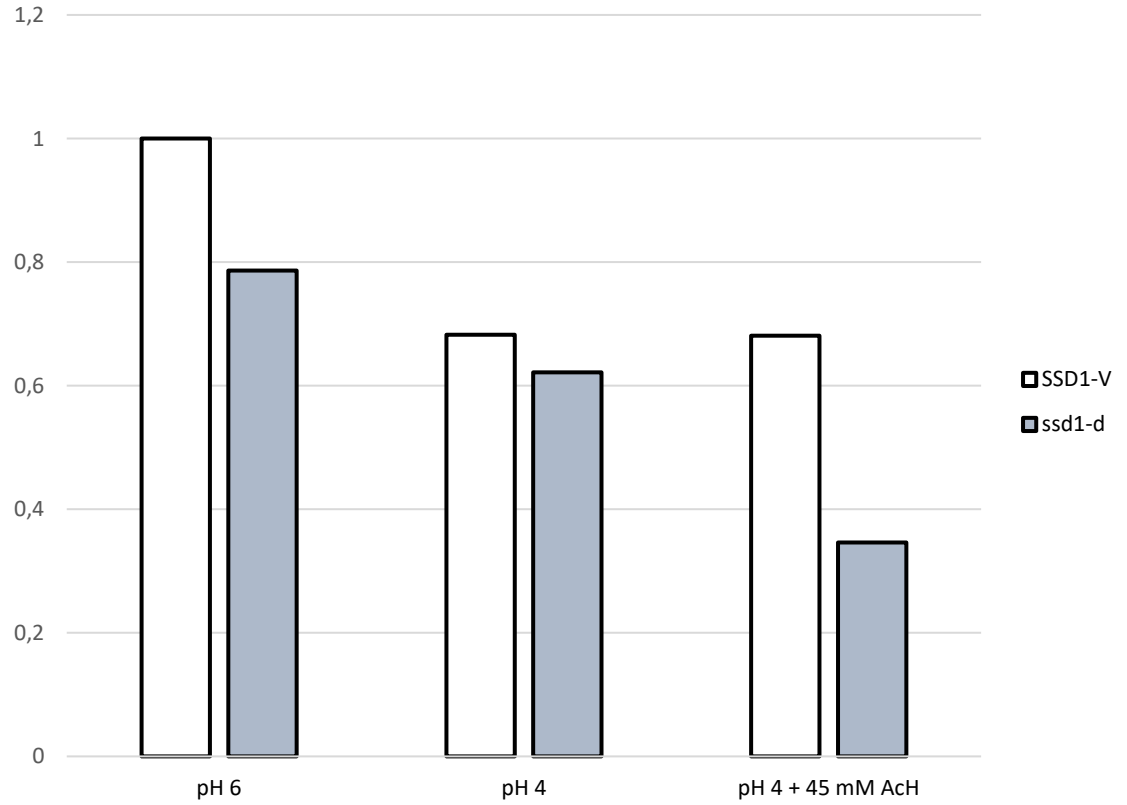

Figura 2: Valores de tasa de crecimiento relativas en *Bioscreen* para las diferentes cepas y condiciones. Dos cepas son analizadas: *SSD1-V* y *ssd1-d*. Se visualizan los valores de tasa de crecimiento en función de la cepa y condición control: *SSD1-V* pH 6. Las columnas se agrupan por condiciones: pH 6, pH 4 y pH 4 + 45 mM AcH. No se introducen barras de error debido a que los valores del mismo eran del orden del 0.01 % del valor normalizado.

Finalmente, en cuanto a la duración de la fase de latencia, se observa un ligero aumento de la misma entre pH 6 y pH 4 para ambas cepas. En condiciones de estrés por ácido acético, la duración de la fase de latencia se ve incrementada un 120 % en el caso de *SSD1-V* respecto a la observada en condiciones normales. El alelo *ssd1-d* en estas condiciones presenta la mayor duración, extendiéndose el doble en el tiempo que la cepa *SSD1-V*.

Por lo tanto, de acorde a los resultados obtenidos, se puede observar una tolerancia al ácido acético presentada por la cepa *SSD1-V* en medio mínimo en contraposición a la cepa *ssd1-d*, descrita en TFG anteriores (TFG de Manuel Bernabeu, 2015). Esto es evidenciado por la tasa de crecimiento equivalente entre las condiciones pH 4 y pH 4 + 45 mM AcH para *SSD1-V* en contraposición a *ssd1-d*, la cual se ve reducida del orden del 50 %. Asimismo, una menor duración de la fase de latencia y un mayor rendimiento de *SSD1-V* frente a *ssd1-d* en pH 4 + 45 mM AcH lo refuerzan.

4.2. EXTRACCIÓN Y PURIFICACIÓN DE ARN PARA RNAseq

Debido a que la cepa *SSD1-V* presenta tolerancia a ácido acético en medio mínimo, característica ausente en la cepa *ssd1-d*, se realizaron qPCR para comprobar los niveles de expresión de *CLN1*, *CLN2* y *CLN3*, ciclinas del ciclo celular (TFG de Celia Canales, 2017). Se seleccionaron estas ciclinas debido a que están implicadas en la transción G_1-S , junto al hecho de que la interacción entre Ssd1 y el mensajero de *CLN2* había sido descrita previamente (Ohyama *et al.*, 2010). Debido que no se obtuvieron resultados significativos, se opta por realizar un estudio transcriptómico mediante RNAseq. La finalidad es individualizar genes candidatos que, debido a la presencia y ausencia de *SSD1*, se encuentren diferencialmente expresados y sean responsables del crecimiento diferencial observado entre *SSD1-V* y *ssd1-d* en presencia de ácido acético.

Para ello, se realiza una extracción y purificación del ARN total presente en las cepas *SSD1-V* y *ssd1-d* en fase exponencial crecidas en medio SD en las condiciones de SD pH 4 y SD pH 4 + 45 mM AcH. Esto se realiza para 4 réplicas biológicas por cada cepa y condición de acorde a los apartados 3.4.1 y 3.4.2. Los valores de concentración tras la purificación pueden ser consultados en la Tabla 3.

Para poder ser aceptadas las muestras y enviadas al servicio de secuenciación, es necesario que estas posean 20 µg de ARN total en 10-20 µL de muestra. Se verifica la cantidad de ARN por *Nanodrop* y la calidad del mismo por electroforesis en gel de agarosa. Tras la comprobación de que las muestras cumplen los requisitos requeridos, son entregadas al servicio, el cual procede a analizar nuevamente la calidad de las mismas de acuerdo al apartado 3.4.3.

El informe de resultados rendido por *Bioanalyzer* se presenta en la Figura 3. Todas las muestras poseen buena calidad, al manifestarse en el electroferograma resumen generado la doble banda correspondiente al ARNr, indicativo de que el ARN total extraído no está degradado.

De igual manera, el *Bioanalyzer* rinde el RIN de cada muestra. Los resultados del valor RIN se pueden consultar en la Tabla 3. Se necesita un RIN superior a 7 para ser considerados adecuados. Dado que el promedio del valor RIN de las muestras supera el umbral, se procede a su secuenciación masiva. No obstante, las muestras 1, 7 y 14, al poseer un RIN inferior a las demás, pueden suponer una fuente de variabilidad y ha de comprobarse la reproducibilidad de las réplicas posteriormente mediante un análisis PCA.

Tabla 3: Resumen de las condiciones y réplicas del experimento, junto con los resultados de la purificación de ARN total y el RIN para cada muestra. Cuatro réplicas biológicas son utilizadas para cada condición en medio SD. En la primera columna se muestra orden de las réplicas en el experimento. En la segunda, se describe la cepa y condición de cada réplica. En la tercera se resumen los valores de la concentración del ARN total tras la purificación. Finalmente, la cuarta columna muestra los valores de calidad del ARN total en valor RIN determinados por *Bioanalyzer*.

Número de la muestra	Descripción	Concentración (µg/ml)	RIN
1	*SSD1-V* pH 4 + 0mM AcH (1)	130	6,6
2	*SSD1-V* pH 4 + 0mM AcH (2)	743	8,3
3	*SSD1-V* pH 4 + 0mM AcH (3)	284	7
4	*SSD1-V* pH 4 + 0mM AcH (4)	336	7,7
5	*ssd1-d* pH 4 + 0mM AcH (1)	417	7,7
6	*ssd1-d* pH 4 + 0mM AcH (2)	278	7,8
7	*ssd1-d* pH 4 + 0mM AcH (3)	322	6
8	*ssd1-d* pH 4 + 0mM AcH (4)	303	7,8
9	*SSD1-V* pH 4 + 45mM AcH (1)	140	9
10	*SSD1-V* pH 4 + 45mM AcH (2)	220	8,2
11	*SSD1-V* pH 4 + 45mM AcH (3)	145	8,5
12	*SSD1-V* pH 4 + 45mM AcH (4)	137	8,4
13	*ssd1-d* pH 4 + 45mM AcH (1)	90	7,7
14	*ssd1-d* pH 4 + 45mM AcH (2)	43	6,2
15	*ssd1-d* pH 4 + 45mM AcH (3)	44	7,2
16	*ssd1-d* pH 4 + 45mM AcH (4)	363	8,2

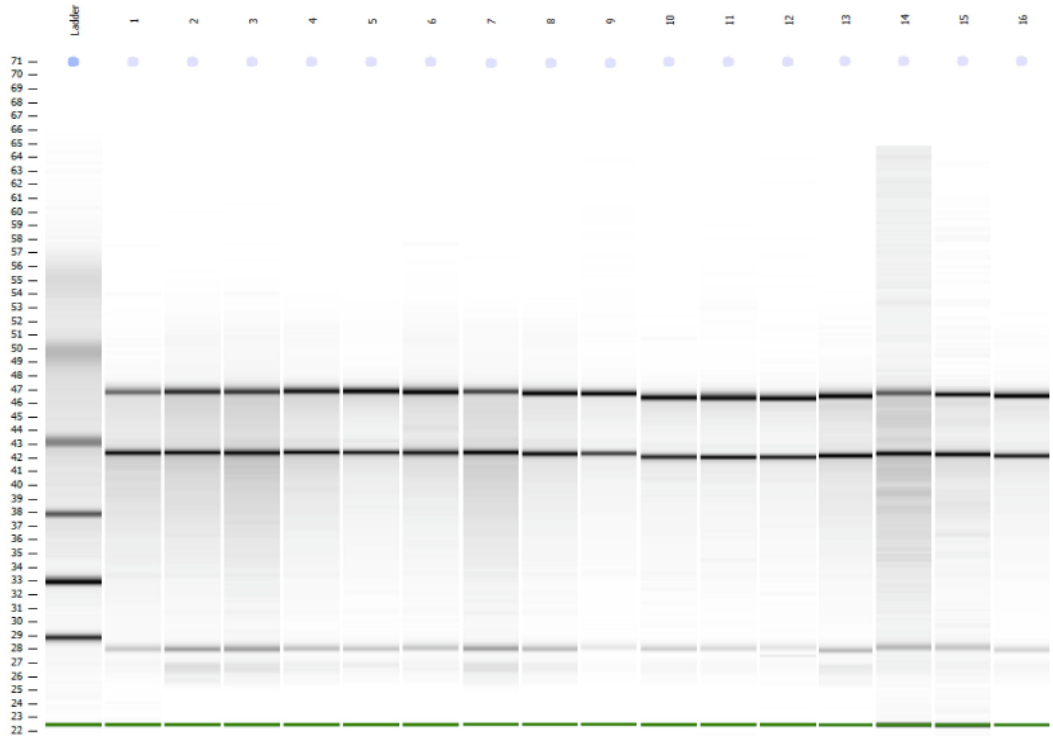

Figura 3: Resultados obtenidos del *Bioanalyzer*. Se muestran los resultados de la electroforesis en gel de agarosa. Se puede apreciar que las 16 muestras no presentan degradación, denotado por la doble banda correspondiente al ARNr.

4.3. LIMPIEZA CONTROL DE CALIDAD DE LAS SECUENCIAS

Para poder mapear las secuencias, las lecturas han de ser tratadas con el fin de eliminar los adaptadores en los extremos de las lecturas, posibles adaptadores quiméricos unidos entre ellos, y lecturas con poca calidad. Esto se realiza de acorde al apartado 3.4.4. De esta forma, se eliminan zonas de lecturas con calidad inferior a 30 en escala Phred, se eliminan los adaptadores junto con lecturas inferiores a 20 pb de longitud y aquellas lecturas ambiguas, donde el secuenciador haya determinado la base nitrogenada como "N". De esta forma, menos del 2% de las lecturas son descartadas, siendo el 98% restante sometida al control de calidad por *FASTQC*.

Dado que *FASTQC* está diseñado para experimentos ómicos de diversa índole, se espera que algunos de sus análisis rindan un resultado negativo. Un ejemplo claro de ello es el análisis de duplicaciones de secuencia, donde el programa espera que haya poca duplicación, dado que considera que son lecturas destinadas al ensamblaje de genoma. Por lo tanto, en RNAseq, siempre dará un fallo en este test.

Todas las réplicas superan satisfactoriamente el control de calidad, presentando una calidad media superior a 34 en escala Phred. Los ficheros individuales rendidos por el programa pueden encontrarse en Anexos, en el apartado 7.2.

4.4. RESULTADOS DEL MAPEO

Los datos procesados en el apartado anterior son mapeados con el programa *TopHat2* frente al genoma de referencia de la cepa S288C, disponible en *SGD* (S288C GENOME REFERENCE, 2018). De esta manera, se obtiene una media de 98% de lecturas que mapean de manera única en el genoma para cada réplica y condición del experimento, siendo indicativo de un correcto mapeo. Ello puede ser observado en el visualizador de mapeos *IGV*.

4.5. CONTAJE DE LECTURAS Y ANÁLISIS DE LA REPRODUCIBILIDAD ENTRE MUESTRAS

Utilizando el programa *HTSeq-count*, se genera una matriz de conteo con el número de lecturas mapeadas para cada gen en cada condición y en cada réplica, junto con sus valores RPKM. De acorde con lo expuesto en el apartado 1.8.3.4, es necesario realizar un análisis de reproducibildad para poder identificar posibles réplicas que no sigan el patrón de las demás para la misma cepa y condición.

Un análisis PCA donde las muestras se separen por tratamiento es un buen indicador de un diseño correcto para un experimento de RNAseq enfocado en la detección de expresión diferencial. Por lo tanto, mediante lenguaje de programación R, se realiza un análisis PCA a partir de los valores RPKM de la matriz de conteo, resultando en la Figura 4.

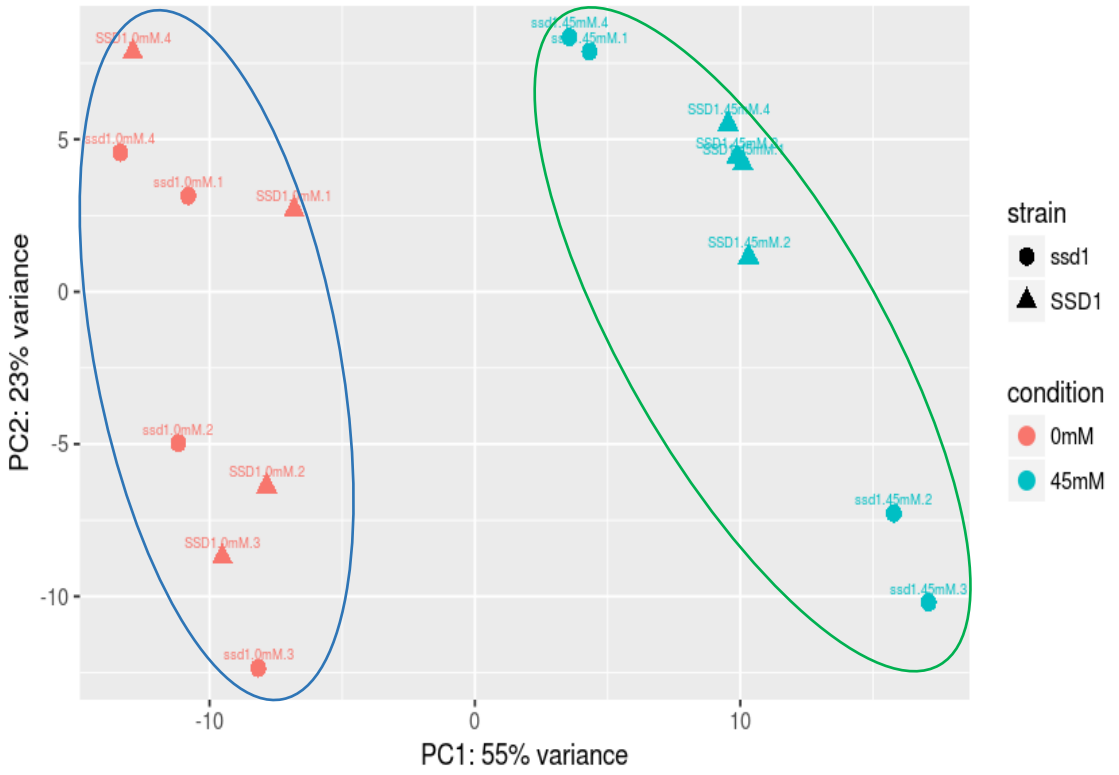

Figura 4: Análisis PCA realizado a partir de los valores RPKM de la matriz de conteo para las cepas *SSD1-V* y *ssd1-d* en ausencia (pH 4) y presencia (pH 4 + 45 mM AcH) de ácido acético. En él se muestran los puntos correspondientes al conjunto de los valores RPKM para todos los genes en cada réplica. Siendo los triángulos correspondientes a la cepa *SSD1-V* y los círculos a la cepa *ssd1-d*, se pueden observar en naranja la condición pH 4 y en azul la condición pH 4 + 45 mM AcH.

En la Figura 4 se puede visualizar la separación de las muestras por el tratamiento con ácido acético por la componente PC1, que engloba el 55 % de la varianza. Este hecho es indicativo de que el diseño del experimento para la determinación de expresión diferencial ha sido el correcto y, por lo tanto, se puede continuar con el análisis de expresión diferencial.

Además, se puede observar cómo en la componente PC2 las réplicas de la cepa *SSD1-V*, tras el tratamiento con ácido acético, tienden a agruparse presentando menor variabilidad entre ellas. Esto encaja con el hecho de que sólo una subpoblación de células, tras el tratamiento con ácido acético, es capaz de adaptarse al estrés y proliferar tras un periodo de latencia. Esta adaptación a un patrón de expresión génica concreto, responsable de la agrupación mencionada anteriormente, confirma lo descrito en publicaciones anteriores (Swinnen *et al.*, 2014).

4.6. CONTAJE DE LECTURAS Y RESULTADOS DE LA EXPRESIÓN DIFERENCIAL

A partir de la matriz de conteo obtenida con *HTSeq-count*, mediante los programas *DESeq2* y *edgeR*, se realiza el análisis de expresión diferencial. Ambos programas utilizan un umbral de *adjusted p-value* $\leq$ 0,05 (*DESeq2*) o FDR $\leq$ 0,05 (*edgeR*) para la determinación de la expresión diferencial, rindiendo ambos una lista ordenada de menor a mayor en los valores mencionados anteriormente. Ambos valores refieren al mismo concepto, *FDR-adjusted p-value*. En la actualidad de la publicación del presente trabajo, sólo el análisis de la expresión diferencial mediante *DESeq2* pudo completarse, dejando pendiente la comparación con los resultados de *edgeR*.

La comparación entre cepas de interés es: *SSD1-V* pH 4 + 45 mM AcH vs *ssd1-d* pH 4 + 45 mM AcH. En el análisis con *DESeq2*, el programa devuelve una lista de genes de los cuales 11 se consideran diferencialmente expresados. Esto es debido a que su *p-value* ajustado por FDR es menor o igual a 0,05. La lista de genes, con los valores de los parámetros para cada uno, puede ser encontrada la Tabla 4.

Tabla 4: Resultados del análisis de expresión diferencial con el programa *DESeq2* para la comparación *SSD1-V* 45 mM AcH vs *ssd1-d* 45 mM AcH. Se muestran aquellos genes cuyo p-value ajustado por FDR sea menor a 0,05, valor para el cual son considerados diferencialmente expresados. Valores negativos de Log_2FC indican menor expresión en *ssd1-d* pH 4 + 45 mM AcH frente a *SSD1-V* pH 4 + 45 mM AcH.

ID	Log_2FC	FC	p-value	p-value ajustado por FDR
YGR249W	1,273799537	2,41797535	6,13E-09	3,71E-05
YNL152W	-0,433888982	0,74026361	1,31E-08	3,97E-05
YLR142W	1,002178711	2,00302262	1,05E-06	0,00212037
YLR367W	0,505319235	1,41943741	3,90E-06	0,00588689
YDL127W	-0,887011992	0,54073289	1,54E-05	0,01327235
YGR006W	-0,52196188	0,69642414	1,37E-05	0,01327235
YOR338W	1,439532029	2,71232871	1,11E-05	0,01327235
YGL262W	1,854442428	3,61611968	2,14E-05	0,01614963
YKR075C	0,778378789	1,71520235	3,58E-05	0,02404534
YGR035C	0,903018987	1,86997501	4,67E-05	0,02705548
snR86	0,610208566	1,52647987	4,93E-05	0,02705548

De esta lista destaca el gen *PCL2* (YDL127W). A partir de la matriz de conteo, se realiza el promedio de los valores RPKM del gen *PCL2* para cada una de las réplicas en las siguientes condiciones: condiciones normales (pH 4) y en estrés por ácido acético (45 mM AcH). Esto se realiza con el fin de visualizar las variaciones brutas de los transcritos.

Conforme lo explicado en el apartado 1.8.3.5.2.a, el valor de RPKM es una normalización de las lecturas mapeadas para el gen. Por lo tanto, puede correlacionarse con la expresión del gen y ser usado para la visualización de los cambios en dicha expresión en las diferentes condiciones. Se escoge *SSD1-V* pH 4 como condición de expresión normal y se normalizan los demás valores en función de los obtenidos para esa condición, representándose en la Figura 5.

En la Figura 5 se puede observar que, sin estrés por ácido acético, el valor de RPKM para *ssd1-d* es un 23% inferior a *SSD1-V*. Este patrón se repite en condiciones de estrés por ácido acético, pero siendo dicha diferencia el doble en valor de RPKM entre *SSD1-V* y *ssd1-d*. De igual manera, se puede observar un incremento en el valor RPKM en ambas cepas en condiciones de estrés por ácido acético frente a condiciones normales. Esto es indicativo de un incremento en la expresión del gen *PCL2* frente a este estrés.

PCL2 codifica para una ciclina que interacciona con quinasas dependientes de ciclinas (CDKs), concretamente Pho85, para formar un complejo ciclina-CDK con actividad G_1 periódica (Measday *et al.*, 1994, p.2). Otras rutas conocidas de función análoga son la interacción de las ciclinas *Cln1, Cln2* y *Cln3* con *Cdc28*, otra CDK.

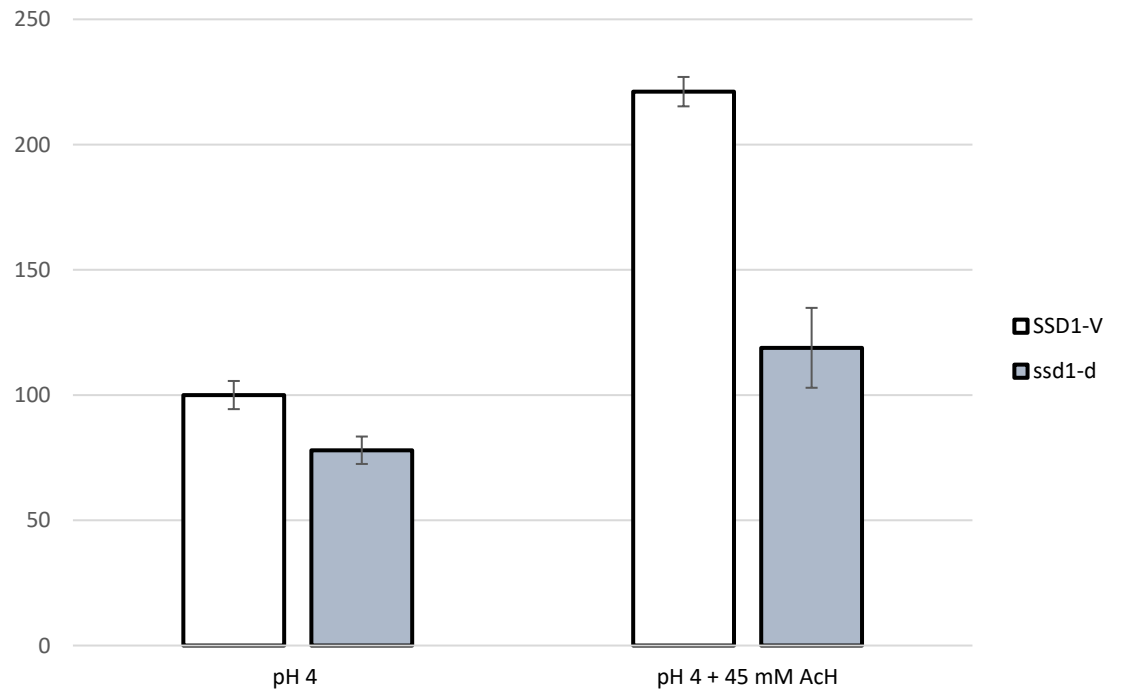

Figura 5: Valor relativo de RPKM de las lecturas para el gen PCL2 para cada cepa y condición en función de *SSD1-V* pH 4. Se observa el valor relativo de RPKM en función de la condición control: *SSD1-V* pH 4. Los valores están agrupados por condiciones: pH 4 y pH 45 mM AcH.

Dichas ciclinas fueron objeto de estudio en trabajos anteriones (TFG de Celia Canales, 2017), donde no fue posible relacionar directamente la falta de *SSD1* con el crecimiento deferencial observado entre las cepas con alelo *SSD1-V* y *ssd1-d*, debido a que el nivel de expresión de las mismas se mantenía invariable frente al tratamiento con ácido acético. En ambos casos, el complejo formado ayuda a la entrada en el ciclo mitótico de la célula, también llamado *Start*. Mientras que la proteína Pho85 no es esencial para la viabilidad celular, el complejo formado con Pcl2 sí lo es, en ausencia del complejo análogo formado con Cdc28 (Measday *et al.*, 1997). Asímismo, el complejo Pcl2-Pho85 es esencial para la morfogénesis celular (Lenburg y O'Shea, 2001).

Así pues, se selecciona el gen *PCL2* como candidato para futuros experimentos en el laboratorio, tales como la sobreexpresión de su actividad y su medición del nivel de transcripción mediante qPCR, entre otros.

Otras comparaciones, como las dobles mencionadas en el apartado 3.4.5.1.2, no pueden ser realizadas debido a que la comparación *SSD1-V* pH 4 vs *SSD1-V* pH 4 + 45 mM AcH no rinde ningún gen diferencialmente expresado. Ello puede ser debido a una falta de refinamiento en los parámetros de análisis, o a que no hay ningún cambio a nivel de expresión entre esas dos condiciones. Esto será objeto de estudio en análisis posteriores.

5. CONCLUSIONES

A partir de lo expuesto anteriormente se puede concluir:

- Se ha caracterizado el patrón de crecimiento de las cepas *SSD1-V* y *ssd1-d* en medio mínimo a pH 6, pH 4 y pH 4 + 45 mM AcH en *Bioscreen*. De esta manera, se ha podido observar la mayor tolerancia a ácido acético presente en la cepa *SSD1-V* respecto a la cepa *ssd1-d*.

- El análisis de reproducibilidad mediante el análisis PCA refleja un correcto diseño del experimento para la detección de la expresión diferencial. Ello acrecenta la confianza puesta en los resultados finales del análisis de expresión diferencial.

- A la luz de los resultados rendidos por *DESeq2*, se selecciona como gen candidato responsable del fenotipo diferencial entre *SSD1-V* y *ssd1-d* en condiciones de estrés por ácido acético al gen *PCL2*, a la espera de ser complementado con los resultados de la expresión diferencial obtenidos con *edgeR*. La interacción entre el gen *PCL2* y el gen *SSD1* ha de ser corroborada en experimentos posteriores.

6. BIBLIOGRAFÍA

ANDERS, S. AND HUBER, W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, 11, R106.

ANDERS, S., PYL, P.T. AND HUBER, W. (2015) HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31, 166–169.

ANDREWS, S. (2014) FastQC A Quality Control tool for High Throughput Sequence Data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

ASHBURNER, M., BALL, C.A., BLAKE, J.A., *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 25, 25–29.

AUER, P.L. AND DOERGE, R.W. (2010) Statistical Design and Analysis of RNA Sequencing Data. *Genetics*, 185, 405–416.

BAINBRIDGE, M.N., WARREN, R.L., HIRST, M., *et al*. (2006) Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics*, 7, 246.

BARATHOVA, M., TAKACOVA, M., HOLOTNAKOVA, T., *et al*. (2008) Alternative splicing variant of the hypoxia marker carbonic anhydrase IX expressed independently of hypoxia and tumour phenotype. *Br. J. Cancer*, 98, 129–136.

BARKER, J., KHAN, M. A. A. AND SOLOMOS, T. (1964) Mechanism of the Pasteur Effect. *Nature*, 201, 1126–1127.

BENJAMINI, Y. AND HOCHBERG, Y. (1995) Controlling The False Discovery Rate - A Practical And Powerful Approach To Multiple Testing. *J. Royal Statist. Soc., Series B*, 57, 289–300.

BERNABEU LORENZO, M. (2015) Aumento de dosis génica de los genes DPL1, SSD1 y SRP101 en Saccharomyces cerevisiae y fenotipo de tolerancia a acidificación intracelular. Available at: https://riunet.upv.es/handle/10251/54365 [Accessed July 4, 2018].

BIDLINGMAIER, S., WEISS, E.L., SEIDEL, C., DRUBIN, D.G. AND SNYDER, M. (2001) The Cbk1p pathway is important for polarized cell growth and cell separation in Saccharomyces cerevisiae. *Mol. Cell. Biol.*, 21, 2449–2462.

BLAND, J.M. AND ALTMAN, D.G. (1995) Multiple significance tests: the Bonferroni method. *BMJ*, 310, 170.

BOLGER, A.M., LOHSE, M. AND USADEL, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30, 2114–2120.

BOTSTEIN, D. AND FINK, G.R. (2011) Yeast: an experimental organism for 21st Century biology. *Genetics*, 189, 695–704.

BRACHMANN, C.B., DAVIES, A., COST, G.J., CAPUTO, E., LI, J., HIETER, P. AND BOEKE, J.D. (1998) Designer deletion strains derived from Saccharomyces cerevisiae S288C: a useful set of strains and plasmids for PCR-mediated gene disruption and other applications. *Yeast*, 14, 115–132.

BRADLEY, A.J., LIM, Y.Y. AND SINGH, F.M. (2011) Imaging features, follow-up, and management of incidentally detected renal lesions. *Clin Radiol*, 66, 1129–1139.

BRENNER, S., JOHNSON, M., BRIDGHAM, J., *et al*. (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.*, 18, 630–634.

BUERMANS, H.P.J. AND DEN DUNNEN, J.T. (2014) Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1842, 1932–1941.

BULLARD, J.H., PURDOM, E., HANSEN, K.D. AND DUDOIT, S. (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, 11, 94.

BYRNE, A., BEAUDIN, A.E., OLSEN, H.E., *et al*. (2017) Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nature Communications*, 8, 16027.

CANALES QUILIS, C. (2017) Mecanismos de homeostasis del pH intracelular en levadura: activación de la bomba de protones Pma1 y estabilidad de RNA mensajeros de ciclinas G1. Available at: https://riunet.upv.es/handle/10251/86398 [Accessed July 4, 2018].

CONESA, A., MADRIGAL, P., TARAZONA, S., *et al*. (2016) A survey of best practices for RNA-seq data analysis. *Genome Biol.*, 17, 13.

COSTANZO, M., BARYSHNIKOVA, A., BELLAY, J., *et al*. (2010) The genetic landscape of a cell. *Science*, 327, 425–431.

COSTA-SILVA, J., DOMINGUES, D. AND LOPES, F.M. (2017) RNA-Seq differential expression analysis: An extended review and a software tool. *PLOS ONE*, 12, e0190152.

DAI, M., THOMPSON, R.C., MAHER, C., CONTRERAS-GALINDO, R., KAPLAN, M.H., MARKOVITZ, D.M., OMENN, G. AND MENG, F. (2010) NGSQC: cross-platform quality analysis pipeline for deep sequencing data. *BMC Genomics*, 11 Suppl 4, S7.

DAMAGHI, M., WOJTKOWIAK, J.W. AND GILLIES, R.J. (2013) pH sensing and regulation in cancer. *Front Physiol*, 4. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3865727/ [Accessed June 17, 2018].

DANG, W., SUTPHIN, G.L., DORSEY, J.A., *et al*. (2014) Inactivation of yeast Isw2 chromatin remodeling enzyme mimics longevity effect of calorie restriction via induction of genotoxic stress response. *Cell Metab.*, 19, 952–966.

DE VIRGILIO, C. AND LOEWITH, R. (2006) The TOR signalling network from yeast to man. *Int. J. Biochem. Cell Biol.*, 38, 1476–1481.

DEMBÉLÉ, D. AND KASTNER, P. (2014) Fold change rank ordering statistics: a new method for detecting differentially expressed genes. *BMC Bioinformatics*, 15, 14.

DILLIES, M.-A., RAU, A., AUBERT, J., *et al*. (2013) A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinformatics*, 14, 671–683.

DING, L., RATH, E. AND BAI, Y. (2017) Comparison of Alternative Splicing Junction Detection Tools Using RNA-Seq Data. *Curr Genomics*, 18, 268–277.

DOBIN, A., DAVIS, C.A., SCHLESINGER, F., DRENKOW, J., ZALESKI, C., JHA, S., BATUT, P., CHAISSON, M. AND GINGERAS, T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29, 15–21.

DOHM, J.C., LOTTAZ, C., BORODINA, T. AND HIMMELBAUER, H. (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.*, 36, e105.

DOREY, F. (2010) In Brief: The P Value: What Is It and What Does It Tell You? *Clin Orthop Relat Res*, 468, 2297–2298.

DU, L.-L. AND NOVICK, P. (2002) Pag1p, a novel protein associated with protein kinase Cbk1p, is required for cell morphogenesis and proliferation in Saccharomyces cerevisiae. *Mol. Biol. Cell*, 13, 503–514.

DUNNING, M.J., BARBOSA-MORAIS, N.L., LYNCH, A.G., TAVARÉ, S. AND RITCHIE, M.E. (2008) Statistical issues in the analysis of Illumina data. *BMC Bioinformatics*, 9, 85.

EFRON, B., TIBSHIRANI, R., J.D, S. AND TUSHER, V. (2001) Empirical Bayes Analysis of a Microarray Experiment. *Journal of the American Statistical Association*, 96, 1151–1160.

ENSEMBL GENOME BROWSER. Available at: https://www.ensembl.org/index.html [Accessed June 18, 2018]

FASTX-TOOLKIT. Available at: http://hannonlab.cshl.edu/fastx_toolkit/ [Accessed June 12, 2018]

FAULKNER, G.J., FORREST, A.R.R., CHALK, A.M., SCHRODER, K., HAYASHIZAKI, Y., CARNINCI, P., HUME, D.A. AND GRIMMOND, S.M. (2008) A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE. *Genomics*, 91, 281–288.

FIUME, M., SMITH, E.J.M., BROOK, A., STRBENAC, D., TURNER, B., MEZLINI, A.M., ROBINSON, M.D., WODAK, S.J. AND BRUDNO, M. (2012) Savant Genome Browser 2: visualization and analysis for population-scale genomics. *Nucleic Acids Res.*, 40, W615-621.

FIUME, M., WILLIAMS, V., BROOK, A. AND BRUDNO, M. (2010) Savant: genome browser for high-throughput sequencing data. *Bioinformatics*, 26, 1938–1944.

FRIEDLÄNDER, M.R., CHEN, W., ADAMIDI, C., MAASKOLA, J., EINSPANIER, R., KNESPEL, S. AND RAJEWSKY, N. (2008) Discovering microRNAs from deep sequencing data using miRDeep. *Nat. Biotechnol.*, 26, 407–415.

GALLAGHER, F.A., KETTUNEN, M.I., DAY, S.E., *et al*. (2008) Magnetic resonance imaging of pH in vivo using hyperpolarized 13C-labelled bicarbonate. *Nature*, 453, 940–943.

GARBER, M., GRABHERR, M.G., GUTTMAN, M. AND TRAPNELL, C. (2011) Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Methods*, 8, 469–477.

GATENBY, R.A. AND GILLIES, R.J. (2004) Why do cancers have high aerobic glycolysis? *Nat. Rev. Cancer*, 4, 891–899.

GHAEMMAGHAMI, S., HUH, W.-K., BOWER, K., HOWSON, R.W., BELLE, A., DEPHOURE, N., O'SHEA, E.K. AND WEISSMAN, J.S. (2003) Global analysis of protein expression in yeast. *Nature*, 425, 737–741.

GIAEVER, G., CHU, A.M., NI, L., *et al*. (2002) Functional profiling of the Saccharomyces cerevisiae genome. *Nature*, 418, 387–391.

GILLIES, R.J. (2002) In vivo molecular imaging. *J. Cell. Biochem. Suppl.*, 39, 231–238.

GOFFEAU, A., BARRELL, B.G., BUSSEY, H., *et al*. (1996) Life with 6000 genes. *Science*, 274, 546, 563–567.

GOTTLIEB, R.A., NORDBERG, J., SKOWRONSKI, E. AND BABIOR, B.M. (1996) Apoptosis induced in Jurkat cells by several agents is preceded by intracellular acidification. *Proc. Natl. Acad. Sci. U.S.A.*, 93, 654–658.

GRABHERR, M.G., HAAS, B.J., YASSOUR, M., *et al*. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome., Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat Biotechnol*, 29, 29, 644, 644–652.

HANAHAN, D. AND WEINBERG, R.A. (2011) Hallmarks of Cancer: The Next Generation. *Cell*, 144, 646–674.

HANAHAN, D. AND WEINBERG, R.A. (2000) The hallmarks of cancer. *Cell*, 100, 57–70.

HANSEN, K.D., BRENNER, S.E. AND DUDOIT, S. (2010) Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res.*, 38, e131.

HÄNZELMANN, S., CASTELO, R. AND GUINNEY, J. (2013) GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*, 14, 7.

HARTLEY, S.W. AND MULLIKIN, J.C. (2016) Detection and visualization of differential splicing in RNA-Seq data with JunctionSeq. *Nucleic Acids Res*, 44, e127.

HASHIM, A.I., ZHANG, X., WOJTKOWIAK, J.W., MARTINEZ, G.V. AND GILLIES, R.J. (2011) Imaging pH and metastasis. *NMR Biomed*, 24, 582–591.

HE, K., LI, M., FU, Y., GONG, F. AND SUN, X. (2018) A direct approach to false discovery rates by decoy permutations. Available at: http://arxiv.org/abs/1804.08222 [Accessed June 13, 2018].

HELLER, M.J. (2002) DNA microarray technology: devices, systems, and applications. *Annu Rev Biomed Eng*, 4, 129–153.

HILKER, R., STADERMANN, K.B., DOPPMEIER, D., KALINOWSKI, J., STOYE, J., STRAUBE, J., WINNEBALD, J. AND GOESMANN, A. (2014) ReadXplorer--visualization and analysis of mapped sequences. *Bioinformatics*, 30, 2247–2254.

HOOSE, S.A., RAWLINGS, J.A., KELLY, M.M., *et al*. (2012) A Systematic Analysis of Cell Cycle Regulators in Yeast Reveals That Most Factors Act Independently of Cell Size to Control Initiation of Division. *PLOS Genetics*, 8, e1002590.

HUANG, D.W., SHERMAN, B.T. AND LEMPICKI, R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, 4, 44–57.

HUBER, W., CAREY, V.J., GENTLEMAN, R., *et al*. (2015) Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods*, 12, 115–121.

HUH, W.-K., FALVO, J.V., GERKE, L.C., CARROLL, A.S., HOWSON, R.W., WEISSMAN, J.S. AND O'SHEA, E.K. (2003) Global analysis of protein localization in budding yeast. *Nature*, 425, 686–691.

S288C GENOME REFERENCE. Available at: https://downloads.yeastgenome.org/sequence/S288C_reference/ [Accessed July 4, 2018]

ITO, T., CHIBA, T., OZAWA, R., YOSHIDA, M., HATTORI, M. AND SAKAKI, Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. U.S.A.*, 98, 4569–4574.

IYER, V.R., HORAK, C.E., SCAFE, C.S., BOTSTEIN, D., SNYDER, M. AND BROWN, P.O. (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, 409, 533–538.

JABBARI, K. AND BERNARDI, G. (1998) CpG doublets, CpG islands and Alu repeats in long human DNA sequences from different isochore families. *Gene*, 224, 123–127.

JAITIN, D.A., KENIGSBERG, E., KEREN-SHAUL, H., *et al*. (2014) Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science*, 343, 776–779.

JANSEN, J.M., WANLESS, A.G., SEIDEL, C.W. AND WEISS, E.L. (2009) Cbk1 regulation of the RNA-binding protein Ssd1 integrates cell fate with translational control. *Curr. Biol.*, 19, 2114–2120.

JAZAYERI, S.M., MUÑOZ, M., MARINA, L. AND ROMERO, H.M. (2015) RNA-SEQ: A GLANCE AT TECHNOLOGIES AND METHODOLOGIES. *Acta Biológica Colombiana*, 20, 23–35.

JEAN, G., KAHLES, A., SREEDHARAN, V.T., DE BONA, F. AND RÄTSCH, G. (2010) RNA-Seq read alignments with PALMapper. *Curr Protoc Bioinformatics*, Chapter 11, Unit 11.6.

JEWETT, M.C., HOFMANN, G. AND NIELSEN, J. (2006) Fungal metabolite analysis in genomics and phenomics. *Curr. Opin. Biotechnol.*, 17, 191–197.

JONES, G.M., STALKER, J., HUMPHRAY, S., WEST, A., COX, T., ROGERS, J., DUNHAM, I. AND PRELICH, G. (2008) A systematic library for comprehensive overexpression screens in Saccharomyces cerevisiae. *Nat. Methods*, 5, 239–241.

JORGENSEN, P., NELSON, B., ROBINSON, M.D., CHEN, Y., ANDREWS, B., TYERS, M. AND BOONE, C. (2002) High-resolution genetic mapping with ordered arrays of Saccharomyces cerevisiae deletion mutants. *Genetics*, 162, 1091–1099.

KATZ, Y., WANG, E.T., AIROLDI, E.M. AND BURGE, C.B. (2010) Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods*, 7, 1009–1015.

KENT, W.J., SUGNET, C.W., FUREY, T.S., ROSKIN, K.M., PRINGLE, T.H., ZAHLER, A.M. AND HAUSSLER, D. (2002) The human genome browser at UCSC. *Genome Res.*, 12, 996–1006.

KIM, D., LANGMEAD, B. AND SALZBERG, S.L. (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*, 12, 357–360.

KIM, D., PERTEA, G., TRAPNELL, C., PIMENTEL, H., KELLEY, R. AND SALZBERG, S.L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, 14, R36.

KONG, A., GUDBJARTSSON, D.F., SAINZ, J., *et al*. (2002) A high-resolution recombination map of the human genome. *Nat. Genet.*, 31, 241–247.

KROGAN, N.J., CAGNEY, G., YU, H., *et al*. (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, 440, 637–643.

KUMAR, S., VO, A.D., QIN, F. AND LI, H. (2016) Comparative assessment of methods for the fusion transcripts detection from RNA-Seq data. *Scientific Reports*, 6, 21597.

KURISCHKO, C., KIM, H.K., KURAVI, V.K., PRATZKA, J. AND LUCA, F.C. (2011) The yeast Cbk1 kinase regulates mRNA localization via the mRNA-binding protein Ssd1. *J. Cell Biol.*, 192, 583–598.

KURISCHKO, C., KURAVI, V.K., HERBERT, C.J. AND LUCA, F.C. (2011) Nucleocytoplasmic shuttling of Ssd1 defines the destiny of its bound mRNAs. *Mol. Microbiol.*, 81, 831–849.

ŁABAJ, P.P., LEPARC, G.G., LINGGI, B.E., MARKILLIE, L.M., WILEY, H.S. AND KREIL, D.P. (2011) Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling. *Bioinformatics*, 27, i383-391.

LAGADIC-GOSSMANN, D., HUC, L. AND LECUREUR, V. (2004) Alterations of intracellular pH homeostasis in apoptosis: origins and roles. *Cell Death Differ.*, 11, 953–961.

LANDER, E.S., LINTON, L.M., BIRREN, B., *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, 409, 860–921.

LASHKARI, D.A., DERISI, J.L., MCCUSKER, J.H., NAMATH, A.F., GENTILE, C., HWANG, S.Y., BROWN, P.O. AND DAVIS, R.W. (1997) Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc. Natl. Acad. Sci. U.S.A.*, 94, 13057–13062.

LENBURG, M.E. AND O'SHEA, E.K. (2001) Genetic evidence for a morphogenetic function of the Saccharomyces cerevisiae Pho85 cyclin-dependent kinase. *Genetics*, 157, 39–51.

LEVIN, J.Z., YASSOUR, M., ADICONIS, X., NUSBAUM, C., THOMPSON, D.A., FRIEDMAN, N., GNIRKE, A. AND REGEV, A. (2010) Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat. Methods*, 7, 709–715.

LI, J., LIU, J., WANG, X., ZHAO, L., CHEN, Q. AND ZHAO, W. (2009) A waterbath method for preparation of RNA from Saccharomyces cerevisiae. *Anal. Biochem.*, 384, 189–190.

LI, P., PIAO, Y., SHON, H.S. AND RYU, K.H. (2015) Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data. *BMC Bioinformatics*, 16, 347.

LI, B., RUOTTI, V., STEWART, R.M., THOMSON, J.A. AND DEWEY, C.N. (2010) RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, 26, 493–500.

LIAO, Y., SMYTH, G.K. AND SHI, W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30, 923–930.

LIEB, J.D., LIU, X., BOTSTEIN, D. AND BROWN, P.O. (2001) Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nat. Genet.*, 28, 327–334.

LIU, Y., FERGUSON, J.F., XUE, C., SILVERMAN, I.M., GREGORY, B., REILLY, M.P. AND LI, M. (2013) Evaluating the impact of sequencing depth on transcriptome profiling in human adipose. *PLoS ONE*, 8, e66883.

LIU, Y., ZHOU, J. AND WHITE, K.P. (2014) RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics*, 30, 301–304.

LOVE, M.I., HUBER, W. AND ANDERS, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, 15, 550.

LUDOVICO, P., SOUSA, M.J., SILVA, M.T., LEÃO, C. AND CÔRTE-REAL, M. (2001) Saccharomyces cerevisiae commits to a programmed cell death process in response to acetic acid. *Microbiology (Reading, Engl.)*, 147, 2409–2415.

MAHMOUD, S., PLANES, M.D., CABEDO, M., TRUJILLO, C., RIENZO, A., CABALLERO-MOLADA, M., SHARMA, S.C., MONTESINOS, C., MULET, J.M., SERRANO, R. (2017) TOR complex 1 regulates the yeast plasma membrane proton pump and pH and potassium homeostasis. *FEBS Lett.*, 591, 1993–2002.

MARCO-SOLA, S., SAMMETH, M., GUIGÓ, R. AND RIBECA, P. (2012) The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat. Methods*, 9, 1185–1188.

MARIONI, J.C., MASON, C.E., MANE, S.M., STEPHENS, M. AND GILAD, Y. (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, 18, 1509–1517.

MARTIN, M. (2011) CUTADAPT removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17.

MAZA, E., FRASSE, P., SENIN, P., BOUZAYEN, M. AND ZOUINE, M. (2013) Comparison of normalization methods for differential gene expression analysis in RNA-Seq experiments: A matter of relative size of studied transcriptomes. *Commun Integr Biol*, 6, e25849.

MCINERNY, C.J. (2016) Cell cycle regulated transcription: from yeast to cancer. *F1000Res*, 5. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4870989/ [Accessed June 16, 2018].

MEASDAY, V., MOORE, L., OGAS, J., TYERS, M. AND ANDREWS, B. (1994) The PCL2 (ORFD)-PHO85 cyclin-dependent kinase complex: a cell cycle regulator in yeast. *Science*, 266, 1391–1395.

MEASDAY, V., MOORE, L., RETNAKARAN, R., LEE, J., DONOVIEL, M., NEIMAN, A.M. AND ANDREWS, B. (1997) A family of cyclin-like proteins that interact with the Pho85 cyclin-dependent kinase. *Mol. Cell. Biol.*, 17, 1212–1223.

MEDINA, I., CARBONELL, J., PULIDO, L., *et al*. (2010) Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. *Nucleic Acids Res.*, 38, W210-213.

MEDINA, I., SALAVERT, F., SANCHEZ, R., MARIA, A. DE, ALONSO, R., ESCOBAR, P., BLEDA, M. AND DOPAZO, J. (2013) Genome Maps, a new generation genome browser. *Nucleic Acids Res.*, 41, W41-46.

MEYERS, B.C., VU, T.H., TEJ, S.S., GHAZAL, H., MATVIENKO, M., AGRAWAL, V., NING, J. AND HAUDENSCHILD, C.D. (2004) Analysis of the transcriptional complexity of Arabidopsis thaliana by massively parallel signature sequencing. *Nat. Biotechnol.*, 22, 1006–1011.

MORIN, R., BAINBRIDGE, M., FEJES, A., *et al*. (2008) Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *BioTechniques*, 45, 81–94.

MORTAZAVI, A., WILLIAMS, B.A., MCCUE, K., SCHAEFFER, L. AND WOLD, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, 5, 621–628.

MUNOZ, A.J., WANICHTHANARAK, K., Meza, E. and Petranovic, D. (2012) Systems biology of yeast cell death. *FEMS Yeast Res.*, 12, 249–265.

NOBLE, W.S. (2009) How does multiple testing correction work? *Nat Biotechnol*, 27, 1135–1137.

OHYAMA, Y., KASAHARA, K. AND KOKUBO, T. (2010) Saccharomyces cerevisiae Ssd1p promotes CLN2 expression by binding to the 5'-untranslated region of CLN2 mRNA. *Genes Cells*, 15, 1169–1188.

OSHLACK, A. AND WAKEFIELD, M.J. (2009) Transcript length bias in RNA-seq data confounds systems biology. *Biol. Direct*, 4, 14.

PALMQVIST, E. AND HAHN-HÄGERDAL, B. (2000) Fermentation of lignocellulosic hydrolysates. II: inhibitors and mechanisms of inhibition. *Bioresource Technology*, 74, 25–33.

PAMPULHA, M.E. AND LOUREIRO, V. (1989) Interaction of the effects of acetic acid and ethanol on inhibition of fermentation in *Saccharomyces cerevisiae*. *Biotechnol Lett*, 11, 269–274.

PARKHOMCHUK, D., BORODINA, T., AMSTISLAVSKIY, V., BANARU, M., HALLEN, L., KROBITSCH, S., LEHRACH, H. AND SOLDATOV, A. (2009) Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res.*, 37, e123.

PATRO, R., MOUNT, S.M. AND KINGSFORD, C. (2014) Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat. Biotechnol.*, 32, 462–464.

PETRANOVIC, D., TYO, K., VEMURI, G.N. AND NIELSEN, J. (2010) Prospects of yeast systems biology for human health: integrating lipid, protein and energy metabolism. *FEMS Yeast Res.*, 10, 1046–1059.

POLLEN, A.A., NOWAKOWSKI, T.J., SHUGA, J., *et al*. (2014) Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.*, 32, 1053–1058.

RACKI, W.J., BÉCAM, A.M., NASR, F. AND HERBERT, C.J. (2000) Cbk1p, a protein similar to the human myotonic dystrophy kinase, is essential for normal morphogenesis in Saccharomyces cerevisiae. *EMBO J.*, 19, 4524–4532.

RAPAPORT, F., KHANIN, R., LIANG, Y., PIRUN, M., KREK, A., ZUMBO, P., MASON, C.E., SOCCI, N.D. AND BETEL, D. (2013) Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol.*, 14, R95.

REDDY, R. (2015) A Comparison of Methods: Normalizing High-Throughput RNA Sequencing Data. *bioRxiv*, 026062.

REFSEQ: NCBI REFERENCE SEQUENCE DATABASE. Available at: https://www.ncbi.nlm.nih.gov/refseq/ [Accessed June 13, 2018]

REUTER, J.A., SPACEK, D. AND SNYDER, M.P. (2015) High-Throughput Sequencing Technologies. *Mol Cell*, 58, 586–597.

RICHARDSON, R., DENIS, C.L., ZHANG, C., *et al*. (2012) Mass spectrometric identification of proteins that interact through specific domains of the poly(A) binding protein. *Mol. Genet. Genomics*, 287, 711–730.

ROBINSON, M.D., MCCARTHY, D.J. AND SMYTH, G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26, 139–140.

ROBINSON, M.D. AND OSHLACK, A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.*, 11, R25.

ROGÉ, X. AND ZHANG, X. (2014) RNAseqViewer: visualization tool for RNA-Seq data. *Bioinformatics*, 30, 891–892.

SACCHAROMYCES GENOME DATABASE | SGD. Available at: https://www.yeastgenome.org/ [Accessed June 13, 2018]

SCHENA, M., SHALON, D., DAVIS, R.W. AND BROWN, P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270, 467–470.

SCHROEDER, A., MUELLER, O., STOCKER, S., *et al*. (2006) The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol. Biol.*, 7, 3.

SEQC/MAQC-III CONSORTIUM (2014) A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.*, 32, 903–914.

SHARMA, M., ASTEKAR, M., SOI, S., MANJUNATHA, B.S., SHETTY, D.C. AND RADHAKRISHNAN, R. (2015) pH Gradient Reversal: An Emerging Hallmark of Cancers. *Recent Pat Anticancer Drug Discov*, 10, 244–258.

SIMS, D., SUDBERY, I., ILOTT, N.E., HEGER, A. AND PONTING, C.P. (2014) Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.*, 15, 121–132.

STOREY, J.D. (2002) A Direct Approach to False Discovery Rates. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64, 479–498.

SUTTON, A., IMMANUEL, D. AND ARNDT, K.T. (1991) The SIT4 protein phosphatase functions in late G1 for progression into S phase. *Mol. Cell. Biol.*, 11, 2133–2148.

SWINNEN, S., FERNÁNDEZ-NIÑO, M., GONZÁLEZ-RAMOS, D., MARIS, A.J.A. VAN AND NEVOIGT, E. (2014) The fraction of cells that resume growth after acetic acid addition is a strain-dependent parameter of acetic acid tolerance in Saccharomyces cerevisiae. *FEMS Yeast Res.*, 14, 642–653.

TARASSOV, K., MESSIER, V., LANDRY, C.R., *et al*. (2008) An in vivo map of the yeast protein interactome. *Science*, 320, 1465–1470.

TARAZONA, S., GARCÍA-ALCALDE, F., DOPAZO, J., FERRER, A. AND CONESA, A. (2011) Differential expression in RNA-seq: a matter of depth. *Genome Res.*, 21, 2213–2223.

THORVALDSDÓTTIR, H., ROBINSON, J.T. AND MESIROV, J.P. (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinformatics*, 14, 178–192.

TONG, A.H., EVANGELISTA, M., PARSONS, A.B., *et al*. (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, 294, 2364–2368.

TRAPNELL, C., PACHTER, L. AND SALZBERG, S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25, 1105–1111.

TRAPNELL, C., WILLIAMS, B.A., PERTEA, G., MORTAZAVI, A., KWAN, G., BAREN, M.J. VAN, SALZBERG, S.L., WOLD, B.J. AND PACHTER, L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, 28, 511–515.

TUERK, A., WIKTORIN, G. AND GÜLER, S. (2017) Mixture models reveal multiple positional bias types in RNA-Seq data and lead to accurate transcript concentration estimates. *PLoS Comput Biol*, 13. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5448817/ [Accessed June 4, 2018].

UCSC GENOME BROWSER HOME. Available at: https://genome.ucsc.edu/ [Accessed June 13, 2018]

UESONO, Y., TOH-E, A. AND KIKUCHI, Y. (1997) Ssd1p of Saccharomyces cerevisiae associates with RNA. *J. Biol. Chem.*, 272, 16103–16109.

VELCULESCU, V.E., ZHANG, L., VOGELSTEIN, B. AND KINZLER, K.W. (1995) Serial analysis of gene expression. *Science*, 270, 484–487.

VILLAS-BÔAS, S.G., MOXLEY, J.F., AKESSON, M., STEPHANOPOULOS, G. AND NIELSEN, J. (2005) High-throughput metabolic state analysis: the missing link in integrated functional genomics of yeasts. *Biochem. J.*, 388, 669–677.

WAGNER, G.P., KIN, K. AND LYNCH, V.J. (2012) Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.*, 131, 281–285.

WANG, K., SINGH, D., ZENG, Z., *et al*. (2010) MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.*, 38, e178.

WANG, L., WANG, S. AND LI, W. (2012) RSeQC: quality control of RNA-seq experiments. *Bioinformatics*, 28, 2184–2185.

WANG, X. AND CAIRNS, M.J. (2013) Gene set enrichment analysis of RNA-Seq data: integrating differential expression and splicing. *BMC Bioinformatics*, 14 Suppl 5, S16.

WANG, Z., GERSTEIN, M. AND SNYDER, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10, 57–63.

WANLESS, A.G., LIN, Y. AND WEISS, E.L. (2014) Cell morphogenesis proteins are translationally controlled through UTRs by the Ndr/LATS target Ssd1. *PLoS ONE*, 9, e85212.

WIDMANN, C., GIBSON, S., JARPE, M.B. AND JOHNSON, G.L. (1999) Mitogen-activated protein kinase: conservation of a three-kinase module from yeast to human. *Physiol. Rev.*, 79, 143–180.

WILHELM, B.T., MARGUERAT, S., WATT, S., SCHUBERT, F., WOOD, V., GOODHEAD, I., PENKETT, C.J., ROGERS, J. AND BÄHLER, J. (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, 453, 1239–1243.

WILSON, R.B., BRENNER, A.A., WHITE, T.B., ENGLER, M.J., GAUGHRAN, J.P. AND TATCHELL, K. (1991) The Saccharomyces cerevisiae SRK1 gene, a suppressor of bcy1 and ins1, may be involved in protein phosphatase function. *Mol. Cell. Biol.*, 11, 3369–3373.

WINZELER, E.A., SHOEMAKER, D.D., ASTROMOFF, A., *et al*. (1999) Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science*, 285, 901–906.

WU, T.D. AND NACU, S. (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26, 873–881.

WU, Z. AND IRIZARRY, R.A. (2005) Stochastic models inspired by hybridization theory for short oligonucleotide arrays. *J. Comput. Biol.*, 12, 882–893.

YOUNG, M.D., WAKEFIELD, M.J., SMYTH, G.K. AND OSHLACK, A. (2010) Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.*, 11, R14.

ZHAO, S. AND ZHANG, B. (2015) A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. *BMC Genomics*, 16, 97.

ZHENG, W., CHUNG, L.M. AND ZHAO, H. (2011) Bias detection and correction in RNA-Sequencing data. *BMC Bioinformatics*, 12, 290.

ZHU, H., BILGIN, M., BANGHAM, R., *et al*. (2001) Global analysis of protein activities using proteome chips. *Science*, 293, 2101–2105.

7. ANEXOS

7.1. *SCRIPTS* GENERADOS

La colección de programas generados mediante programación en *Python*, se puede descargar desde los siguientes enlaces:

https://mega.nz/#F!SHBCwCqD!RJQ8HLIAHNIuISb-rYCVTw

7.2. RESULTADOS DEL CONTROL DE CALIDAD DE LAS LECTURAS POR *FASTQC*

En el enlace dispuesto a continuación se pueden descargar los archivos rendidos por *FASTQC* para cada una de las réplicas del experimento de RNAseq.

https://mega.nz/#F!SHZHjCxY!rNqV4q1MFddfbZ_iFK8d0w