PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
FACULTAD DE CIENCIAS BIOLÓGICAS
DEPARTAMENTO DE GENÉTICA MOLECULAR Y MICROBIOLOGÍA
PROGRAMA DE DOCTORADO EN CIENCIAS BIOLÓGICAS

MODELAMIENTO DE LOS ESTADOS DEL TRANSCRIPTOMA DE *A. THALIANA* FRENTE A PERTURBACIONES

Tesis entregada a la Pontificia Universidad Católica de Chile en cumplimiento parcial de los requisitos para optar al Grado de Doctor en Ciencias con mención en Genética Molecular y Microbiología.

Por: TOMÁS CUSTODIO MOYANO YUGOVIC
Director de Tesis: DR.RODRIGO ANTONIO GUTIÉRREZ ILABACA

AGOSTO 2018

AGRADECIMIENTOS

Agradezco a mi director de tesis, el Dr. Rodrigo Gutiérrez, por sus consejos, apoyo, paciencia, dedicación e interés por mi trabajo desde mi llegada al laboratorio hace 10 años. A Antoine de Daruvar por ayudarme a darle sentido, ordenar y animarme para realizar este trabajo. A Elena Vidal por su ayuda desde el principio de los tiempos, sus críticas, gran cooperación en el escrito, desarrollo de las ideas y formación como científico. Al Dr. Alvaro Soto por su ayuda y tiempo especialmente en la parte de análisis de texto. A Manolo Cabello por su ayuda en programación. A Jano por mantener funcionando los servidores en que se analizaron los datos. A Susana Cabello y Eleodoro por las correcciones en el escrito. A Pinwino y Joncito por la ayuda en las correcciones finales del trabajo. A todos los psblitos que me escucharon, corrigieron y presionaron para que terminara la tesis.

Agradezco también a los miembros de la comisión por sus críticas constructivas, especialmente al Dr. Alejandro Maass por darme consejos e invitarme a diferentes cursos que me sirvieron para el desarrollo del trabajo.

A todos los PSBLitos tanto antiguos como nuevos que en todos estos años me han apoyado tanto en el trabajo como en el ánimo para seguir adelante en la tesis y su paciencia para aguantarme. Especialmente a Karem por su paciencia de tantos años sentados juntos.

Agradezco a toda mi familia y amigos, especialmente a los que ya han partido, por darme su apoyo y compañia, que sin ellos esto no hubiese funcionado.

A las fuentes de financiameinto de este trabajo :

MIISSB Iniciativa Científica Milenio-MINECON, Beca de Doctorado Nacional Conicyt 2011, Fondo de Desarrollo de Areas Prioritarias (FONDAP) Center for Genome Regulation (15090007), Fondo Nacional de Desarrollo Científico y Tecnológico (FONDECYT) 1180759, y EvoNet (DE-SC0014377).

INDICE

INDICE DE FIGURAS

INDICE DE TABLAS

1. RESUMEN

Hoy en día es habitual el análisis de cambios en la expresión génica para identificar genes relevantes en la respuesta a una perturbación o a una transición en el desarrollo. Sin embargo, muchos de los genes claves para la respuesta de un organismo, no están regulados a nivel de la expresión génica (por ejemplo, genes tempranos en las vías de señalización). Esos genes actualmente están ocultos a los enfoques tradicionales basados en transcriptomas. En este trabajo, abordamos el problema de identificar genes funcionalmente relevantes para una condición "A", independiente si es que cambian sus niveles de expresión en una condición experimental contrastante. En esta tesis, proponemos un nuevo marco teórico basado en entropía para identificar estos genes. En primer lugar, determinamos los estados del transcriptoma y sus restricciones a partir de una gran cantidad de datos de expresión génica. Encontramos restricciones inherentes a la expresión génica a nivel global que revelan posibles nuevas relaciones funcionales para los genes, que no se obtienen por otros métodos ampliamente utilizados, tales como redes de correlación. Nuestro enfoque, además nos permitió encontrar nuevos genes relevantes en respuesta a perturbaciones, los cuales no necesariamente cambian su expresión en respuesta a dicha perturbación. Nuestro marco conceptual para analizar datos de transcriptomas fue evaluado en dos organismos modelos, *Arabidopsis thaliana* y *Saccharomyces cerevisiae.* Esta nueva metodología permite generar nuevas hipótesis acerca de redes regulatorias, las cuales no pueden ser alcanzadas con los métodos existentes. Adicionalmente, con el fin de contextualizar biológicamente listas de genes, se adaptó una metodología basada en análisis de texto que puede complementar la información de herramientas ya existentes, tales como Gene Ontology. Estas metodologías pueden ser fácilmente aplicadas a cualquier organismo que cuente con un número importante de datos transcriptómicos e información.

2. ABSTRACT

It is second nature nowadays to use changes in gene expression to identify relevant genes in response to a perturbation or in a developmental transition. However, many key genes for an organism's response are not regulated at the gene expression level (e.g. early genes in signaling pathways). These genes are hidden to conventional molecular profiling approaches such as transcriptome analysis. Here we sought to address the problem of finding functionally relevant genes for the condition "A" regardless of whether they change at the gene expression level under contrasting experimental conditions to evaluate the perturbation. In this thesis we propose a new framework based in entropy to identify these genes. In order to identify them, we first modeled transcriptome states and boundaries using large public expression databases. Using a novel entropy-based framework, we uncovered inherent restrictions in gene expression at the genome-wide level, which reveal novel functional relationships for genes. These associations had not been found by widely used methods such as correlation networks. Moreover, our approach allowed the identification of key genes involved in the response to perturbations. Interestingly, only some of them were transcriptionally regulated in response to the perturbation whilst others were not. Our novel transcriptomic analysis framework was evaluated in two model organisms, *Arabidopsis thaliana* and *Saccharomyces cerevisiae*. This new approach enables the formulation of novel hypotheses regarding gene regulatory networks, which are not attainable by conventional bioinformatics methods. Additionally, in order to aid the biological contextualization of gene lists, a methodology based on text analysis was adapted. It complements the information of existing tools, such as Gene Ontology. These two novel approaches can be easily applied for any organism with large transcriptome databases.

3. DEFINICIONES Y ABREVIATURAS.

Distribución de fondo: distribución de las razones de un par de genes en la colección completa del transcriptoma estudiado.

GPE (Gene Pair Entropy): Entropía de Shannon asociada a la distribución de razones para un par de genes dados.

ΔGPE: Es el valor que se obtiene al evaluar la diferencia entre el GPE inicial y el GPE al añadir la nueva relación entre genes dada por la nueva condición a la distribución de fondo y su valor de GPE, ajustado por el valor de GPE inicial.

EPI sum: (sum of Entropy Perturbation Index) Es la suma de todos los cambios en GPE al añadir una nueva condición a la distribución de fondo.

EPI: (Entropy Perturbation Index). Índice de perturbación de entropía, es la contribución individual de un gen al EPI sum, considerando la contribución en el GPE en cada par que participa.

RNA: ácido ribonucleico

DNA: ácido desoxi-ribonucleico

mRNA: RNA mensajero

PDB: Protein Data Bank

C: Carbono

N: Nitrógeno

R(i): Coeficiente de correlación del par i.

DLS: Discriminative local subspace.

KEGG: base de datos Kyoto Encyclopedia of Genes and Genomes.

GO: Gene Ontology.

LDA: Latent Dirichlet Allocation.

SE: entropía de Shannon (SE).

RSA: arquitectura del sistema radicular.

Nota: En este trabajo debido al idioma de los sistemas operativos y programas en que se realizan los cálculos es en inglés, el separador decimal es "." y el separador de grupo es ",".

4. INTRODUCCIÓN

Para sobrevivir, los organismos han evolucionado mecanismos que permiten modificar su desarrollo, crecimiento, metabolismo y fisiología para enfrentar y adaptarse a un entorno cambiante. La adaptación a un entorno variable es parcialmente dependiente del cambio en la expresión génica. Desde el inicio de la era genómica se ha estimado que más de la mitad de los genes de un organismo pueden modificar su expresión debido a cambios en su entorno (Causton et al., 2001). El desarrollo de tecnologías que permiten cuantificar el nivel de RNA de todos los genes de un organismo, ha producido una considerable cantidad de información acerca del estado transcripcional del genoma en diferentes situaciones. Considerando el estado del transcriptoma como el nivel de expresión de todos los genes en una condición y tiempo en particular, el número teórico de posibles estados del transcriptoma podría considerarse como ilimitado. Sin embargo, se ha demostrado que los patrones de distribución de la expresión génica siguen un número limitado de modelos de expresión. Esto indica que hay restricciones que gobiernan la expresión génica en los sistema biológicos (Aceituno et al., 2008; Krouk et al., 2009). Un ejemplo drástico de esta situación sería que todos los genes estén “apagados” o en

forma contraria, que todos los genes se expresen al máximo nivel de expresión. Estos estados extremos obviamente no se observan y existe amplia evidencia que las redes regulatorias, a nivel transcripcional por ejemplo, imponen restricciones sobre los niveles posibles de expresión observados para los genes de un organismo (Guo et al., 2011; Petricka et al., 2012). Esas restricciones son establecidas por la arquitectura intrínseca de las redes regulatorias génicas que orquestan los procesos celulares en los sistemas biológicos y que dictan las relaciones entre los genes (Kashtan y Alon, 2005; Guo et al., 2011), los cuales generan diferentes motivos de regulación (Alon, 2007).

4.1. Respuesta de un organismo al ambiente.

La mayoría de los organismos, especialmente las plantas debido a su naturaleza sésil están constantemente expuestos a diferentes tipos de estímulos ambientales. Para sobrevivir, los organismos deben responder a las señales externas requiriendo una fina interacción entre el genotipo y el ambiente para montar una respuesta (Friedel et al., 2012). Para la mayoría de las respuestas ambientales abióticas, el primer paso es detectar el cambio en la condición, lo que, dependiendo del estímulo y el sensor, genera una señal secundaria. Existen diferentes señales secundarias, como por ejemplo, cambios en los niveles de calcio en el citoplasma, incremento de especies reactivas de oxígeno, alteración de niveles de hormonas, entre otros. Estas alteraciones inician cascadas de señalización que provocan cambios en los niveles de transcritos que activan o reprimen funciones relevantes para la respuesta al estímulo (Kilian et al., 2007; Hahn et al., 2013). Las proteínas y otros productos génicos, en conjunto con los metabolitos forman redes de interacción funcional, las cuales constituyen la base molecular funcional de un

organismo vivo (Huang, 1999). Variaciones en alguna condición generan que la célula responda de manera global, orquestando cambios en la expresión génica, estados de proteínas, interacciones entre diferentes componentes y cambios en concentraciones de los elementos que están interactuando entre otras alteraciones (Huang, 1999). Todos estos cambios organizados permiten que el organismo pueda adaptarse o responder a la nueva condición.

4.2. Cuantificación de la expresión génica.

La expresión génica, puede ser entendida como el proceso en el cual un organismo utiliza la información codificada en el DNA para generar algún producto génico. La información codificada en los genes puede tener múltiples destinos, por ejemplo, producir una proteína o algún tipo de RNA no codificante. Existen a lo menos 4 procesos celulares que están involucrados en la expresión génica: la transcripción y degradación de RNA; y la traducción y degradación de proteínas. Cada uno de estos procesos, puede ser modificado en diversas etapas (Schwanhausser et al., 2011), lo que sugiere que desde que se transcribe el RNA hasta que una proteína está preparada para cumplir su función, pueden ocurrir muchos eventos de regulación. Esos procesos fundamentales de la célula están estrechamente coordinados y, en términos simples, su relación determina los niveles de productos génicos en un estado estacionario así como también las cinéticas de cambio de los transcriptomas (Opitz et al., 2010; Miller et al., 2011; Shalem et al., 2011).

El análisis global de transcritos, ya sea por microarreglos o secuenciación de RNA (RNA-seq), ha sido una herramienta útil para generar hipótesis que permitan a los investigadores entender el crecimiento, desarrollo o respuesta de organismos a señales internas o externas

(Usadel y Fernie, 2013; Malik, 2016; Moustafa y Cross, 2016). El creciente uso de metodologías de cuantificación masiva de RNA ha sido catalizado por la continua reducción de costos en las tecnologías de medición (Wetterstrand, 2016). Si bien existen métodos para cuantificar de forma masiva el nivel de proteínas, principalmente mediante el uso de espectrómetro de masas, la complejidad de las muestras excede la capacidad de secuenciación disponible (Rost et al., 2015; Schubert et al., 2017). Existen diversos estudios que han tratado de dilucidar cuál es la correlación entre los niveles de mRNA y proteína, en el trabajo de Schwanhusser et al. 2011 (Schwanhausser et al., 2011) encontraron que esta es relativamente baja. Sin embargo, nuevos estudios revisados en el trabajo de Liu et al 2016 (Liu et al., 2016), muestran que la correlación entre mRNA y proteína puede ser explicada entre un 55 y 84% por la variación en la abundancia de mRNA lo que indica que a pesar de que la relación entre mRNA y proteína puede variar, la medición de los niveles de transcrito a nivel global es una correcta aproximación de la expresión génica (Li et al., 2014).

4.3. Modificación de los niveles de RNA

La dinámica del transcriptoma es gobernada principalmente por dos procesos antagónicos, la biosíntesis y la degradación del ácido ribonucleico. Cuando se cuantifica el nivel de un transcrito en un momento y condición determinada, lo que uno observa, es la sumatoria de muchos eventos de regulación en que se activan y reprimen múltiples procesos para llegar a ese estado. Cuando hay un cambio en el ambiente, este es detectado, las tasas de transcripción y degradación de algunos genes es modificada para responder a la nueva condición, lo cual puede ser detectado en una nueva cuantificación del transcrito. Las modificaciones en la expresión

deben seguir ciertas reglas, describiendo trayectorias y restricciones (Huang, 2010; Wang et al., 2010; Huang, 2012). Los cambios en los niveles de RNA y sus restricciones pueden ilustrarse en un ejemplo teórico descrito en Huang 2007 (Huang et al., 2007). Tenemos 2 genes, X e Y, los cuales interactúan entre sí y van a determinar el destino de una célula. Estos genes se inducen a sí mismos y se reprimen entre ellos, además de degradarse en el tiempo. En este ejemplo, los niveles de estos genes dependerán de constantes de transcripción, degradación y de la concentración de los elementos del sistema. Dependiendo de las concentraciones de X e Y, estos pueden variar siguiendo ciertas trayectorias para llegar a algún estado estable. Debido a que X e Y son moléculas que se reprimen mutuamente, es poco probable que X e Y se encuentren en su máxima expresión, esto correspondería a un estado inestable o poco probable. En cambio, si es que uno de ellos posee expresión alta y el otro baja, es probablemente un estado más estable o de baja pseudo-energía libre. Esta energía es llamada por Huang (2007, 2009, 2012) como energía quasi-potencial, de manera que cada estado posee su propio quasi-potencial de energía.

Escalar lo anterior al genoma completo puede entregar información importante que posibilite determinar las trayectorias de expresión que siguen las células en su desarrollo y frente a cambios en condiciones experimentales. Para analizar esto necesitamos una métrica que nos permita transformar los datos para llevarlos a un espacio que nos facilite la cuantificación de cuál es el costo de cambiar de estado.

4.4. Aproximaciones de Biología de Sistemas para comprender el funcionamiento de organismos.

El análisis de la biología a nivel sistémico, tiene como objetivo central entender la estructura y la dinámica del comportamiento que emerge de los componentes moleculares de un sistema y sus relaciones funcionales (Ideker et al., 2001; Kitano, 2002; Gutiérrez et al., 2005). Un enfoque de biología de sistemas para estudiar la fisiología de las plantas u otro organismo vivo implica modelar el sistema como un todo, en lugar de un conjunto de partes. Sin embargo, la precisión de este enfoque depende en gran medida del conocimiento existente sobre los componentes e interacciones de las partes que constituyen el sistema, así como de métodos fiables para manejar, integrar, analizar y visualizar grandes conjuntos de datos.

Durante la última década, los avances en métodos experimentales que generan grandes conjuntos de datos aceleraron el desarrollo de recursos que integran información en diferentes especies modelo. El desarrollo de las plataformas de microarreglos y de secuenciación masiva ha sido particularmente importante para la generación de datos. Esto principalmente, debido a sus diversas aplicaciones, que incluyen la secuenciación de genomas, la medición del nivel de RNA, inmuno-precipitación de cromatina y el análisis de marcas epigenéticas (Lister et al., 2009) entre otras muchas aplicaciones. Otras fuentes de datos biológicos que proporcionan información importante acerca de las relaciones funcionales entre genes, son los conjuntos de datos de interacción proteína-proteína a gran escala determinados por ensayos de doble-híbrido en levadura, espectrometría de masas e inmuno-precipitación (Bracha-Drori et al., 2004; Ciruela, 2008). Además, las asociaciones de proteínas y DNA proporcionan un punto de partida para construir redes regulatorias. Estas asociaciones se predicen a menudo basadas en elementos

regulatorios en cis y sitios de unión conocidos de factores de transcripción (Weirauch et al., 2014; O'Malley et al., 2016).

Un desafío importante es la integración e interpretación de estos conjuntos de datos masivos con el fin de generar hipótesis sobre las redes de interacción que rigen los comportamientos del sistema (por ejemplo, los mecanismos moleculares subyacentes a las respuestas a las señales ambientales). La teoría de la red aplicada a los datos biológicos ha demostrado ser extremadamente útil para integrar tipos de datos heterogéneos y descubrir principios organizativos en sistemas biológicos (Barabasi y Oltvai, 2004). Una red de genes captura las dependencias entre las entidades moleculares que forman parte de un sistema. Las redes génicas se representan generalmente como gráficos en los que cada nodo representa una entidad molecular (por ejemplo, genes, proteínas, metabolitos) y las líneas representan relaciones funcionales entre ellos (por ejemplo, interacciones proteína-proteína, interacciones proteína-DNA, microRNA:blanco, coexpresión). La integración de diferentes tipos de datos a gran escala mejora la reconstrucción de la red y permite una mejor comprensión de la estructura y regulación del sistema (Joyce y Palsson, 2006; Karlebach y Shamir, 2008).

Las redes de interacción permiten entender la estructura de importantes procesos biológicos en plantas. Una de las primeras redes cualitativas de Arabidopsis fue construida integrando distintos tipos de datos, incluyendo interacciones regulatorias y metabólicas de 6176 genes y 1459 metabolitos (Gutiérrez et al., 2007b). Esta red incluye 230,900 conexiones representando diferentes relaciones funcionales (regulatorias, metabólicas, físicas), la cual fue inicialmente usada para determinar módulos en la red controlados por carbono (C) y/o metabolitos de

nitrógeno (N) (Gutiérrez et al., 2007b). En ese estudio, el análisis de la red condujo a la hipótesis de que la señalización de auxina estaba implicada en la respuesta de la raíz de Arabidopsis a los metabolitos C y / o N (Gutiérrez et al., 2007b). Esta hipótesis se confirmó posteriormente mediante otros análisis informáticos y experimentales (Vidal et al., 2010a; Krouk et al., 2010b; Vidal et al., 2010b; Vidal et al., 2013; Canales et al., 2014; Vidal et al., 2014).

A pesar de ser cualitativo e incompleto, este modelo de red resultó ser extremadamente útil para generar hipótesis comprobables concretas en esta y una serie de estudios que siguieron (Gutiérrez et al., 2007a; Gutiérrez et al., 2007b; Krouk et al., 2010a; Ruffel et al., 2010; Vidal et al., 2010a; Vidal et al., 2013; Alvarez et al., 2014). Por ejemplo, el análisis de redes sugirió un circuito regulador entre el reloj circadiano y la nutrición por N en Arabidopsis (Gutiérrez et al., 2008). El análisis sistémico mostró que *CIRCADIAN CLOCK ASSOCIATED 1* (CCA1), uno de los reguladores maestros del reloj circadiano, coordina la respuesta de genes asimiladores de N por unión directa a los promotores de *BASIC REGION/LEUCINE ZIPPER TRANSCRIPTION FACTOR 1,* el cual regula la expresión de *ASPARAGINE SYNTHETASE 1*, *GLUTAMINE SYNTHETASE 1.3*, y GLUTAMATE DEHYDROGENASE 1 *GLUTAMATE DEHYDROGENASE 1* (Gutiérrez et al., 2008), todos estos, genes importantes en la asimilación de N orgánico. A su vez, los metabolitos de N pueden actuar como un regulador del reloj circadiano a través de la modulación de la expresión del gen CCA1 (Gutiérrez et al., 2008).

La generación de redes depende en gran medida del análisis computacional para poder manejar y emplear adecuadamente datos heterogéneos y presentados en diferentes formatos. Se han desarrollado varias herramientas y recursos en línea que nos ayudan a integrar y utilizar los

datos disponibles en plantas, así como en otros organismos, por ejemplo, VirtualPlant (Katari et al., 2010), CORNET (De Bodt et al., 2010; De Bodt et al., 2012), STRING (Franceschini et al., 2013), GeneMania (Zuberi et al., 2013), ATTED-II (Obayashi et al., 2014).

Estas herramientas y bases de datos nos muestran que existen numerosos elementos conectados, aunque la información sobre las interacciones es limitada. Estudios genómicos y proteómicos de redes moleculares, nos indican que las interacciones moleculares en la célula, forman un gran componente conectado, que podría abarcar el 90% de los genes del genoma revisado por Huang et al, 2005 (Huang et al., 2005). Esto sugiere que aún quedan muchas relaciones que caracterizar y que cambios en la expresión de un gen, puede afectar genes que hasta el momento no se han señalado como relacionados.

4.5. Métodos para calcular relaciones entre los genes.

Las relaciones entre los genes son usualmente abordadas mediante herramientas matemáticas que miden relaciones utilizando la información de expresión génica. Estas relaciones, generalmente se expresan en forma de redes, en las cuales se representan genes conectados por aristas que representan la relación entre sus patrones de expresión. Existe una gran variedad de métodos que intentan reconstruir redes génicas basados en datos de expresión. En general, los métodos más utilizados y más sencillos para la reconstrucción *in-silico* de redes son los basados en medidas de similaridad (Yan et al., 2017). Uno de los principales objetivos de encontrar patrones similares de expresión frente a diferentes condiciones, es que si un grupo de genes posee patrones de expresión similares, es probable que compartan una función o que tengan algún mecanismo de control común (Eisen et al., 1998; Cheng et al., 2011). Existen

diferentes métodos para encontrar similitudes en los patrones de expresión, dentro de los primeros utilizados debido a su facilidad de implementación está la distancia euclidiana. La distancia euclidiana mide la distancia absoluta entre 2 patrones de expresión, la cual entrega una información intuitiva, de la relación entre los genes. Sin embargo, la distancia euclidiana puede tener sesgos debido a escalamiento de los datos y diferencias en las medias de niveles de expresión (D'haeseleer et al., 2000). Otro método que mide similaridad, el cual es uno de los más utilizados, es el coeficiente de correlación. El valor del coeficiente de correlación fluctúa entre 1 y -1 de acuerdo con el comportamiento de los genes, lo que hace intuitivo comprender utilizando el valor de correlación que valores cercanos a 1 indican un patrón de alta similitud y un valor cercano a -1 indica un comportamiento en que los genes se comportan totalmente opuestos. El método más utilizado para determinar la correlación entre la expresión de un par de genes es el coeficiente de correlación de Pearson, el cual mide la relación lineal entre 2 variables. La correlación de Pearson se puede definir como:

$$R(i) = \frac{\sum_{k=1}^{m}(x_{k,i} - \bar{x}_i)(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{m}(x_{k,i} - \bar{x}_i)^2 \sum_{k=1}^{m}(y_k - \bar{y})^2}}$$

donde el numerador es la covarianza y el denominador es la raíz del producto de las varianzas de cada variable (Guyon y Elisseeff, 2003). Una variante para el cálculo de correlación de Pearson es el coeficiente de Spearman. Este es un método no paramétrico basado en el coeficiente de correlación de Pearson, donde se reemplaza el valor de expresión por el ranking (Li et al., 2015). Esta aproximación ha sido utilizada en múltiples análisis de datos biológicos

dado que al utilizar el ranking en vez del valor de expresión lo hace más robusto a datos muy alejados y no es necesaria la normalización (Song et al., 2012). Sin embargo, se pierde información debido a la conversión de valores numéricos a rankings (Wang y Huang, 2014). Acompañando al coeficiente de correlación, generalmente se calcula la significancia de la correlación para descartar altas correlaciones debidas al azar (Li et al., 2015).

Existen diferentes herramientas basadas en correlación, una de las más utilizada es el paquete de bioconductor WGCNA (Langfelder y Horvath, 2008), el cual tiene como objetivo encontrar módulos o grupos de genes que posean una misma regulación mediante correlaciones ponderadas. Esta herramienta no solo cuantifica la correlación entre 2 genes sino que también se extiende a genes que comparten los mismos vecinos, por lo que se detectan módulos que comparten el mismo patrón de expresión (Langfelder y Horvath, 2008; DiLeo et al., 2011).

Típicamente, el análisis de correlación se realiza sobre todos los datos disponibles en un gran número de experimentos. Una variante a los análisis de correlación global fue desarrollada en nuestro laboratorio en el trabajo de Puelma et al 2012,2016 (Puelma et al., 2012; Puelma et al., 2017). En el cual mediante el método DLS (Discriminative Local Subspaces) (Puelma et al., 2012) permite la identificación de nuevos genes involucrados en procesos de interés utilizando una red de correlación, pero basándose solo en los experimentos en que se puedan diferenciar los genes discriminativos del proceso de interés. Este método fue aplicado con éxito en el trabajo de Araus 2016 (Araus et al., 2016).

En cuanto a métodos basados en la teoría de la información, la medida de relación más utilizada es la información mutua. La información mutua se basa en la entropía de Shannon y permite evaluar dependencia en el comportamiento de 2 genes (Bansal et al., 2007). Tiene como característica que permite detectar tanto relaciones lineales como no lineales entre variables. Si

la información mutua es 0 o cercana a 0, quiere decir que las variables analizadas no comparten información, por lo que son independientes. Si es mayor que 0, dice que la información de una variable aporta información sobre la otra variable analizada, así mientras mayor sea el valor sugiere mayor relación entre los genes. Matemáticamente la información mutua se define como:

$$I(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

donde p(x,y) es la probabilidad conjunta de que X =x e Y = y, y p(x) es la probabilidad de que X=x, y p(y), es la probabilidad de que Y=y (Guyon y Elisseeff, 2003). La información mutua se ha utilizado en múltiples trabajos y en la herramienta más popularmente utilizada es el paquete ARACNE (Margolin et al., 2006), el cual a la fecha posee más de 3000 citaciones.

Existen otro tipo de redes inferidas a partir de datos de expresión, son las redes bayesianas, las cuales combinan la probabilidad con la teoría de grafos para construir la red (Li et al., 2015). Para estas redes se calcula la probabilidad de que los datos observados se ajusten a posibles redes y se selecciona la topología que mejor explique el conjunto de datos observados (Yan et al., 2017). Como se pueden generar un gran número de topologías posibles, las cuales aumentan a medida que se le añaden más genes a la red, el tiempo de cálculo aumenta exponencialmente (Li et al., 2015). Las redes bayesianas son redes regulatorias que indican direccionalidad, un gen influye sobre otro indicando causalidad. Una variante de redes bayesianas son las redes bayesianas dinámicas, las que requieren para su inferencia series de datos temporales largos (Yan et al., 2017) los cuales no son comunes en biología.

Con el objetivo de de detectar la relación entre genes se utilizan reglas de asociación, las que determinan como la expresión de un gen en particular podría afectar la expresión de otros genes (Creighton y Hanash, 2003). Un importante logro de este método es poder predecir o encontrar reglas que describen el comportamiento de un gen a partir del comportamiento de otro grupo de genes. Con esta información se asume que los genes implicados podrían ser parte de una misma red génica, asumiendo que los genes pertenecientes a esta red poseen patrones de expresión comunes, así como también determinar cómo se expresan los genes en diferentes condiciones celulares (Creighton y Hanash, 2003). Esto ha ido desarrollándose en los últimos años, pero existen dos grandes problemas: su alto costo computacional debido a que se requiere de un gran conjunto de series temporales para su inferencia y que en este tipo de análisis se generan reglas ambiguas o que se contradicen, por lo que en la mayoría de los algoritmos son descartadas, quedando como válidas solo un pequeño número de ellas (Chen et al., 2015).

Cualquier medida de dependencia del perfil de expresión entre 2 genes puede sugerir una posible relación biológica, pero no necesariamente interacciones físicas entre los miembros del par. El análisis de 2 variables independientes puede tener muchas interpretaciones y resultados según el método que se utilice para analizarlo. Aunque se han desarrollado múltiples herramientas para analizar posibles dependencias entre patrones de expresión, la correlación sigue siendo la herramienta más utilizada para evaluar posibles dependencias entre genes (Li et al., 2015).

4.6. Búsqueda da genes relevantes.

Desde la década de 1990, donde se comienza a obtener datos de forma "masiva" de expresión génica, se busca ver que conjuntos de genes se expresan diferencialmente. Uno de los objetivos claves para los cuales se utiliza la obtención de datos transcriptómicos es para la identificación de genes relevantes, ya sea para conocer la respuesta a un cambio en una condición ambiental, cuáles son los genes que están relacionados con alguna enfermedad o estímulo entre otras posibilidades. A la fecha, al realizar una consulta en el buscador de Pubmed con las palabras "transcriptome differential expression" nos da como resultado alrededor de 96200 artículos relacionados con estas palabras, lo que nos muestra la importancia en su utilización. Generalmente los experimentos se hacen tomando en cuenta a lo menos 2 condiciones, un control y un tratamiento, y se realiza un análisis utilizando alguna prueba estadística (Audic y Claverie, 1997). Existen otras formas de encontrar genes relevantes basados en la expresión génica, una de ellas es el análisis de redes parenclíticas, en las cuales se compara la red formada por la expresión de los genes en cada condición y se buscan las diferencias topológicas para obtener genes relevantes a la condición estudiada (Zanin et al., 2014). En el trabajo de Wang *et al.*, (2014) postulan que, si bien la expresión diferencial es valiosa, tiene la limitación que solo detecta grandes alteraciones en los niveles de expresión ya que diferencias en el promedio de los niveles de expresión entre tratamiento y control pueden no ser detectados. Proponen un método en que utilizan la variabilidad y la entropía de Shannon del nivel de expresión en vez de la diferencia entre las medias de expresión, concluyendo que otros métodos permiten obtener información que no se obtiene por los típicos análisis de expresión diferencial (Wang et al., 2014).

4.7. Análisis de grupos

Uno de los resultados más frecuentes en análisis de datos ómicos son conjuntos o listas de genes. Como se trabaja generalmente con un gran número de genes, es necesario utilizar herramientas que permitan resumir la información contenida en estas listas. Generalmente se realiza un análisis funcional de estos grupos de genes, identificando así, por ejemplo, que rutas o procesos biológicos están sobre-representados, es decir, que se encuentren en una proporción mayor a lo esperado por azar. Estos enriquecimientos funcionales nos permiten comprender biológicamente lo que está ocurriendo en el experimento que se está evaluando. La clave para derivar información biológica de estas listas de genes es detectar la presencia de conjuntos de genes que compartan alguna propiedad biológica, como por ejemplo una función o un mecanismo regulatorio. El análisis de enriquecimiento de vías o funciones estaría detectando estos bloques de genes que comparten un papel en la célula para realizar una función (Ponzoni et al., 2014).

Para realizar estos análisis funcionales se requiere recopilación de conocimientos previos, generalmente contenidas en bases de datos como por ejemplo, la base de datos de KEGG que contiene información de genes asociados a vías metabólicas en diferentes organismos (Kanehisa, 2002). Una de los recursos de información de genes más importantes y ampliamente utilizado es la base de datos de Gene Ontology (GO,http://www.geneontology.org)(Gene Ontology Consortium, 2004). El proyecto GO reúne información biológica de genes o sus productos, de una forma ordenada utilizando una estructura de árbol jerárquico y un vocabulario definido que puede ser comprendido y utilizado para integrar y analizar información de forma computacional (Yon Rhee et al., 2008). Este proyecto posee información depositada por diferentes colaboradores para múltiples organismos, siendo una de las bases de datos más

completa de información de función biológica de genes, la cual es ampliamente utilizada para hacer análisis de enriquecimiento funcional (Yi et al., 2013). Sin embargo, está importante base de datos tiene el problema que la cobertura de genes anotados es baja si no se consideran anotaciones en los primeros niveles del árbol (Yi et al., 2013). Para buscar la función o ver la relación que tiene un gen con alguna condición, es posible buscar en literatura, sin embargo, la profundidad con que se hace la búsqueda puede generar diferentes resultados. Existen diferentes herramientas utilizadas en análisis de semántica de texto que permiten describir artículos mediante métodos que analizan la semántica usando modelos probabilísticos. Estos análisis han sido ampliamente utilizados en el ámbito de las comunicaciones (Blei et al., 2012) permitiendo resaltar la temática del texto de forma objetiva y sin necesidad de lectura, mientras en el ámbito biomédico se ha utilizado para observar coocurrencia entre drogas y genes (Wu et al., 2012). En este trabajo, adaptaremos el método conocido como "Latent Dirichlet Allocation" (LDA) descrito por Blei (Blei et al., 2012), para analizar listas de genes de manera objetiva, con el mínimo sesgo del investigador.

4.8. Una nueva estrategia de análisis de datos de expresión.

Como en una red de interacciones el comportamiento de un gen influye sobre el resto, la información conjunta de grupos de genes puede entregar más información que el comportamiento de cada gen por separado. Con el fin de caracterizar relaciones entre los genes e identificar límites que establecen restricciones sobre los posibles estados del transcriptoma, en esta tesis hemos desarrollado una métrica basada en la entropía de Shannon (Shannon, 1948). La entropía de Shannon (SE) es una medida del contenido o complejidad de la información, originalmente desarrollada para la tecnología de las comunicaciones (Shannon, 1948). La SE

puede proporcionar una medida de la información contenida en un patrón de expresión génica. En biología, se ha utilizado previamente como indicadores para identificar posibles blancos de fármacos (Cunningham et al., 2000; Fuhrman et al., 2000; Wang et al., 2014), para cuantificar y clasificar la especificidad de los genes en un tejido (Schug et al., 2005) y como algoritmos para la agrupación de datos, así como también para detectar genes importantes para la respuesta frente a un estímulo (Wang et al., 2014).

Como mencionamos anteriormente, al hacer un estudio de expresión génica frente a un estímulo, generalmente el análisis de datos se centra en identificar los genes que cambian su expresión frente a un estímulo. Sin embargo, puede que existan genes importantes para el estímulo que no cambien significativamente su expresión. Si cada gen o producto génico de un organismo puede situarse en una red génica, este puede ser influido por el resto de los genes. En este trabajo, analizaremos el comportamiento de cada pareja de genes, para lo cual utilizaremos la entropía de Shannon de la distribución de la razón de expresión de cada par de genes. A este valor de entropía le llamaremos "entropía del par de genes" o GPE (Gene Pair Entropy). Este valor lo utilizaremos para caracterizar relaciones entre genes y dar un valor objetivo a posibles restricciones de comportamiento entre ellos en un contexto global. Esto lo realizaremos en dos organismos, *Arabidopsis thaliana* y *Saccharomyces cerevisiae.* Observaremos que modificaciones en las condiciones ambientales pueden provocar perturbaciones en los valores de GPE, estas perturbaciones las cuantificamos y las llamamos EPI (Entropy Perturbation Index). Esta nueva métrica puede ayudar a resaltar genes importantes en la respuesta a un estímulo ambiental, lo que en algunos casos no puede ser observado como un cambio significativo en los niveles de mRNA.

5. HIPÓTESIS Y OBJETIVOS

5.1. Hipótesis de trabajo.

Los cambios del transcriptoma pueden ser evaluados en base a las restricciones de los estados del transcriptoma observados.

5.2. Objetivo general.

Determinar y modelar los posibles estados del transcriptoma y evaluar sus cambios frente a perturbaciones basado en sus restricciones observadas.

5.3. Objetivo específico 1.

Evaluar los estados del transcriptoma reportados en bases de datos disponibles públicamente.

En esta etapa se recopiló y organizó información de bases de datos de expresión global de genes para conocer el comportamiento de estos frente al mayor número de condiciones experimentales posibles. Esta información se organizó para seleccionar grupos de experimentos que mejor se ajustaran a las interrogantes del trabajo y de acuerdo con la condición experimental. Se descartaron experimentos redundantes y de baja calidad. Se eliminó del análisis sondas que poseían bajo nivel de señal o saturadas, así como también aquellas que mostraron poca variación en las diferentes condiciones experimentales evaluadas. Con esto se generó una matriz que contiene la información con la que se desarrolló el resto del trabajo.

5.4. Objetivo específico 2.

Desarrollar modelos para evaluar las restricciones de los posibles estados del transcriptoma de *Arabidopsis thaliana*.

Bajo la premisa de que existen estados del transcriptoma más frecuentes o representativos, deben existir reglas que mantengan la relación entre la expresión de los genes. Para determinar si es que hay reglas que gobiernan el comportamiento entre los genes, se determinaron las relaciones de expresión entre todos los pares de genes analizados. Para caracterizar estas relaciones, se analizó la distribución de razones entre los genes en las diferentes condiciones experimentales y se calculó la entropía de Shannon de cada distribución como medida de

relación entre genes. Se obtuvo una medida de co-expresión que representa las restricciones en el comportamiento de un par de genes. Nuestros resultados muestran que esta aproximación puede entregar información biológica complementaria a los métodos ampliamente utilizados en literatura.

5.5. Objetivo específico 3.

Explorar cambios entre estados del transcriptoma utilizando el modelo antes planteado.

Del objetivo anterior se obtuvieron las relaciones más frecuentes y la distribución de razones de expresión de cada par de genes evaluados. Debido a que nos interesa conocer la forma en que evoluciona un transcriptoma frente a un estímulo, se propuso un método basado en la diferencia de entropía generada por un estímulo para identificar los genes relevantes para el estímulo analizado. Con este fin, se utilizó una selección de experimentos en que se evalúa el transcriptoma luego de un estímulo y así determinar el comportamiento de los genes. El resultado obtenido, se comparó con un método ampliamente utilizado previamente para conocer los genes diferencialmente expresados ante un estímulo. Luego de la comparación, se encontró que el método propuesto, el cual considera las relaciones de un gen con el resto, entrega información adicional de la respuesta del transcriptoma frente a un estímulo, sugiriendo que es un método complementario a los previamente descritos.

5.6. Objetivo específico 4.

Aplicación del modelo para determinar los procesos asociados a cambios de estado del transcriptoma.

En esta etapa se desarrollaron y adaptaron herramientas que permiten el análisis de los procesos biológicos y funciones moleculares asociadas a los genes encontrados en el objetivo anterior, ya sea regulados diferencialmente o que presenten perturbaciones en su entropía. Con estos genes, se construyeron redes y se identificaron genes o grupos de genes involucrados en procesos biológicos asociados al estímulo con el que fueron tratados. Para esto se realizó un análisis de sobrerrepresentación de procesos biológicos anotados en la base de datos de Gene Ontology. Adicionalmente, de manera de poder asignar posibles funciones a genes no anotados en Gene Ontology, se desarrolló una herramienta de análisis semántico. Esta herramienta permite asignar “temas” o “tópicos” relacionados a cada gen basado en textos de resúmenes de publicaciones. Una versión web de esta herramienta está en desarrollo y estará prontamente disponible para la comunidad científica.

6. MÉTODOS

6.1. Recopilación de datos de expresión.

Para *Arabidopsis thaliana*, los datos de expresión se obtuvieron del repositorio de NASCArrays (Craigon et al., 2004), desde el cual se descargaron un total de 249 series experimentales, incluyendo 4815 hibridaciones basadas en el ampliamente utilizado chip ATH1 de Affymetrix. Para levadura, los datos se descargaron de la base de datos ArrayExpress (Brazma et al., 2003; Kolesnikov et al., 2015). Se descargaron un total de 117 series experimentales, incluyendo 2078 hibridaciones del Affymetix GeneChip Yeast Genome 2.0 Array. Solamente sondas marcadas como provenientes de la especie *S. Cerevisiae* fueron utilizadas.

6.2. Pre-procesamiento de datos.

Los archivos de datos crudos .CEL se procesaron utilizando R (R Core Team, 2015). Un primer control de calidad se realizó usando el algoritmo IQRray (Rosikiewicz y Robinson-rechavi, 2011), con el cual se eliminó un 5% de los archivos .CEL que presentaban más baja calidad de hibridación. Luego de este filtro, se eliminaron las series experimentales que tuviesen menos de 4 archivos ".CEL", así como también archivos duplicados. Los datos filtrados se normalizaron utilizando el método Robust Multiarray Analysis (RMA) (Irizarry et al., 2003) del paquete *affy* de Bioconductor. Un segundo filtro a nivel de sondas (probesets) se realizó, eliminando sondas de asignación ambigua al gen correspondiente (p.ej, sondas asignados a múltiples genes). Si un gen es medido por más de una sonda, se utilizó el promedio de las sondas involucradas. Adicionalmente, se descartaron sondas de acuerdo con los siguientes filtros: niveles de expresión del gen con una desviación estándar menor a 0.25, intensidad de expresión menor a 6 y con varianza de expresión menor a 0.5 y genes correspondientes al 1% más expresado, de manera de evitar saturación de la sonda.

6.3. Recopilación de datos de Interacción biológica

Datos de interacción biológica fueron descargados desde bases de datos públicas. Para Arabidopsis se descargó la base de datos ANAP (V1.2) (Wang et al., 2012), la cual recopila información de interacciones funcionales de la mayoría de las bases de datos existentes de Arabidopsis. Un total de 146455 pares de genes interactores fueron recolectados incluyendo interacciones confirmadas de forma experimental y/o predichas. Para *Saccharomyces cerevisiae*, los datos de interacción fueron obtenidos de la base de datos BioGRID (Stark et al.,

2006; Chatr-aryamontri et al., 2015) en la que se encontraron 304198 de pares de genes interactuantes.

6.4. Recopilación de datos de vías y procesos metabólicos.

La información de vías y procesos metabólicos tiene como fuente la base de datos KEGG (Kanehisa, 2002), la cual fue obtenida de la base de datos PlantGSEA (Yi et al., 2013) para *A. thaliana* y de Saccharomyces genome database (Cherry et al., 2012) para levadura.

6.5. Análisis de redes.

La visualización, exploración y análisis de redes fue desarrollado utilizando Cytoscape (Shannon et al., 2003; Kohl et al., 2011). El análisis topológico fue realizado con el algoritmo Antipole (Ferro et al., 2006). El análisis de enriquecimiento funcional para identificar procesos biológicos sobrerrepresentados fue realizado con el complemento BiNGO (Maere et al., 2005) y ClueGO (Bindea et al., 2009) con valores por defecto y utilizando principalmente los niveles 5 y 6 desde la raíz del árbol de Gene Ontology y un valor de P ajustado menor a 0.05.

6.6. Determinación de valores de co-expresión y entropía ente pares de genes

Las dos medidas más usadas de co-expresión de genes fueron calculadas: Correlación lineal e Información Mutua. La correlación fue calculada mediante la función nativa de R “cor”, la información mutua se calculó con el programa de MATLAB FastPairMI.m descrito por Qiu y cols. (Qiu et al., 2009).

En este trabajo, se adaptó un método para definir dependencia entre pares de genes a partir de los perfiles de expresión, al cual hemos llamado entropía de pares de genes (GPE, gene pair entropy) (Figura 1). Después del pre-procesamiento de los datos, se obtiene la matriz de expresión en base $\log_2$ de todas las sondas en las diferentes condiciones. Se calculó la diferencia en el valor de expresión de cada par de genes y se redondeó al entero más cercano quedando de forma discreta. Debido a que la matriz de expresión está en base logarítmica, la diferencia corresponde a la razón de expresión entre cada par de genes evaluados. Lo anterior se realizó para todos los pares presentes en la matriz por lo que se obtiene una matriz de razones de expresión (o diferencia en base $\log_2$). Estos valores fueron tabulados para obtener una matriz de frecuencia de razones de expresión de todos los pares posibles a la cual llamamos distribución de fondo ("background distribution") para cada par de genes. La dependencia o restricciones entre los niveles de expresión de pares de genes fue evaluada calculando la Entropía de Shannon (Shannon, 1948) de cada distribución mediante la siguiente formula utilizando un script de perl:

$$\text{GPE} = H = -\sum p_i \log_2 p_i$$

siendo p_i la probabilidad que una razón de expresión discretizada ocurra dada por la frecuencia en las observaciones.

6.7. Cálculos del estado del transcriptoma y sus perturbaciones.

La suma de valores GPE de todos los genes fue usada como una medida para definir el estado basal del transcriptoma de una especie en una condición experimental determinada. Debido a un cambio en el estado del desarrollo o a una variación en las condiciones ambientales, los genes pueden regularse, induciéndose o reprimiéndose, respecto a una condición de

referencia. Consecuentemente, las razones de cambio entre la expresión de estos genes regulados y todo el resto es potencialmente afectada. Cuando se añade esta nueva condición, a la que se definió anteriormente como distribución de fondo, el valor de GPE puede cambiar en diferentes magnitudes dependiendo de la forma de la distribución. Para resaltar este cambio, el valor de cambio de GPE, es ajustado por el valor de GPE inicial. Lo cual fue llamado como ΔGPE, el cual está definido en la siguiente formula:

$$\Delta GPE = \frac{GPE_0 - GPE_n}{GPE_0}$$

Siendo GPE_0 el valor de GPE antes de añadir el nuevo experimento y GPE_n el valor de GPE después de añadir la nueva condición. Este valor fue calculado para todos los pares de genes del transcriptoma. La suma de todos los cambios en GPE fue llamada suma de índices de perturbación de entropía (sum of entropy perturbation index , EPI sum). Para un cambio en las condiciones ambientales es posible medir el EPI sum de la condición. La contribución individual de un gen al EPI sum, es llamada índice de perturbación de entropía (entropy perturbation index, EPI). La cual puede ser expresada en la siguiente formula:

$$EPI_j = \sum_j \Delta GPE$$

donde el EPI para el gen "j" está dado por la sumatoria de todos los valores de ΔGPE en que el gen "j" está involucrado. Luego de un cambio en las condiciones experimentales, se pueden identificar diferentes clases de respuesta de los genes. Los genes que muestran una expresión diferencial (DE) y los que muestran un cambio en EPI diferencial (Differential entropy perturbation index, DEPI). Estas clases de respuestas fueron estadísticamente definidas usando

la librería Rankprod (Hong et al., 2006) de Bioconductor. La significancia fue definida con un porcentaje de falsos positivos (pfp) < 0.05 para DE y < 0.1 para DEPI (solo inducidos).

6.8. Genes de respuesta múltiple.

Es definido como genes de respuesta múltiple (Multi Response Genes, MRG), el 5% de los genes de Arabidopsis que según el trabajo de Aceituno y cols. (Aceituno et al., 2008), que aparecen en un mayor número de experimentos como diferencialmente regulados en las condiciones tratamiento/control analizadas.

6.9. Conjunto de genes de referencia de Gene Ontology.

Para cada condición experimental probada (frío, osmótico, estrés oxidativo, salino, nitrato), una lista de genes de Arabidopsis anotados en términos de Gene Ontology (GO) asociados al estímulo fueron utilizados para evaluar los genes asociados a cada señal. Para los diferentes tratamientos se utilizaron los siguientes términos GO:

Nitrato	Frio	Osmótico	Oxidativo	Salino
cellular response to nitrate	cellular response to cold	cellular response to osmotic stress	cellular response to oxidative stress	cellular response to salt
nitrate assimilation	cold acclimation	regulation of response to osmotic stress	Regulation of response to oxidative stress	cellular response to salt stress
nitrate transport	response to cold	response to osmotic stress	response to oxidative stress	regulation of response to salt stress
response to nitrate				response to salt
				response to salt stress

6.10. Asociación de genes con publicaciones.

La relación entre genes y las publicaciones asociadas fueron descargadas utilizando la herramienta Thalemine de Araport (https://apps.araport.org/thalemine/) (Krishnakumar et al., 2017). Para extraer el resumen de la publicación asociada se utilizó la herramienta "efetch" de Entrez Programming Utilities (https://www.ncbi.nlm.nih.gov/books/NBK25501/) (Wheeler et al., 2008).

7. RESULTADOS.

7.1. Definiendo un estado.

Cada uno de los transcritos en una célula se pueden representar de igual manera tal que en una condición arbitraria *i* el estado del transcriptoma se puede definir:

$$s_i = [x_1, x_2, x_3, \ldots, x_n]$$

Donde "x" corresponde a la expresión de un gen y el conjunto de todos los *s* definen una matriz S que representa todos los estados posibles que puede adoptar el transcriptoma de un organismo. Los valores que pueden abarcar el vector *s* pueden ser muy variados, pero existen restricciones. Por ejemplo, si tenemos un genoma de 5 genes que pueden tener diferentes niveles de expresión *s*, en este ejemplo discretos (alto, medio, bajo), tendremos 3^5 combinaciones de estados posibles de este grupo de genes. Si pudiéramos obtener un gran número de mediciones de estos genes en un sistema biológico, es esperable que algunas combinaciones de estados sean más frecuentes y otras no se observen nunca. No todas las combinaciones son observables. Por ejemplo, los genes de un genoma no se expresan todos en su máximo o mínimo nivel en un

mismo momento. De esta manera, podemos determinar una probabilidad de ocurrencia de cada estado. En un organismo real, los valores de expresión de los genes son continuos y el número de genes generalmente es cercano a 28,000, por lo que existe un número muy grande de estados y restricciones posibles. De manera de conocer la máxima capacidad de estados diferentes y así aprender de sus restricciones, se necesita recopilar y obtener información de cómo se comportan los genes de un organismo en condiciones que abarquen un gran número de estados, por lo que en este trabajo se propone lo que se describe a continuación que está diagramado en la Figura 1.

7.2. Recopilación de estados y su distribución

Para conocer los diferentes estados del transcriptoma de *Arabidopsis thaliana*, se obtuvieron las hibridaciones del repositorio de NASCArrays (Craigon et al., 2004), el cual contiene 249 series experimentales, incluyendo 4,815 hibridaciones basadas en la plataforma ATH1 de Affymetrix. Estos experimentos fueron clasificados según el factor experimental evaluado y el tejido de procedencia, mostrando que existe una amplia gama de condiciones y tejidos estudiados (Figura 2 A y B). Dentro de los factores experimentales analizados, el estudio de la función de genes y de cambios en las condiciones abióticas y bióticas son los principales motivos de estudio que se han realizado en este organismo. En cuanto a los tejidos evaluados, predomina el análisis de las partes aéreas de las plantas (Figura 2B). Todo lo antes expuesto otorga un amplio panorama de la expresión génica de Arabidopsis.

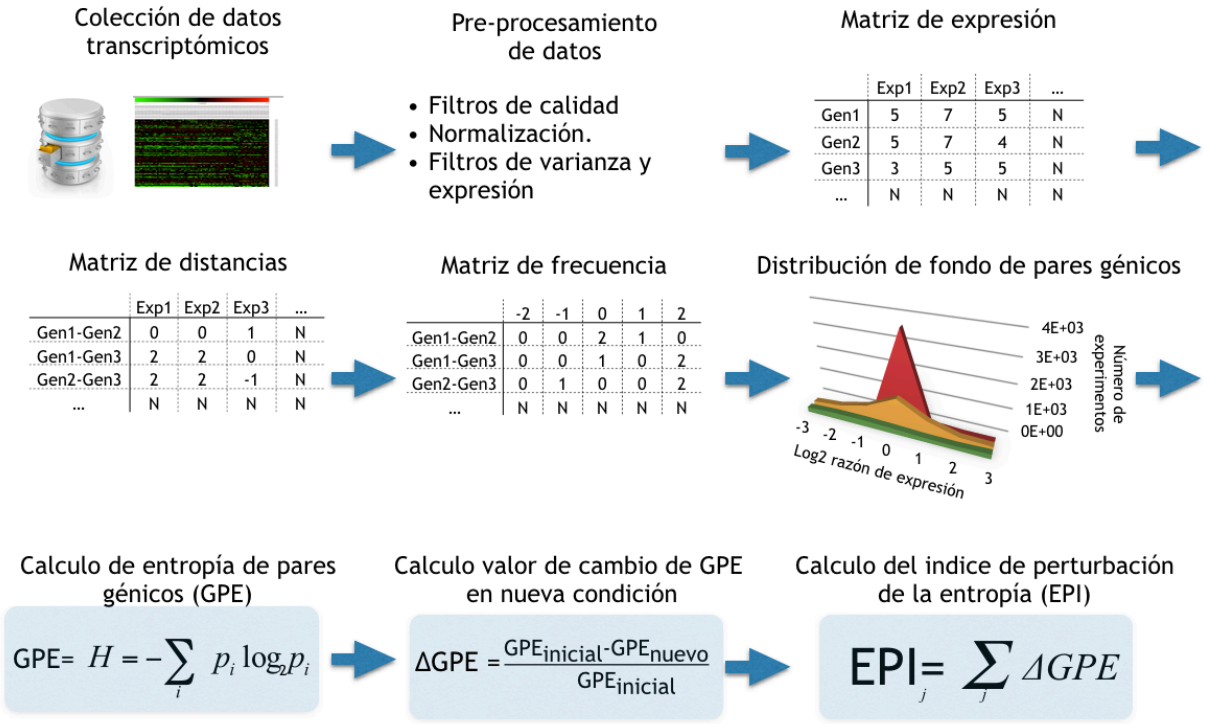


Figura 1. Propuesta para capturar información de las interacciones y/o restricciones entre pares de genes.

Se muestra el método propuesto para el desarrollo de este trabajo. Ver texto para más detalles

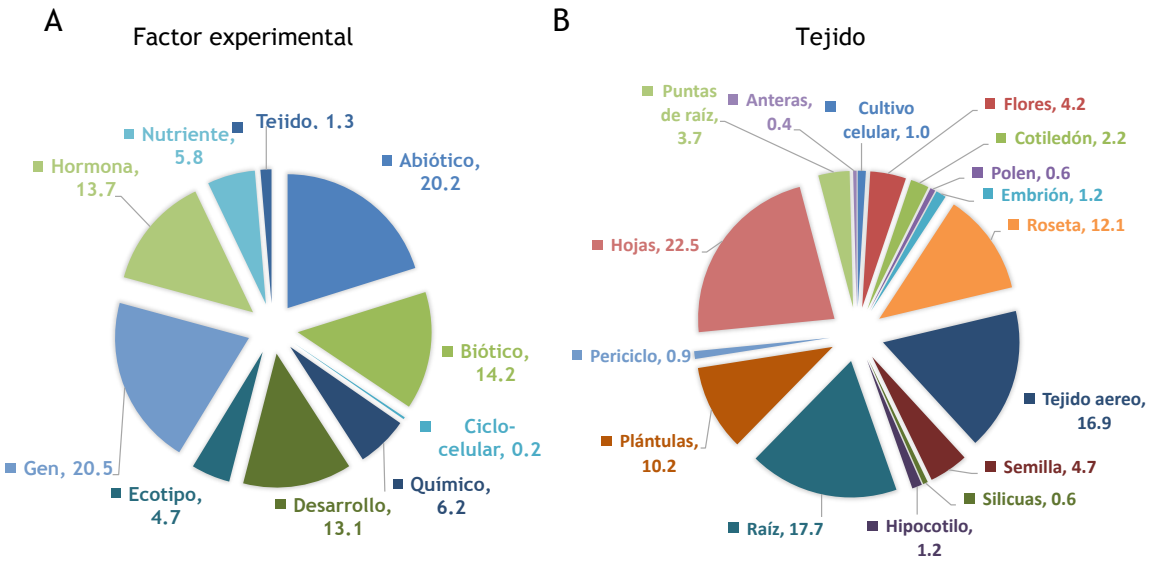


Figura 2. Clasificación de experimentos utilizados para transcriptoma de Arabidopsis.

Clasificación del factor experimental (A) y el tejido de procedencia (B) de la colección de transcriptomas descargados de la base de datos Nascarrays.

7.3. Pre-procesamiento de los datos.

Una vez obtenidos los datos se necesita leerlos, limpiarlos y normalizarlos. Los archivos de datos crudos ".CEL" se leyeron utilizando R (R Core Team, 2015), específicamente con el paquete de Bioconductor Affy (Gautier et al., 2004). Lo primero que se realizó fue descartar hibridaciones de baja calidad debido a que aportarían ruido al análisis. Este paso se realizó utilizando el algoritmo IQRray (Rosikiewicz y Robinson-rechavi, 2011), el cual utiliza el valor de intensidad de cada sonda y lo transforma en un ranking; como existe un grupo de sondas que miden un mismo fragmento de gen, si el ranking de intensidad de cada miembro de este grupo es similar, el valor de calidad entregado por IQRray es mayor que si son diferentes. Este cálculo se realiza para todos los conjuntos de sondas medidas y se obtiene una distribución de calidades de las hibridaciones. Con esta distribución, se eliminó el 5% de las hibridaciones con más baja calidad. Luego de este filtro, se eliminaron las series experimentales que tuviesen menos de 4 hibridaciones con calidad sobre el umbral. Se eliminaron hibridaciones duplicadas, es decir que poseen nombres diferentes, pero la correlación con otra hibridación es igual a 1, ya que se sobreestimarían las relaciones en una misma condición. Esto se debe a que algunos investigadores suben la misma información en diferentes series experimentales, por lo que se mantuvo solo uno de estos elementos. Los datos filtrados se normalizaron utilizando el método Robust Multiarray Analysis (RMA) (Irizarry et al., 2003) del paquete *affy* de Bioconductor. Un segundo filtro a nivel de sondas se realizó, eliminando conjuntos de sondas de asignación ambigua al gen correspondiente (por ejemplo, sondas asignadas a múltiples genes). Si un gen es medido por más de un conjunto de sondas, se utilizó el promedio de las sondas involucradas. Adicionalmente, se descartaron sondas de acuerdo con los siguientes filtros: niveles de

expresión del gen con una desviación estándar menor a 0.25, intensidad de expresión menor a 6 y con varianza de expresión menor a 0.5 y genes correspondientes al 1% más expresado, de manera de evitar sondas con mucho ruido, que no cambien o presenten saturación.

7.4. Análisis global de la entropía de los pares génicos.

De manera de determinar las restricciones en expresión génica que existen en los organismos es necesario tener una idea de las relaciones que existen entre los genes que conforman estos organismos. En este paso adaptaremos un método que combina la relación de la expresión de los genes y la libertad o restricción entre ellos. Luego, del pre-procesamiento de los datos se obtiene una matriz de expresión en logaritmo base 2 de todos los genes en las diferentes condiciones experimentales que describimos anteriormente (Figura 1). A partir de esta matriz, se calcula la diferencia entre todos los genes en una misma condición, es decir, la razón de expresión entre todos los pares de genes y esto se repite para cada experimento presente en la base de datos. Luego de obtener la matriz, se hace una aproximación al entero más cercano y se hace el conteo del número de ocurrencias obteniendo una distribución de la razón discretizada. A esto le llamamos “distribución de fondo de pares génicos” (Figura 1), la cual se utiliza para conocer posibles limitaciones y grados de incertidumbre en el comportamiento de los genes. Para definir este comportamiento de forma objetiva se calcula la entropía de Shannon de la distribución de razones de cada par de genes, a la cual llamamos entropía de pares génicos (GPE) (Figura 1).

La entropía de Shannon es la sumatoria negativa del logaritmo de la probabilidad ponderada de la ocurrencia de cada razón de expresión (Ver Métodos). La entropía de Shannon entrega una

noción acerca del comportamiento de una distribución, en este caso, de la distribución de razones de un par de genes. Mientras menor es el valor de entropía, menor es la libertad de cambios de expresión que se observa para ese par.

7.5. Distribución de valores de entropía de pares génicos.

Los análisis fueron utilizados para evaluar el comportamiento del transcriptoma de *Arabidopsis thaliana.* De manera de establecer si la estrategia podía ser utilizada en otros organismos, se realizó el mismo procedimiento para *Saccharomyces cerevisiae* (Ver métodos para más detalles). Para conocer cómo se comportan los valores de cada GPE de forma global, se graficó la distribución de GPE todos los pares de Arabidopsis y levadura como se muestra en la Figura 3 (A y B). Los valores de GPE siguen distribuciones similares en el caso de Arabidopsis o levadura, con pocos genes que tienen una marcada alta o baja entropía, asemejándose a una curva normal. En Arabidopsis el par de genes que posee una menor entropía es *40S RIBOSOMAL PROTEIN S2* (AT4G33865) y el *RIBOSOMAL PROTEIN L7Ae/L30e/S12e/Gadd45* (AT3G18740). Como se muestra en la Figura 3C, la razón de expresión para esos genes en la mayoría de los experimentos es representado por una distribución que presenta un máximo estrecho y elevado, indicando que la razón de expresión entre estos dos genes se mantiene en la mayoría de las condiciones experimentales evaluadas. Así, genes de bajo GPE son indicativos de una alta restricción de expresión entre un par de genes. En este caso, es consistente el hecho de que este par de genes codifica para proteínas que

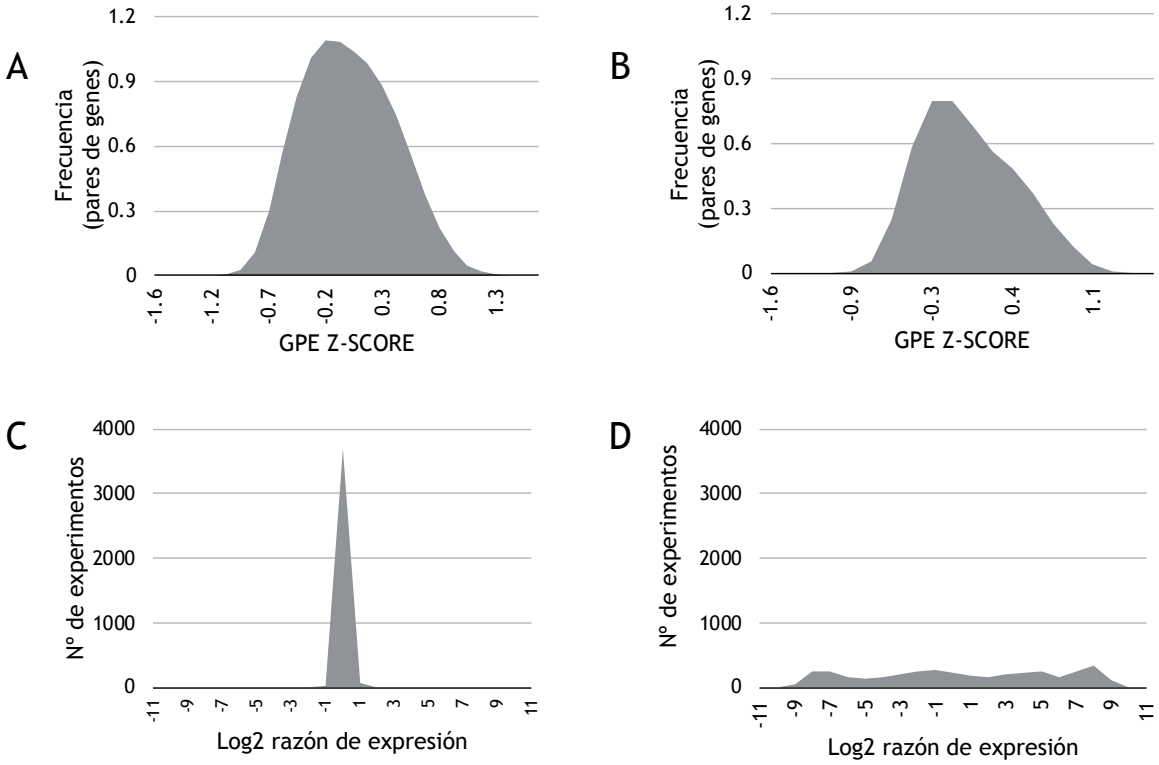


Figura 3. Distribución de valores de entropía de pares de genes siguen una curva de distribución normal en Arabidopsis y levadura.

Entropía de pares de genes(GPE) fueron calculados para todas las combinaciones de genes de *Arabidopsis thaliana* (A) y *Saccharomyces cerevisiae* (B) y graficadas frente al número de pares ge genes. C y D muestran casos extremos de distribuciones de bajo (C) y alto (D) GPE de los datos de Arabidopsis.

forman parte de un mismo complejo, el ribosoma. En contraste, la distancia de expresión de un par de genes con un valor alto de GPE, como los genes *CALMODULIN-LIKE PROTEIN* 10 (AT2G41090) y *BETA GLUCOSIDASE 23 (*AT3G09260) presentan una distribución de razones más plana (Figura 3D), indicando que el nivel de expresión de este par de genes de alta GPE fluctúa casi sin restricciones o de forma independiente. El área bajo la curva en la Figura 3C y 3D es igual ya que es el mismo número de observaciones.

7.6. Medidas de GPE capturan información adicional de dependencia de expresión génica entre pares de genes.

Genes que codifican proteínas que son parte de una misma vía o que forman parte de un mismo complejo son a menudo co-regulados (Li, 2002; Childs et al., 2011; Tegge et al., 2012). Por esta razón, la correlación de expresión frente a un gran número de condiciones son a menudo usados como un método para señalar genes asociados a una función común (Tegge et al., 2012). El valor de información mutua es otro de los métodos más utilizados para determinar dependencia entre genes. Una alta información mutua refleja una asociación no-azarosa entre los miembros del par, por lo tanto, se ha utilizado para determinar relaciones biológicas entre genes. De manera de comparar nuestra aproximación con otros métodos ampliamente utilizados para inferir relaciones funcionales entre genes, se seleccionaron pares de genes que tenían un valor de GPE menor a 3 desviaciones estándar bajo la media del valor de entropía total, es decir, los de menor entropía. En el caso de Arabidopsis, son 8118 pares y para S. cerevisiae 2205 pares. Debido a que estos pares de genes mantienen relaciones sin mucha variación bajo distintas condiciones experimentales, podrían indicar que son genes esenciales para el organismo. Consistente con esta observación, encontramos que la lista de genes contenidos en los 8118

pares de Arabidopsis y los 2205 pares de S. cerevisiae está enriquecida en genes presentes en el grupo de “CEGMA core genes”. Los “CEGMA core genes” codifican para proteínas que están altamente conservadas en un amplio rango de eucariotas, los cuales se asumen como esenciales para las funciones básicas de estos organismos (Parra et al., 2007). La lista de pares con valores bajos de entropía señalada anteriormente se comparó con una lista de igual número de pares de genes de más alta correlación (correlación > 0.93) y de más alta información mutua, calculadas a partir del mismo set de datos. Como se muestra en la Figura 4, cada método es capaz de capturar un conjunto de pares diferentes. Sin embargo, hay una importante intersección entre pares de genes obtenidos usando correlación e Información mutua tanto en Arabidopsis (Figura 4A) como en levadura (Figura 4B) (60.8 % y 60.2% de pares de alta correlación e información mutua son compartidos en Arabidopsis y levadura respectivamente). De manera interesante, la intersección entre pares de bajo GPE y pares obtenidos por los otros dos métodos es muy limitado (0.073 % y 0% de pares de bajo GPE son compartidos con pares de alta información mutua para Arabidopsis y levadura y 1.3% and 6.8% de pares de bajo GPE son también pares de alta correlación en Arabidopsis y levadura respectivamente), indicando que el cálculo de GPE podría capturar información complementaria acerca de relaciones entre genes comparado con otros métodos utilizados comúnmente. Como dijimos anteriormente, un conjunto de datos de 2 variables independientes puede tener muchas interpretaciones y resultados según el método que se utilice para analizarlos. Como en nuestro caso estamos capturando información diferente con los métodos utilizados, calcularemos las 3 medidas utilizando los datos del cuarteto de Anscombe (Hanna et al., 2010; Nielsen, 2016). En estos conjuntos de datos, visualmente se puede observar que son comportamientos muy diferentes (Figura 5). Podemos observar que los valores de correlación de los 4 set de datos son iguales, mientras que el valor de información

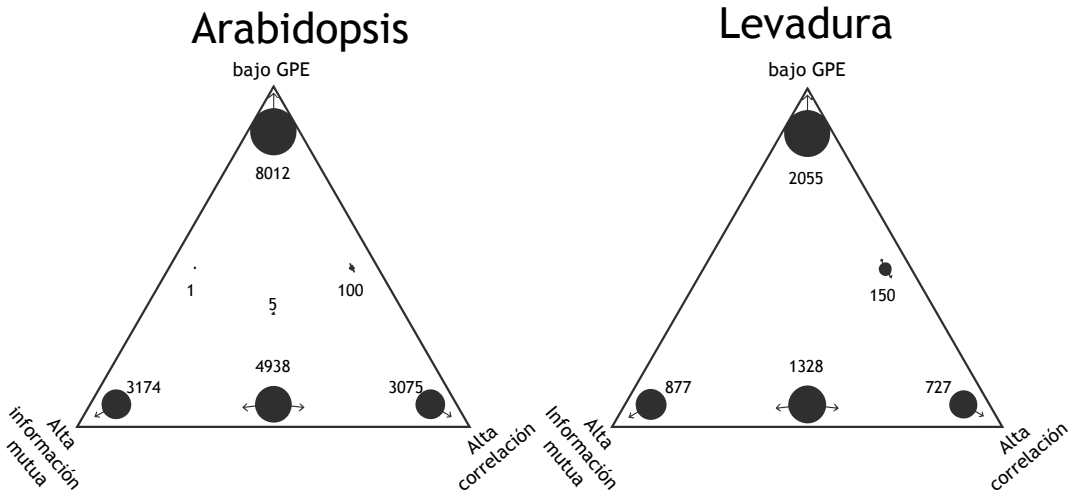
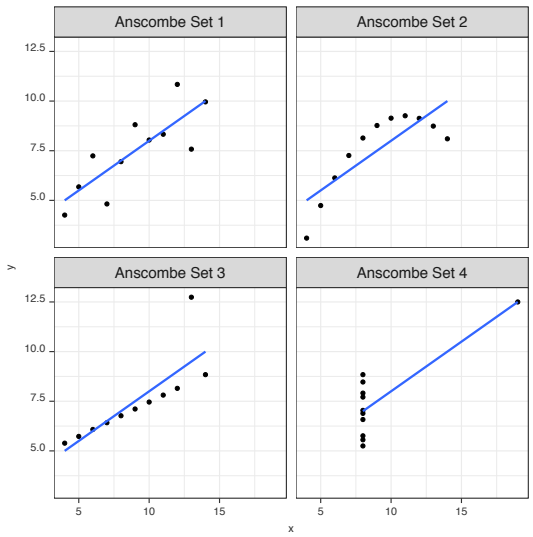


Figura 4. Valores de bajo GPE capturan información diferente comparada con los valores de alta correlación e información mutua.

Pares de genes con bajo GPE (3 desviaciones estándar bajo la media) fueron usados para comparar con el mismo número de pares con mejor correlación e información mutua. El polígono de Sungear muestra el nombre de los 3 conjuntos de datos en los vértices. Los círculos dentro del polígono representan las intersecciones entre los grupos. A) representa datos de Arabidopsis y B) datos de levadura.



Set	Media(x)	Desviación estándar(x)	Media(y)	Desviación estándar(y)	cor(x,y)	Cor.p.val	Información mutua	Entropía	Rank Cor	Rank IM	Rank entropía
1	9	3.316625	7.500909	2.03	0.816	0.00217	1.972247	0.9743148	1	1	1
2	9	3.316625	7.500909	2.03	0.816	0.002179	1.540306	1.153742	1	3	3
3	9	3.316625	7.500909	2.03	0.816	0.002176	1.672625	1.011404	1	2	2
4	9	3.316625	7.500909	2.03	0.816	0.002165	0.3046361	1.277034	1	4	4

Figura 5. Análisis de datos del cuarteto de Anscombe.

Datos del cuarteto de Anscombe, se analizaron con las diferentes métricas utilizadas en este trabajo. Arriba se puede observar visualmente los diferentes conjuntos de datos y abajo la tabla con los valores de la media y desviación estándar de las 2 variables y la correlación (cor), información mutua (IM), entropía y el orden del valor para cada conjunto de datos.

mutua y entropía cambian de acuerdo con el set de datos. Si bien el orden se mantiene entre la información mutua y la entropía, el rango de cambio es diferente.

7.7. Pares de baja entropía capturan información diferente que correlación e información mutua.

De manera de determinar si es que la información de dependencia entre genes obtenida mediante GPE tiene alguna significancia biológica, se buscó información acerca de interacciones reportadas entre los pares de genes en bases de datos públicas de interacción génica. Para Arabidopsis se utilizó la información contenida en la base de datos ANAP (Wang et al., 2012), la cual contiene 219,754 interacciones físicas o inferidas por curador y para levadura se utilizó la base de datos BioGRID (Stark et al., 2006; Chatr-aryamontri et al., 2015) la cual contiene 228,442 interacciones no redundantes experimentalmente identificadas. Se encontró que, con los tres métodos, GPE, correlación e información mutua, se pueden identificar pares de genes que tienen algún tipo de información de interacción biológica. El número de pares de genes para los cuales existe información es además mayor a lo esperado por azar, lo cual es consistente con la utilización de estos métodos para la inferencia de relaciones funcionales entre genes (Tabla 1). Más aún, el número de pares GPE validados es mejor o al menos similar a los obtenidos por correlación o información mutua. En Arabidopsis, de los pares de genes evaluados en que la interacción es apoyada por datos de interacción biológica, 76% (157 pares) son encontrados por bajo GPE, un 6% (33 pares) y 16% (43 pares) por alta correlación e información mutua respectivamente (Figura 6A y Tabla 1). Para levadura, se encontró un número similar de pares validados para bajo GPE, correlación e información mutua, 49% (594 pares) de bajo GPE, 49% (597 pares) de alta correlación y 33% (400 pares) de alta información mutua) (Figura 6B y Tabla1B).

A

método de detección	bajo GPE	Alta correlación	Alta información mutua	Media azar	Azar DE	Total pares en Base de datos
Inferido por curador	125	23	31	0.9	0.93	23781
Bioquímico	48	8	9	0.07	0.27	1924
Electroforésis en condiciones nativas	19	0	0	0.04	0.21	1401
Matriz doble híbrido	3	1	1	0.22	0.48	5590
Cromatografía por afinidad	2	2	2	0.06	0.24	1576
Co-fraccionamiento	1	0	0	0	0	12
Co-sedimentación por densidad	1	0	0	0	0.03	17
Purificación por afinidad	1	0	0	0.04	0.21	1045
Western blot	1	0	0	0	0.04	79
Numero de pares diferentes	157	33	43	-	-	-

B

método de detección	bajo GPE	Alta correlación	Alta información mutua	Media azar	Azar DE	Total pares en Base de datos
Captura por afinidad MS	549	557	372	5.91	2.49	41948
Captura por afinidad-Western	54	70	71	1.21	1.1	8607
Genética positiva	37	12	11	2.87	1.72	20633
Co-purificación	21	18	16	0.6	0.77	4207
Doble híbrido	20	24	16	0.64	0.8	5162
Letalidad sintética	19	15	5	0	0.03	22
Reconstrucción de complejos	17	8	7	0.6	0.8	4340
PCA	12	18	11	0.78	0.9	5795
Co-fraccionamiento	7	4	4	0.11	0.31	788
Co-cristalización	6	3	4	0.07	0.25	435
Rescate por dosis	6	6	7	0.54	0.73	4224
Captura por afinidad-RNA	3	0	0	1.27	1.12	9341
Mejora fenotípica	3	3	5	0.71	0.84	5219
Supresión fenotípica	3	0	0	0.59	0.79	4501
Número de pares diferentes	594	597	400	-	-	-

Tabla 1. Pares de baja entropía capturan información biológica relevante.

Pares seleccionados con bajo GPE, alta correlación y alta información mutua, fueron consultados en la base de datos de ANAP para Arabidopsis (A) y BioGRID para levadura (B). En esta tabla se muestra el número de pares que poseen algún tipo de validación ya sea física, genética o inferida por el curador. También se muestra la media y la desviación estándar para un igual número de pares obtenidos al azar y el número total de interacciones únicas presentes en las bases de datos. El número de diferentes interacciones validadas se muestra en la última fila.

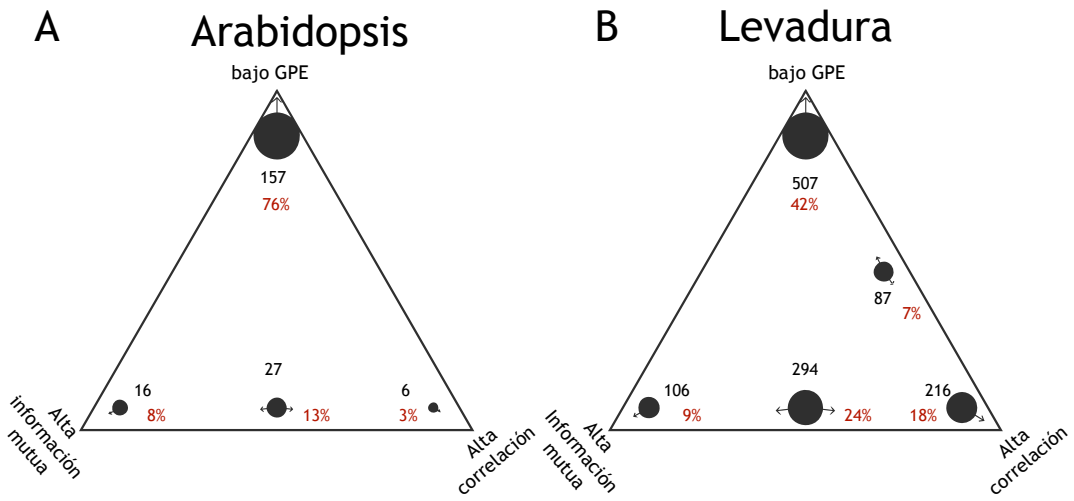


Figura 6. Pares de bajo GPE capturan interacciones biológicamente relevantes.

Pares seleccionados con bajo GPE, alta correlación y alta información mutua, fueron consultados en la base de datos de ANAP para Arabidopsis (A) y BioGRID para levadura (B). Se muestra el número y porcentaje (rojo) de pares que poseen algún tipo de validación ya sea física, genética o inferida por el curador.

Con el fin de evaluar si el número de pares seleccionados para la evaluación (Pares de genes con entropía menor a 3 desviaciones estándar bajo la media) es representativo, se utilizó el mismo criterio con los datos de información mutua y correlación, encontrando resultados similares a los mostrados. Esto indica que valores de bajo GPE pueden ser tan buenos en señalar interacciones biológicas como los ampliamente utilizados de alta correlación e información mutua en ambos organismos.

7.8. Pares de baja entropía capturan información biológica.

De manera de identificar procesos asociados con pares de bajo GPE, se generó una red representando a cada gen presente en un par, como un nodo conectado por una línea al otro miembro del par (Figura 7). Esta red fue agrupada por topología para encontrar subredes de genes altamente conectados utilizando el algoritmo Antipole (Ferro et al., 2006). A estas subredes se les realizó un análisis de enriquecimiento de procesos biológicos, utilizando información de la base de datos de "Gene Ontology". Se encontró que la mayoría de los procesos biológicos sobrerrepresentados corresponden a la maquinaria basal de la célula. Algunos de los cuales, como "ribosome biogénesis" o "proteasome assembly", son conocidos por formar complejos multiproteicos. Esto sugiere que pares de genes dentro de procesos acotados pueden poseer valores de GPE promedio menor que pares al azar.

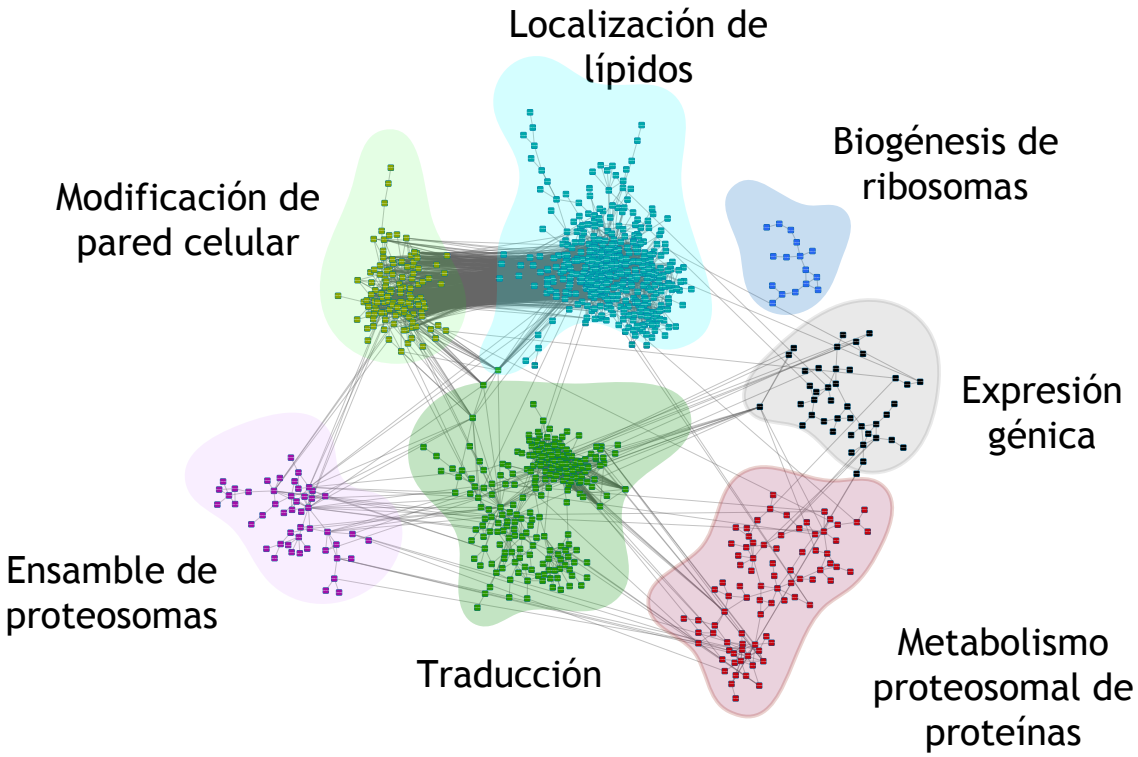


Figura 7. Redes de pares de bajo GPE son enriquecidos en genes conservados y procesos biológicos centrales.

Pares de genes con bajo GPE (menor a 3 desviaciones estándar bajo la media) fueron usados para construir una red. Esta red está enriquecida en genes de CEGMA ($p< 1*10^{-16}$). Análisis de grupos muestra Análisis de clúster con Antipole muestran procesos enriquecidos en cada clúster. Cada clúster fue nombrado un término representativo.

7.9. Distribución de la entropía de pares génicos en procesos biológicos.

Para evaluar la hipótesis, de que pares de genes dentro de procesos definidos tienen más restricciones que pares de genes cualquiera se realizó la siguiente operación: Se calculó la GPE entre pares de genes pertenecientes a un mismo proceso biológico o vía metabólica (definidos según la base de datos KEGG) y entre pares de genes pertenecientes a distintos procesos biológicos. Se encontró que pares de genes que participan en un mismo proceso biológico tienen una menor GPE promedio que pares que participan en distintos procesos (Figura 8). Esto sugiere que debe existir una expresión coordinada entre genes pertenecientes a un mismo proceso biológico o vía metabólica. Para determinar cómo se distribuyen los valores de GPE entre los diferentes procesos biológicos, se calculó la media de valores GPE de todos los pares posibles dentro de cada proceso biológico anotado en la base de datos KEGG (Kanehisa, 2002). Los resultados mostrados en la Figura 9, muestran cómo se distribuyen los valores de GPE en los diferentes procesos biológicos. Los procesos que están entre los de más bajo GPE del total de la distribución están asociados al funcionamiento básico de la célula, los cuales en general están compuestos por genes que forman complejos proteicos, incluyendo el proteosoma, la maquinaria de transcripción basal, biosíntesis de glicanos, ribosomas entre otros (Figura 9). Al contrario, procesos con valor alto de GPE, es decir, pares de genes que fluctúan con más flexibilidad, están involucrados en vías metabólicas relacionadas con condiciones específicas de la célula. Estas últimas incluyen la vía de un precursor del ácido jasmónico, el metabolismo del nitrógeno, metabolismo del metano, entre otros (Figura 9). Resultados similares se observan en levadura. Esto nos sugiere que genes participantes de funciones básicas tienen un nivel más alto de restricciones internas en cuanto a su expresión que genes participantes de procesos que son más dependientes de estímulos internos o externos. Los valores promedios mostrados en la

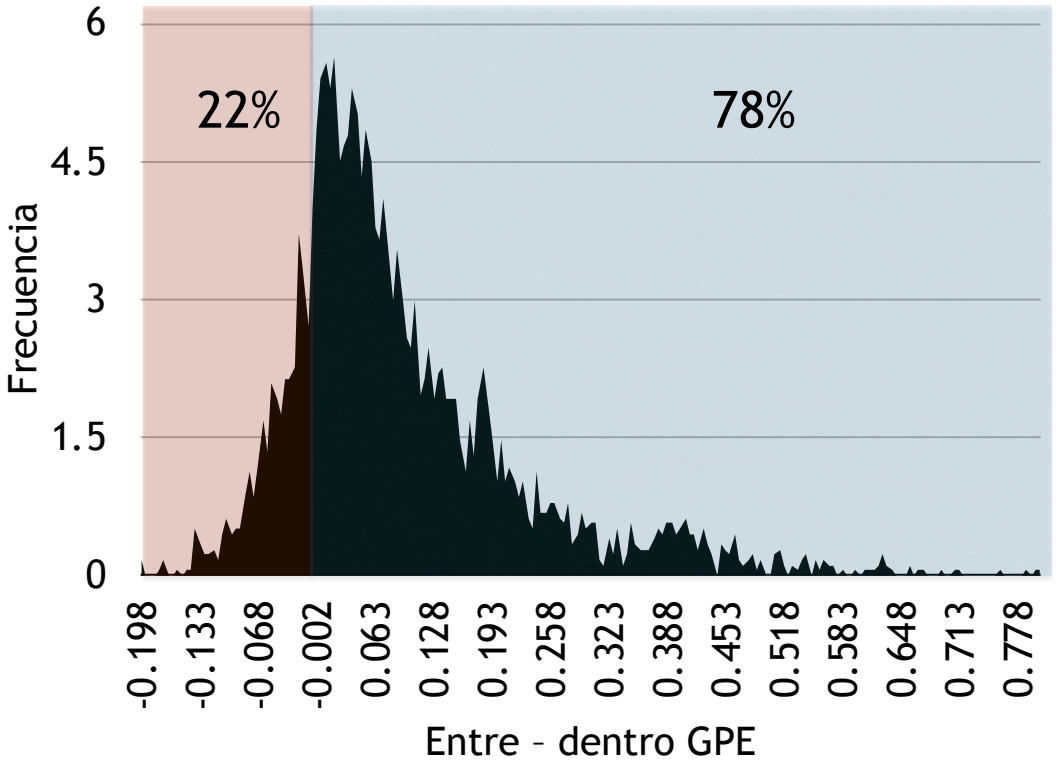
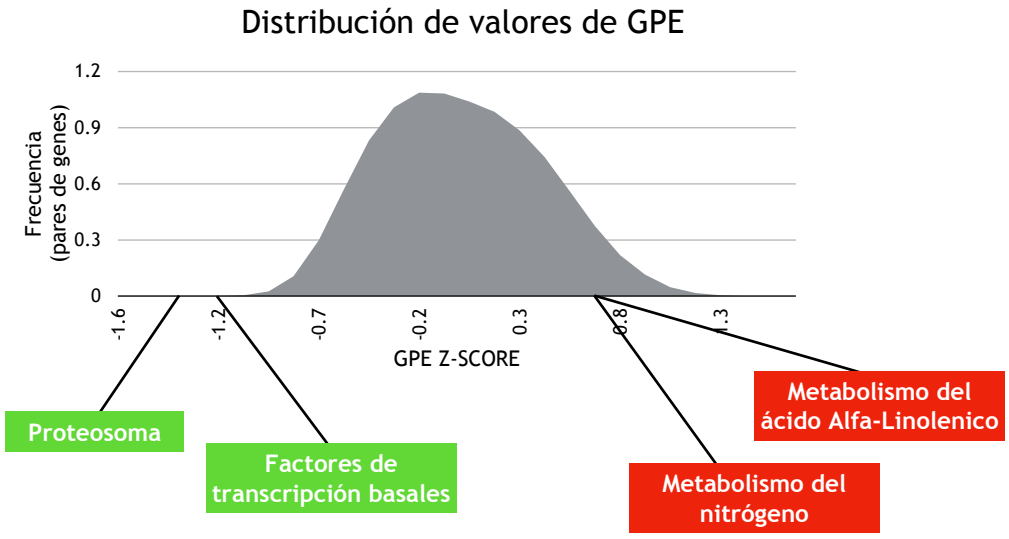


Figura 8. Diferencia de valores de pares Entre-Dentro de proceso biológicos revelan que el comportamiento en genes dentro de un proceso es restringido.

Histograma representa la diferencia entre los valores medios de todos los pares dentro de un proceso KEGG y la media ponderada de pares fuera del proceso para cada comparación. Colores representan valores negativos (rojos) o positivos (azules).



Nombre	Percentil	Número de pares
Proteosoma	0	1653
Factores de transcripción básales	3	528
Regulación de la autofagia	5	45
Biosíntesis de N-Glicanos	5	435
Ribosomas	5	23005
Rna Polimerasas	8	300
Reparación de errores	8	351
Spliceosoma	10	5671

Nombre	Percentil	Número de pares
Biosíntesis de glucosinolatos	78	190
Metabolismo del ácido cianoamino	78	253
Metabolismo de xenobioticos	78	210
Biosíntesis de fenilpropanoides	81	5356
Metabolismo del metano	82	3321
Metabolismo de fenilalanina	82	3160
Metabolismo del nitrógeno	84	861
Metabolismo del ácido Alfa-Linolenico	84	378

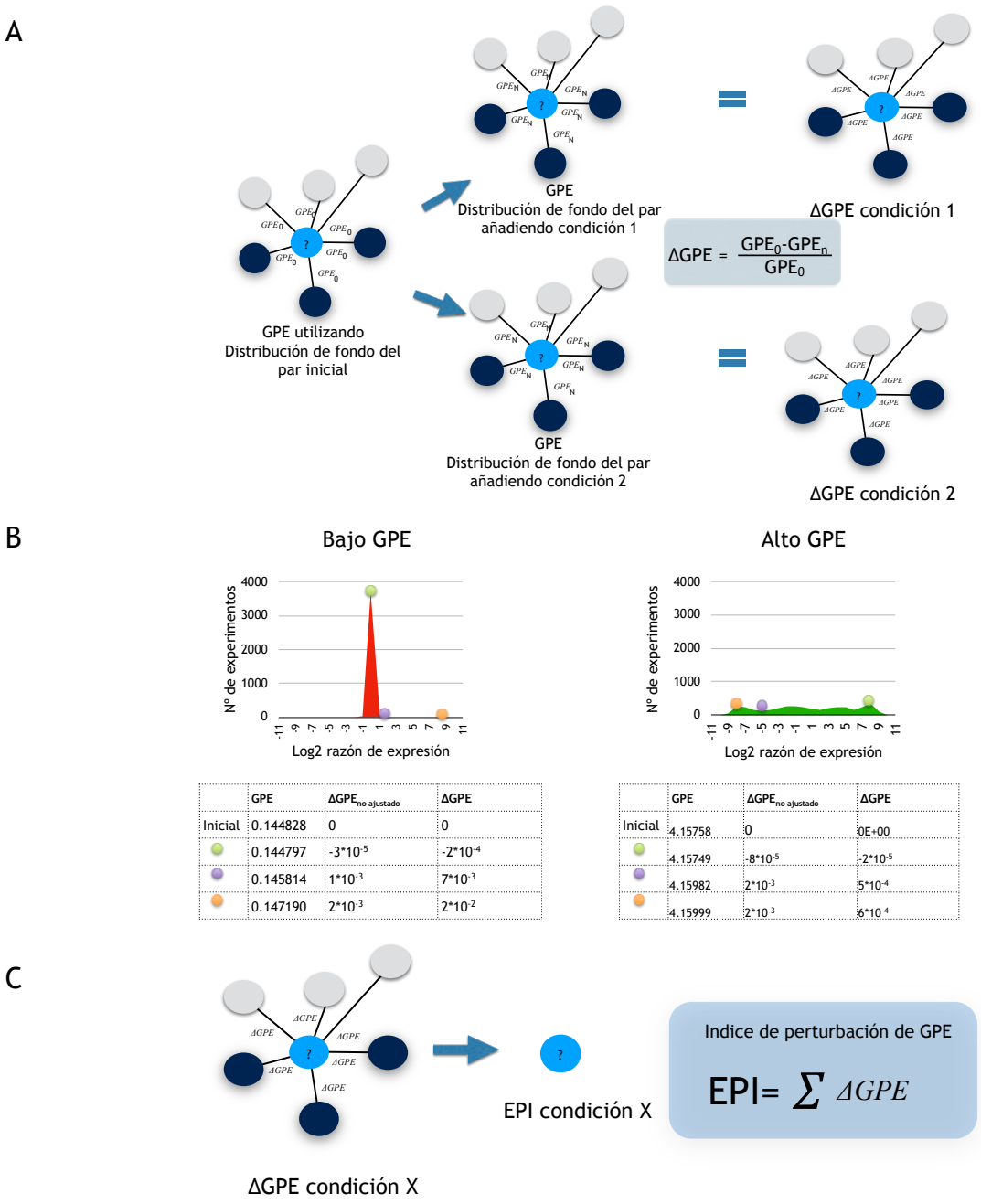
Figura 9. Procesos relacionados al funcionamiento basal de la célula están compuestos de pares de genes de bajo GPE, mientras que procesos relacionados a respuestas del organismo muestran alto GPE.

Se determinaron las medias de valores de GPE de todos los pares de genes presentes en un determinado proceso de la base de datos KEGG. Se muestra la curva global de los valores GPE, los 2 procesos de más bajo valor (verde) y más alto valor (rojo) GPE promedio (verde) y se indica donde se posicionan en la curva de entropía global. En la tabla verde y roja se muestran los procesos biológicos que poseen más bajo y alto valor GPE promedio respectivamente con el percentil en que se ubican en la curva y el número de pares posibles en cada proceso. Los procesos mostrados en las tablas verde y roja poseen un valor promedio significativamente menor y mayor (p empírico <0.01) respectivamente que lo encontrado por azar.

Figura 9 son significativos comparando los valores promedio GPE obtenidos con 10000 conjunto de genes considerando el número de genes del respectivo proceso.

7.10. Definiendo un valor de cambio de la entropía del par génico.

Dada su naturaleza sésil, las plantas son altamente dependientes de sus posibilidades de adaptarse a diferentes estreses ambientales (Hahn et al., 2013). La inducción o represión de la expresión génica en respuesta a cambios en las condiciones ambientales indican que la modulación del transcriptoma es una herramienta clave para su plasticidad. Una perturbación en las condiciones ambientales puede producir cambios en la expresión génica a escala global, lo cual se traduce en cambios en las razones de expresión entre algunos genes del transcriptoma. Al añadir esta nueva razón de expresión a la distribución de fondo (ver Métodos), el valor de GPE de los pares de genes puede cambiar en mayor o menor medida de acuerdo con el comportamiento de este par de genes en el resto de los experimentos. Definimos este cambio respecto al valor de GPE inicial como ΔGPE (Figura 1 y Figura 10). Cada par de genes produce un ΔGPE, por lo que cada gen es responsable de esta variación (Figura 10 A). Como se observa en el ejemplo de la Figura 10 B, con la metodología propuesta si la nueva condición produce una relación que se ha observado muchas veces, el impacto en el valor de GPE va a ser bajo si el par de genes evaluados posee una distribución de fondo de baja entropía. Si el mismo par de genes es evaluado en una condición que provoca un cambio en la entropía que se ha observado poco, el impacto en el valor de GPE de ese par génico va a ser alto. Por otro lado, si la distribución de fondo del par génico tiene una entropía alta, independiente si la nueva observación cambia o no la relación, el cambio en el valor del GPE no será tan drástico.



	GPE	$\Delta GPE_{no\ ajustado}$	ΔGPE
Inicial	0.144828	0	0
	0.144797	$-3*10^{-5}$	$-2*10^{-4}$
	0.145814	$1*10^{-3}$	$7*10^{-3}$
	0.147190	$2*10^{-3}$	$2*10^{-2}$

	GPE	$\Delta GPE_{no\ ajustado}$	ΔGPE
Inicial	4.15758	0	0E+00
	4.15749	$-8*10^{-5}$	$-2*10^{-5}$
	4.15982	$2*10^{-3}$	$5*10^{-4}$
	4.15999	$2*10^{-3}$	$6*10^{-4}$

Figura 10. Descifrando la perturbación del transcriptoma y sus genes involucrados.

En A se muestra la estrategia para cuantificar el cambio del valor de GPE al añadir una nueva condición. En B se muestra como afecta la distribución de fondo al valor de GPE, siendo las bolas de colores posibles nuevas relaciones y la tabla los valores de GPE en la nueva condición. En C se muestra la forma de cuantificar el índice de perturbación de la entropía (EPI) para cada den en una nueva condición.

Con el fin de escalar la información de la relación entre un par de genes a todo el genoma, se propone utilizar la suma de todos los ΔGPE posibles entre los genes del organismo (Figura 1 y 10 C). La suma del ΔGPE de cada gen en el par en que está involucrado lo definimos como Índice de Perturbación de Entropía (Entropy Perturbation Index, EPI). Dado que estímulos ambientales pueden producir cambios en GPE que pueden ser independientes de cambios en la expresión de al menos uno de los genes del par, se razonó que si se determina el EPI diferencial (DEPI, Ver Métodos), se podrían determinar genes que estuviesen involucrados en la respuesta a un estímulo independiente de si cambian o no su expresión.

7.11. Evaluación del valor de perturbación de la entropía (EPI) para identificar genes relevantes.

Una de las más sorprendentes muestras de plasticidad en el desarrollo de las plantas en respuestas al ambiente es la modulación de la arquitectura del sistema radicular (root system architecture, RSA). Cambios en RSA permiten a la planta hacer frente a las variaciones en el suministro de agua y nutrientes, o para interactuar con otros organismos. Aunque se han realizado múltiples esfuerzos para descifrar los reguladores claves del RSA, aún falta mucho por descubrir sobre cómo las raíces perciben y responden a cambios en las señales ambientales. Dentro de los estímulos ambientales más estudiados para cambios en RSA se encuentra la respuesta a nitrato. Nitrato es la mayor fuente de N en los suelos agrícolas y tiene profundos efectos sobre el crecimiento de la raíz (Wang et al., 2004). Genes involucrados en la respuesta a nitrato en raíces han sido determinados mediante metodologías que incluyen genética reversa y directa y biología de sistemas (Gutiérrez, 2012). Para evaluar nuestra metodología, se utilizó

un set de datos correspondientes a un experimento en el cual plantas con déficit de nitrato fueron tratadas con KNO_3 (tratamiento) o KCl (control) por 2 horas, tratamiento que se sabe produce cambios robustos en el transcriptoma de la planta. Se calculó el EPI de cada gen y se comparó tratamiento/control para buscar genes que tuviesen EPI diferencial (DEPI) utilizando la herramienta RankProd (Hong et al., 2006). Adicionalmente, se determinaron genes que tienen expresión diferencial (DE) mediante la misma herramienta (Ver Métodos).

Se obtuvo un total de 594 genes regulados, 462 por DE y 297 por DEPI, con una intersección de 165 genes identificados por ambos métodos (Figura 11). Se realizó un análisis de sobrerrepresentación de procesos biológicos para cada una de estas listas. Dentro de las 3 listas aparecieron sobrerrepresentados los términos "response to nitrate" y "nitrate transport". Los términos "nitrate assimilation" y "nitrate metabolic process" solo aparece en la lista de genes obtenidos por DEPI, lo que indica que ambos métodos son exitosos para identificar genes de respuesta a nitrato y que pueden ser complementarios según la información de la base de datos de Gene Ontology.

Para comprobar que la metodología puede funcionar en otros estímulos, se utilizaron datos publicados del proyecto ATGenExpress (http://arabidopsis.org/info/expression/ATGenExpress.jsp), los cuales incluyen experimentos donde plantas son sometidas a diferentes tipos de stress (frio, osmótico, salino, oxidativo). En estos experimentos, la expresión génica de las plantas es evaluada en raíz a diferentes tiempos de tratamiento (0.5, 1, 3, 6, 12 y 24 h). Se calculó el EPI de cada gen para los datos y se comparó tratamiento/control para identificar genes regulados por DEPI o DE usando RankProd (Hong et al., 2006) al igual que con los datos obtenidos por el tratamiento con nitrato. Como es mostrado en la Figura 11, al igual que con el tratamiento con nitrato, para cada respuesta hay 3

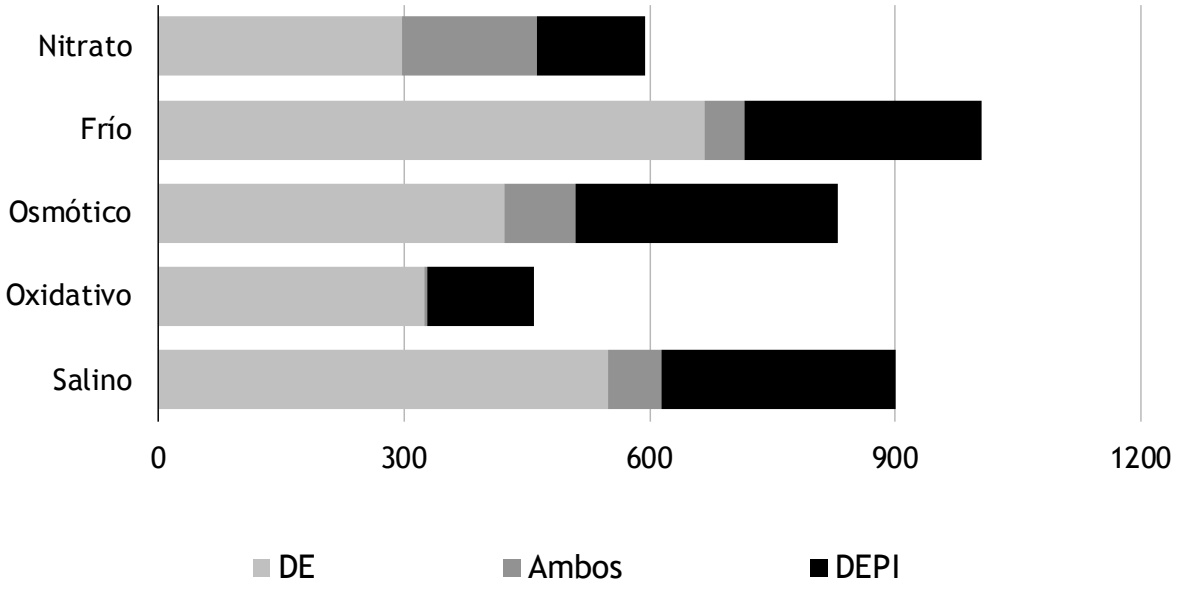


Figura 11. Los genes seleccionados por cada método pueden dar información complementaria respecto al estímulo.

Se muestra en número de genes regulados utilizando RankProd con datos de expresión diferencial DE, índice de perturbación de entropía diferencial (DEPI), o compartidos por ambos métodos (Ambos) lo cual es reportado para cada estímulo evaluado.

grupos de genes. Estos 3 grupos son los genes que solo son identificados por DE, los que solo son identificados por DEPI y los que son identificados por ambos métodos. El número de genes identificados por cada método y el grado de sobrelapamiento entre estas listas cambia de acuerdo con el estímulo (Figura 11), pero en general, el número de genes compartidos por ambos métodos es bajo, lo que podría sugerir que son métodos complementarios.

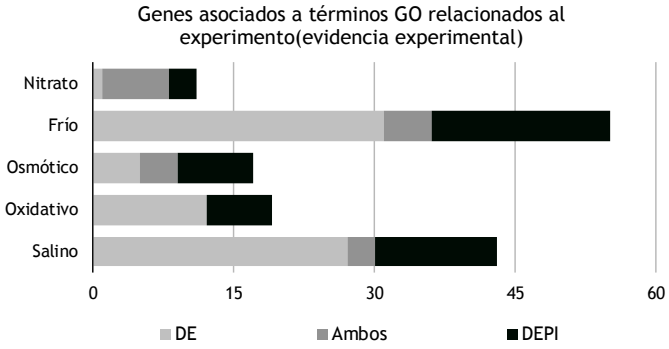
7.12. Relevancia biológica de los genes identificador por el método propuesto.

De manera de evaluar la relevancia biológica de los genes identificados como importantes en la respuesta por el método propuesto, las listas de genes regulados por DEPI o DE fueron comparadas con genes ya anotados a la respectiva respuesta según la base de datos Gene Ontology (Figura 12 y métodos). En la Figura 12B podemos observar que existe un diferente porcentaje de sobrelapamiento dependiendo del estímulo que se utilizó. Esto nos sugiere que hay información que es capturada utilizando el valor de EPI la cual no es capturada por el valor de expresión. Para evaluar si los genes identificados por cada método que están anotados en la base de datos de GO son encontrados o no por azar, se realizó un test hipergeométrico (Figura 12C). Los resultados indican que la identificación de estos términos no es al azar para las diferentes de listas, siendo en algunos casos el valor P más bajo para DEPI y en otros casos para DE, indicando complementariedad de los 2 métodos en la identificación. Basado en esta información y tomando como genes corroborados de la respuesta a lo publicado en Gene Ontology, se calculó la precisión, la exhaustividad y el estadístico F, para los genes DE y DEPI. En la Figura 13 se observa que los resultados son similares para ambos métodos, sugiriendo que ambos métodos son complementarios para comprender la respuesta a un estímulo (Figura 13).

A

Nitrato	Frío	Osmótico	Oxidativo	Salino
Respuesta celular a nitrato	Respuesta celular a frío	Respuesta celular a estrés osmótico	Respuesta celular a estrés oxidativo	Respuesta celular a sal
Asimilación de nitrato	Aclimatación a frío	Regulación de la respuesta a estrés osmótico	Respuesta a estrés oxidativo	Regulación de la respuesta a estrés salino
Transporte de nitrato	Respuesta a frío	Respuesta a estrés osmótico		Respuesta a sal
Respuesta a nitrato				Respuesta a estrés salino

B



C

	DEPI	DE
Nitrato	4E-15	5E-10
Frío	1E-13	1E-14
Osmótico	3E-08	8E-05
Oxidativo	7E-05	3E-05
Salino	2E-06	5E-11

Figura 12. Los genes seleccionados por cada método otorgan información complementaria a cada estímulo.

En A se muestra los principales términos de Gene Ontology asociados a cada estímulo. En B se muestra el número de genes regulados identificados utilizando RankProd con los valores de expresión (DE) y con el índice de perturbación de entropía (DEPI) o por ambos métodos (Ambos) asociados a los términos ontológicos asociados al estímulo con evidencia experimental. C muestra el valor p de sobrerrepresentación del término GO (con evidencia Experimental) para cada estímulo.

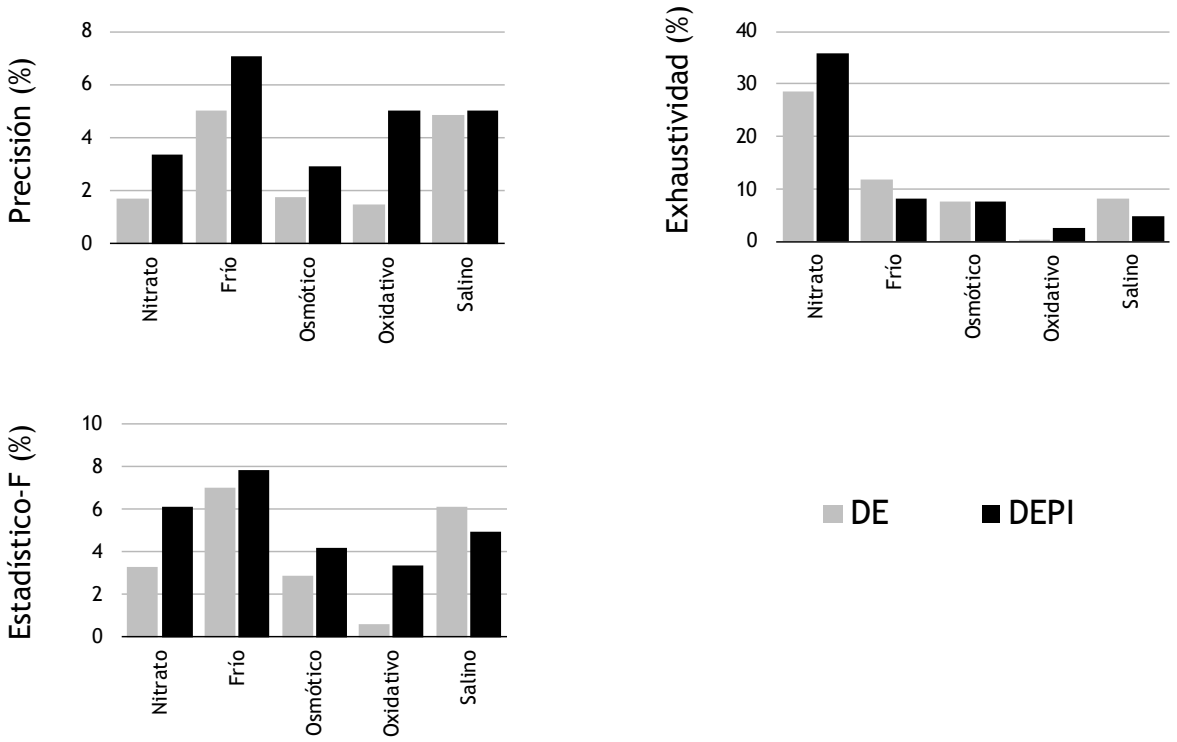


Figura 13. DEPI puede capturar información biológica complementaria a la expresión diferencial.

Para los diferentes estímulos se identifican los genes DEPI y DE. Se calcula la precisión, exhaustividad (recall) y estadístico F, utilizando como verdadero positivo la información entregada por la base de datos Gene Ontology para cada estímulo

Una posible característica de los genes que solo se encuentran diferencialmente regulados en DEPI y no en DE, son los genes que cambian poco su expresión, es decir, que tengan un nivel de restricción alto en su comportamiento. Para conocer cómo se regulan estos genes en diferentes condiciones experimentales, se utilizó la información de las listas de genes regulados del trabajo de Aceituno 2008 (Aceituno et al., 2008) y se seleccionó el 5% de los genes que estaban diferencialmente expresados en un mayor número de condiciones experimentales a los cuales llamaremos "genes multi-respuesta". Comparamos las listas de genes obtenidas en los experimentos de nitrato y de estrés abiótico con los genes multi-respuesta y obtenemos lo que se muestra en la Figura 14, donde podemos observar que los genes identificados por el método DEPI tienen mayor especificidad para el estímulo analizado que los que se encuentran con el método que utiliza solo la expresión. Esto puede sugerir que genes detectados por DEPI, pueden ser genes importantes y más específicos para la respuesta, los cuales no son comúnmente encontrados en los análisis de expresión diferencial.

7.13. Genes seleccionados por ambos métodos pueden ser complementarios para comprender la respuesta al estímulo.

De manera de posicionar los genes obtenidos por DEPI o DE en la respuesta al estímulo evaluado, se construyó una red con el siguiente protocolo: Se seleccionaron los genes obtenidos por ambos métodos y se seleccionaron los que tuviesen alguna anotación relacionada con el estímulo, según Gene Ontology (Figura 12A y métodos). Se buscaron las interacciones factor de transcripción – Blanco de acuerdo a la base de datos DAP-seq (O'Malley et al., 2016) y se generaron las redes para cada estímulo. Para la respuesta a nitrato se construyó la red de acuerdo con las reglas mencionadas anteriormente, quedando una red de 47 nodos y 50 conexiones

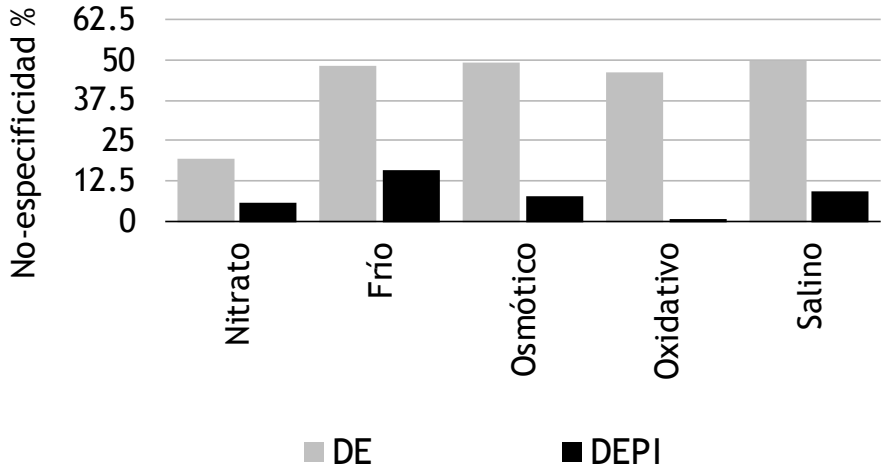
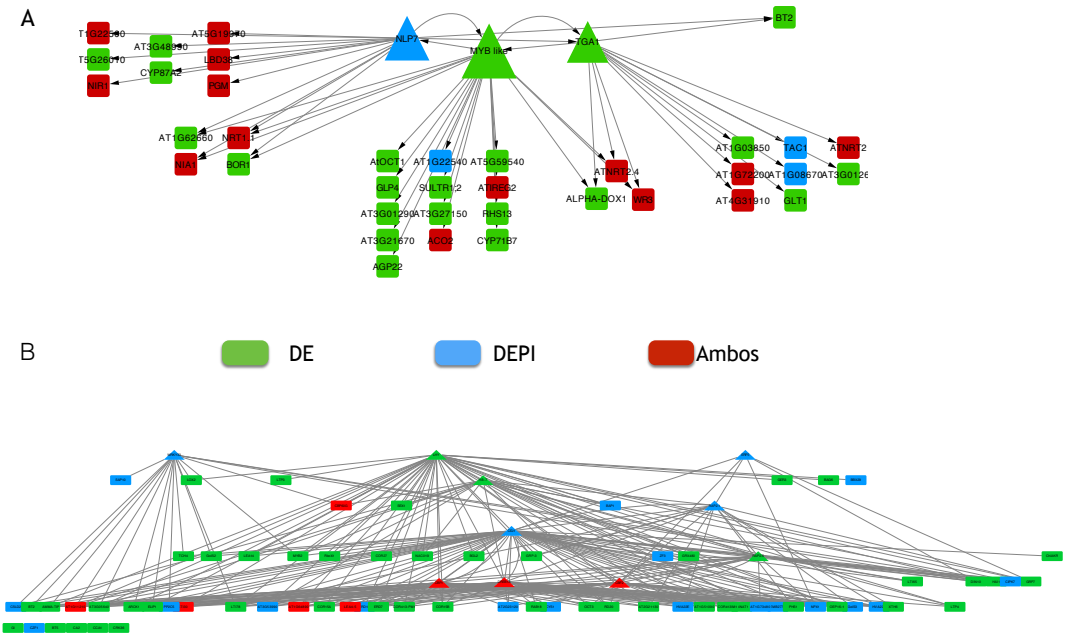


Figura 14. Genes identificados por DEPI tienen menor comportamiento multi-respuesta que los DE.

Comparaciones tratamiento-control usando RankProd con datos de expresión (DE) o con datos de índice de perturbación de entropía (DEPI) en los diferentes estímulos se compararon con el 5% de los datos más respondedores según el trabajo de Aceituno 2008, se muestra el porcentaje de no-especificidad.



A

Figura 15. Red regulatoria generada con genes de respuesta a nitrato y frío.

La red fue generada con los genes de respuesta a nitrato (A) y frío (B) anotados en la base de datos de Gene Ontology y que fueron identificados como importantes por la respuesta con ya sea por expresión (DE, verde) o por el índice de perturbación de entropía (DEPI, azul) o por ambas (Rojo). Las conexiones entre los genes son Factor de transcripción-Blanco presentes en la base de datos DAP-seq. Triángulos representan factores de transcripción.

(Figura 15). De los 47 nodos, 17 están regulados por ambos métodos, 5 sólo por DEPI y 25 sólo por DE. Solo 3 nodos actúan como factores de transcripción en esta red, los que pueden ser los reguladores maestros de la respuesta a nitrato. NLP7, el cual es un gen identificado solamente por el método DEPI en este experimento, el cual modula la mayoría de los genes de la señalización y asimilación de nitrato, y es un factor clave para la respuesta temprana al nutriente (Marchive et al., 2013). Los otros factores de transcripción de esta lista son AT1G25550, un factor de transcripción de la familia G2-like, que se une al promotor de *E2Fa* , el cual es esencial para la iniciación de las raíces laterales (Berckmans et al., 2011; Canales et al., 2014), lo que sugiere que es un importante regulador de la morfología de la raíz en respuesta a la disponibilidad de nitrato (Canales et al., 2014), el cual se encuentra regulado solo por DE. El tercer factor de transcripción de esta lista es TGA1, el cual se ha visto que es un factor importante en la vía de señalización luego de tratamientos con nitrato, afectando el desarrollo de las raíces (Alvarez et al., 2014; Vidal et al., 2015), también se encuentra regulado por DE. Estos resultados nos confirman que ambos métodos pueden ser complementarios para capturar los genes importantes en la respuesta a un estímulo dado.

De forma similar, se hizo el mismo análisis con los otros estímulos abióticos de ATGenexpress. En el caso de respuesta a frío (Figura 15B), podemos observar que se obtienen resultados similares a lo detectado en nitrato, donde se complementa la información obtenida por las diferentes metodologías. Entre los factores de transcripción detectados por DEPI aparece otro importante regulador de procesos en respuesta a stress, el cual está involucrado en la respuesta mediada por ABA, como es el factor de transcripción WRKY33 (Chen et al., 2012). Mutantes de este factor ven afectada la expresión de la familia CBF (Birkenbihl et al., 2012),

que es una de las más importantes en la respuesta a frío (Dong et al., 2011). En este tratamiento, por el lado de los genes solo regulados por DE se encuentra el gen LHY, un importante factor involucrado en el ciclo circadiano, el cual, se ha reportado como involucrado en el control de proteínas de las familias CBF las que en nuestros datos son detectadas como reguladas por ambos métodos, dando otra evidencia que los 2 métodos de detección pueden ser complementarios.

7.14. Señales externas desencadenan cambios en los niveles globales de GPE.

Como mencionamos anteriormente, la regulación positiva o negativa de la expresión génica en respuesta a cambios en las condiciones ambientales perturban el estado del transcriptoma. El valor de GPE de los pares de genes puede cambiar en mayor o menor medida de acuerdo con el comportamiento de este par, es decir a lo que llamamos anteriormente distribución de fondo. La suma del EPI de todos los genes en una condición la llamamos EPI sum, la cual puede servir para seguir de forma global cuan perturbado se encuentra un transcriptoma frente al cambio de la condición experimental.

Para determinar si los cambios ambientales pueden afectar el valor de EPI sum, se utilizaron los datos de tratamientos de stress abióticos del proyecto ATGenExpress. Como describimos anteriormente, incluyen experimentos donde plantas son sometidas a diferentes tipos de stress a diferentes tiempos de tratamiento (0.5, 1, 3, 6, 12 y 24 h). El cambio de valores de GPE fue calculado al introducir la nueva condición a la distribución de fondo, ya sea el tratamiento como el control, para así determinar el efecto que produce la nueva condición en la EPI sum tanto en el tratamiento como en el control (Ver Métodos y Figura 16). Como se muestra en la Figura 16, los estreses abióticos evaluados pueden producir cambios en los valores de EPI sum, tanto en

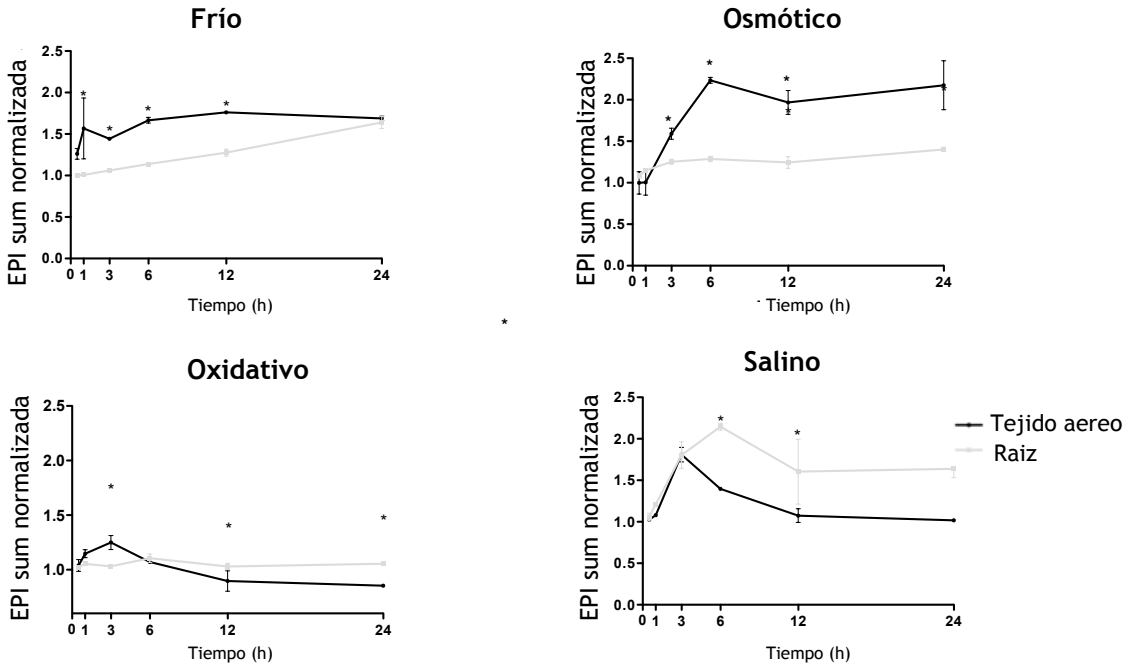


Figura 16. Estímulos ambientales diferentes causan distintos patrones dinámicos de índice de perturbación de entropía global.

Se calculó la sumatoria del índice de perturbación de entropía (EPIsum) en cada tiempo en los diferentes estímulos para raíz y tejido aéreo normalizados por su respectivo control. Diferencias significativas son marcadas con *.

raíz como en tejido aéreo. En el caso de estrés por frío y osmótico, los valores de EPI sum de tejido aéreo es más afectado que en raíces, mientras que, para estrés salino, el EPI sum es más afectado en raíces. Estrés oxidativo afecta igual a raíces que al tejido aéreo. Estos resultados indican que señales abióticas pueden desencadenar cambios en la GPE global y esta es dependiente del órgano, el tratamiento y la duración del tratamiento.

7.15. Evaluación de los genes de respuestas a los diferentes estímulos.

Dado que los resultados obtenidos por los diferentes métodos (DE y DEPI) nos entregan grandes listas de genes, necesitamos conocer relaciones o funciones comunes entre estas listas de genes regulados. Con el fin de conocer de forma global cuales son los procesos biológicos asociados a las diferentes respuestas, se realizó análisis de sobre-representación de proceso biológicos para los diferentes estímulos. Se puede constatar que dependiendo del tratamiento y del método utilizado se identifica un grupo diferente de genes y procesos biológicos sobrerrepresentados. En la mayoría de los casos se observa un número mayor de procesos identificados por DE que por DEPI. Si bien existe sobrelapamiento, muchos procesos son identificados solo por DE o solo por DEPI, lo que sugiere nuevamente que son análisis complementarios.

Para los tratamientos utilizados en este trabajo, es esperable obtener como procesos sobrerrepresentados algunos de los términos GOs mostrados anteriormente en la Figura 12A para el estímulo respectivo. Sin embargo, no aparecen para todos los estímulos estos términos como sobrerrepresentados independiente del método utilizado.

Utilizando como referencia para cada estímulo los genes asociados a los términos GO de la Figura 12A, se cuantificaron los genes que estaban anotados para estos términos lo cual se muestra en la Figura 12B, descrita anteriormente. En todos los casos, se observa que una fracción es detectada por uno o ambos métodos, reafirmando la complementariedad de los métodos. Sin embargo, considerando el total de genes regulados ya sea por DE o por DEPI, la proporción de genes que están asociados a los términos GO más probables del estímulo en estudio es baja. El porcentaje de genes asociados al término respectivo de la Figura 12 A es el siguiente: nitrato 7%, salino 6%, oxidativo 4%, osmótico 2% y frío 7%. Muchos de los genes regulados ya sea por DE o por DEPI que no están anotados como de respuesta al estímulo, están asociados a procesos biológicos que según literatura pueden estar asociados a la respuesta. Sin embargo, hay una gran cantidad de genes en las diferentes listas que no están anotados en procesos biológicos de GO. De manera de obtener una anotación más amplia de la relación de genes con posibles funciones, se utilizó una aproximación por literatura. Al buscar individualmente en literatura la función biológica de un gen y la relación con el estímulo en que se detectó es posible caer en el sesgo de la profundidad en que se busca cada tema. Para Arabidopsis, si bien el porcentaje de anotación es alto (98% considerando RCA (Inferred from Reviewed Computational Analysis) y 31 % sin RCA), solo el 20,8% de los genes anotados tiene evidencia experimental, es decir la mayoría de las anotaciones son obtenidas de forma electrónica.

7.16. Análisis semántico de textos y asociación con genes.

Los resultados anteriores, nos sugieren que, si bien el análisis de enriquecimiento funcional utilizando GO nos ayuda a la interpretación de los datos, es necesario utilizar otros métodos que nos permitan analizar listas de genes con el menor sesgo posible. Para lo cual, en este trabajo utilizaremos una aproximación basada en análisis semántico de textos mediante modelos probabilísticos (Wu et al., 2012). Con esta metodología se puede obtener una recopilación objetiva de la información de cada gen utilizando la información presente en la literatura. Para utilizar esta metodología, se adaptó el método conocido como "Latent Dirichlet Allocation" (LDA) descrito por Blei (Blei et al., 2012), el cual es un método que permite que un conjunto de observaciones puedan ser explicados por grupos inadvertidos (Blei et al., 2012). Para aplicar esta metodología se utilizó la información contenida en la base de datos Araport (Krishnakumar et al., 2015) en la cual se relaciona cada gen con publicaciones (si es que existe asociación en la base de datos) utilizando la herramienta Thalemine (Krishnakumar et al., 2017). Se extraen los resúmenes de cada publicación desde la base de datos de NCBI y se procesa el texto. Se extrajeron 30355 resúmenes de publicaciones que están asociados a algún gene. El procesamiento del texto se hace utilizando el paquete "tm" de R, con el cual se eliminan caracteres no letras y se extrae la raíz de cada palaba, eliminando posibles sufijos o terminaciones. Con este texto ya pre-procesado se utilizó la herramienta desarrollada por Daniel Ramage (2009) http://nlp.stanford.edu/software/tmt/tmt-0.4/ para construir el modelo con el método de LDA. Se ensayó una curva de perplejidad para seleccionar el número óptimo de tópicos para este conjunto de datos, se calculó la perplejidad y la menor perplejidad determinada

fue de 64 tópicos (Figura 17). Sabiendo ya que el optimo matemático para nuestros datos es 64, se añadieron diferentes niveles al análisis para tener información con diferente profundidad. Por lo cual se realizó la determinación de tópicos con diferentes niveles, arbitrariamente en valores de potencia de 2, desde 2 hasta 512, por lo tanto, incluyendo el número óptimo de 64. Se tomaron las palabras con mayor peso dentro de cada tópico, las cuales se muestran en formato "nube de palabras" en la Figura 18. El tamaño de cada palabra está asociado a la importancia del término para el tópico entregado por el modelo. Por otro lado, la herramienta entrega como resultado el peso de cada tópico en cada resumen analizado. Debido a que cada resumen tiene uno o más genes relacionados, se hizo un promedio del peso por tópico de cada gen. De esta forma cada gen posee diferentes distribuciones de peso en cada tópico, lo que nos otorga información objetiva acerca de los posibles roles que tiene el gen en estudio. Para evaluar si el modelo de tópicos propuesto tiene sentido biológico, se utilizaron 2 ejemplos de descripción de genes. Como se muestra en la Figura 19, podemos observar que se consultó por el Gen AT1G77760 (NIA1) y AT1G08090 (NRT2.1), podemos observar que los tópicos con mayor peso están relacionados con la descripción del gen, sugiriendo que puede ser una buena descripción objetiva acerca de cada gen. Se puede destacar que, en los 2 casos, se encuentra un tópico cuyas palabras están relacionadas con el nutriente nitrato, y para la nitrato reductasa tiene asociada una nube de palabras asociadas con procesos de oxido-reducción y para el transportador NRT2.1 se encuentra asociado un tópico asociado con transporte. Para utilizar esta información derivada de los modelos de tópicos latentes está en desarrollo un sitio web que permite hacer consultas de forma interactiva http://networks.bio.puc.cl/linkedgene/#/. Para analizar listas de genes utilizando esta metodología se propone realizar un análisis de sobrerrepresentación de tópicos. Para asignar los tópicos con mayor peso a cada gen, se fijó como umbral el percentil 5% de la

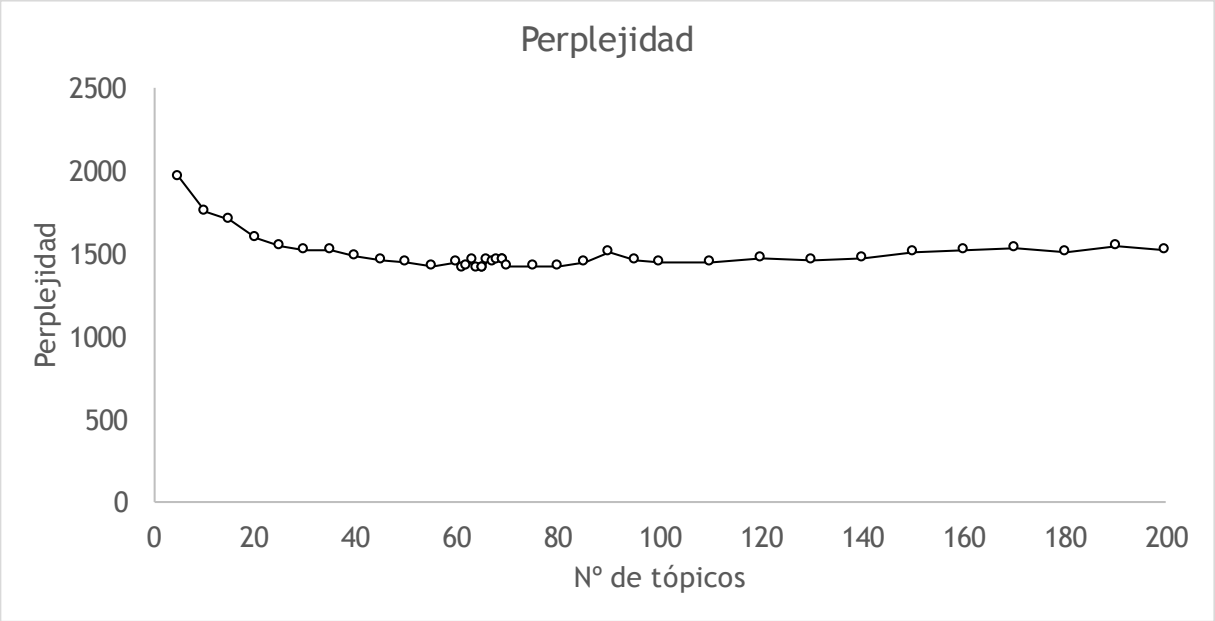


Figura 17. Ensayo de perplejidad.

A partir de los 30355 resúmenes de publicaciones relacionadas a genes de Arabidopsis se calculó una curva de perplejidad para determinar el número óptimo de tópicos a elegir. El ensayo se hizo cada 5 tópicos. Se añadieron nuevos ensayos de números tópicos cercanos al mínimo (64).

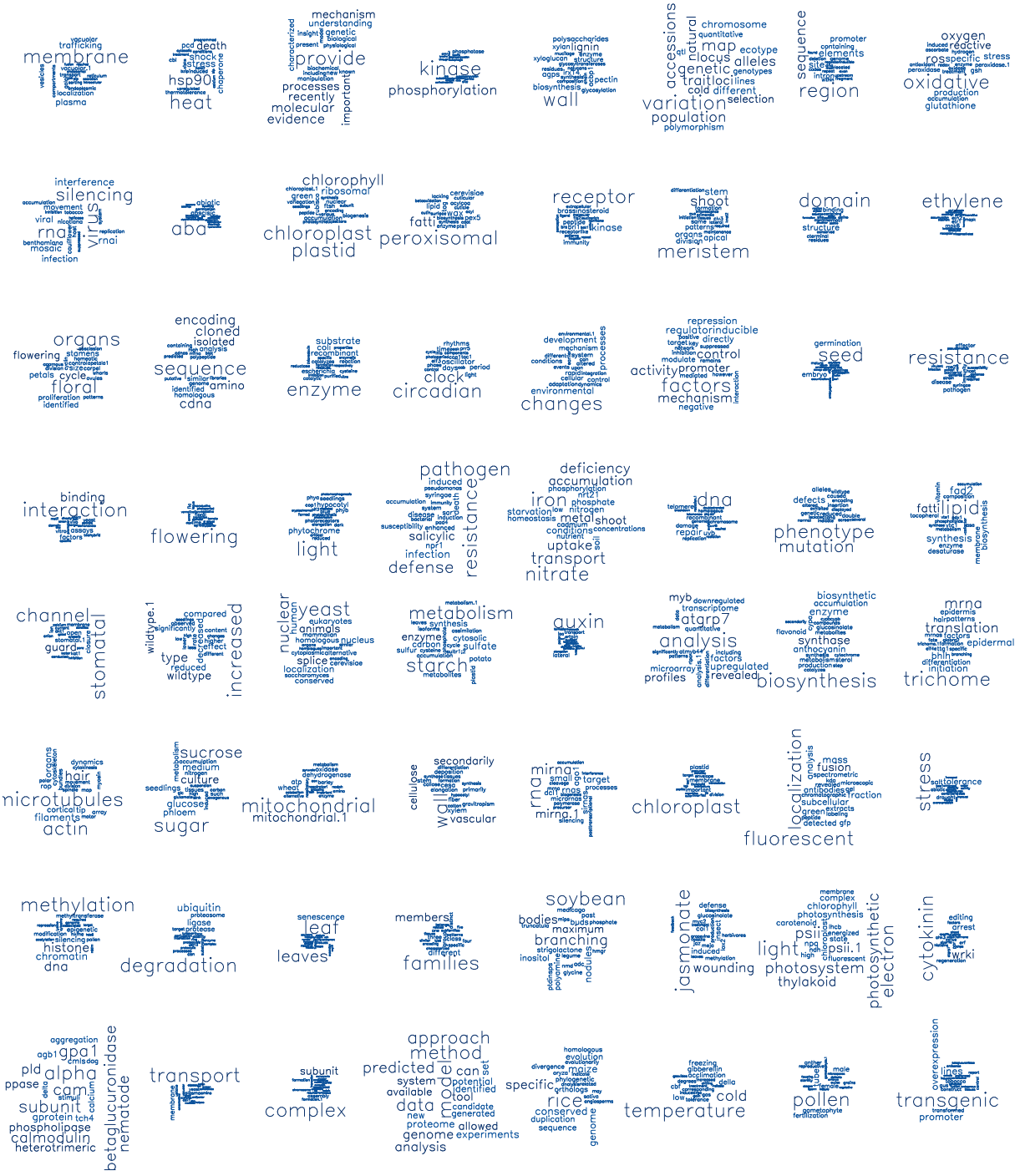


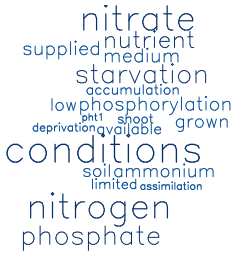
Figura 18. Listado de tópicos en formato de nube de palabras.

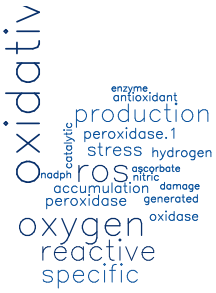
Mediante el análisis de tópicos latentes, se selecciona el número óptimo de 64 posibles tópicos. Se muestra por cada tópico detectado una imagen en formato WordCloud con los conceptos contenidos en cada tópico. El tamaño de la palabra está de acuerdo con su importancia en el modelo.

distribución de pesos, asumiendo así que cada gen queda asociado al o los tópicos más representativos de lo que se ha encontrado en literatura.

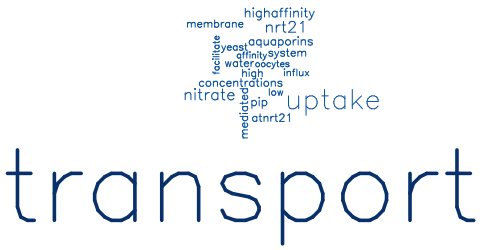
Se realizó un análisis de sobrerrepresentación de tópicos, comparando lo encontrado en las listas de genes DE o DEPI en las condiciones experimentales. Los resultados se muestran en la figura 20, donde podemos ver que lo encontrado como enriquecido significativamente coincide con lo esperado con las condiciones experimentales, relacionando el gen con su tópico independiente de la anotación de Gene Ontology. Por otro lado, podemos observar que, para estos ejemplos, lo encontrado por DEPI , en la mayoría de los casos, es un subconjunto de lo detectado por DE, pero con diferentes grupos de genes, lo que al igual que lo visto anteriormente indica complementariedad. Para evaluar si la nueva estrategia nos aporta información complementaria a lo obtenido por Gene Ontology, se evaluó cuantos de los genes identificados como regulados por alguno de los métodos se puede asociar al tópico que posea el nombre del estímulo en las primeras 10 palabras. En la Figura 21 A se muestra el número de genes identificados por alguno de los métodos propuestos, para los cuales se identificaron cuáles estaban asociados al termino GO más probable para el respectivo estímulo (Ver Métodos y Figura 12). En la Figura 21 B se observa que solo un pequeño porcentaje está asociado al GO más probable y un porcentaje importante mostrado en rojo no lo está. Al hacer la asociación con el tópico más probable se puede observar que el porcentaje de genes asociados a la respuesta aumenta indicando complementariedad y un aporte de información.

AT1G77760
Alias: GNR1, NIA1, NITRATE REDUCTASE 1, NR
Description:Encodes the cytosolic minor isoform of nitrate reductase (NR). Involved in the first step of nitrate assimilation, it contributes about 15% of the nitrate reductase activity in shoots.





AT1G08090
Alias: NITRATE TRANSPORTER 2.1
Description: High-affinity nitrate transporter. Up-regulated by nitrate. Functions as a repressor of lateral root initiation independently of nitrate uptake.



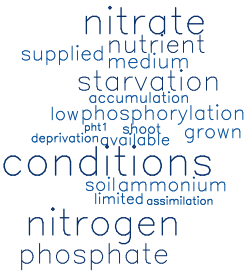


Figura 19. Análisis de tópicos de genes individuales.

Se muestran el nombre, el alias, la descripción de 2 genes de Arabidopsis, con sus respectivos tópicos más representativos utilizando 64 posibles tópicos.

A

Tópico	Nº de genes en DE	Nº de genes en DEPI	P ajustado DE	P ajustado DEPI	Nº genes genoma	Palabras
12	48	36	1.92E-12	3.51E-11	799	nitrat,iron,transport,defici,up
34	24	16	0.000402068	0.008877883	507	cytokinin,wrki,arr,edit
52	35	15	0.000403673	1	915	stress,toler,salt,drought
53	22	8	0.000479115	1	445	sugar,sucros,cultur,glucos
57	29	18	2E-06	0.001289829	528	starch,metabol,enzym,sulfat
60	77	44	7.21E-07	0.01464142	2347	analysi,atg,profil,reveal,myb,

B

Figura 20. Análisis de sobrerrepresentación de tópicos en listas.

La tabla muestra el análisis de sobrerrepresentación de listas, donde se muestra el número del tópico asignado (considerando 64), cuantos genes son asignados en cada lista y en el genoma. Luego se muestra el valor P ajustado utilizando los 2 métodos propuestos DE y DEPI seguido por las principales palabras asociadas a cada tópico. En B se muestran las principales palabras de cada tópico sobrerrepresentado en formato de nube de palabras.

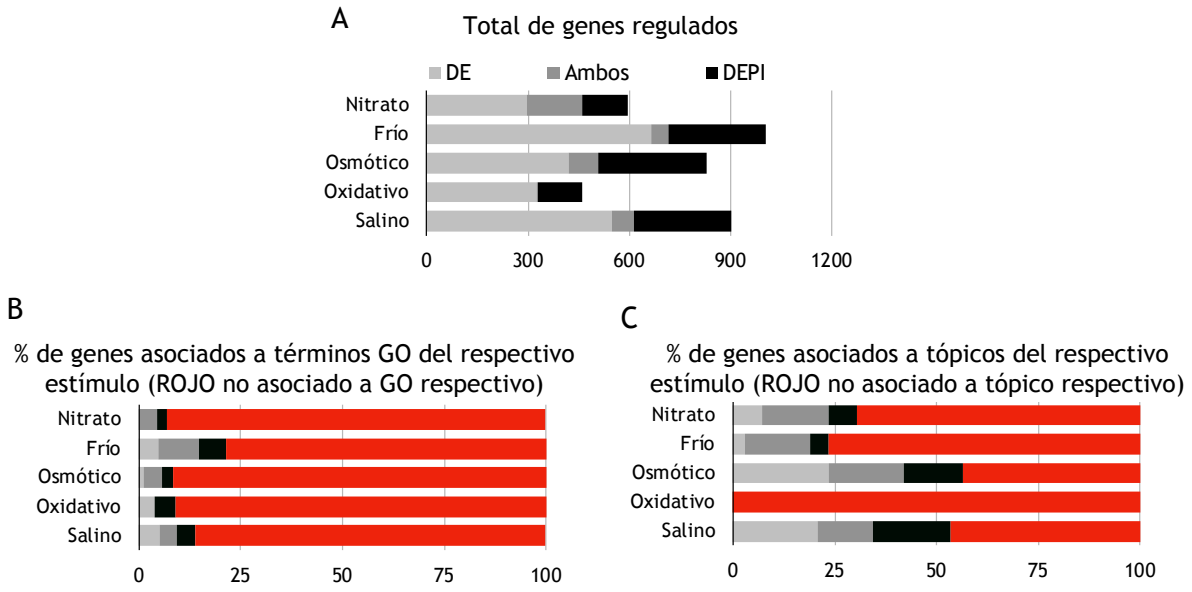


Figura 21. Análisis de tópicos pueden complementar información de Gene Ontology.

En A se muestra el número de genes identificado por los 2 métodos. En B y C se muestra el porcentaje de genes identificados por cada método que están asociados y en rojo lo que no está asociado al GO y al tópico más probables respectivamente.

8. DISCUSIÓN

Es ampliamente aceptado que los cambios ocurridos en los organismos durante el desarrollo o frente a señales ambientales son en parte mediados por cambios en la expresión génica. El transcriptoma de una célula refleja el estado del sistema bajo una condición específica, sin embargo, estos estados del transcriptoma tienen ciertas restricciones que lo definen. Estas limitaciones en el estado del transcriptoma no han sido del todo formalizadas ni sus estructuras sistemáticamente exploradas para ningún sistema biológico. En este trabajo se presenta un modelo que muestra una forma diferente de observar el transcriptoma de un organismo y sus posibles límites. De estos límites se describen posibles estructuras del transcriptoma y sus funciones biológicas asociadas, usando un nuevo marco teórico que combina teoría de la información y datos públicos de *Arabidopsis thaliana* y *Saccharomyces cerevisiae.*

Nuestro trabajo se enfoca en cómo se relacionan los genes, teniendo en cuenta que la unidad básica de relación es el par. La forma que se propone para evaluar el comportamiento de los genes está basada en que tan frecuente es una razón entre los 2 genes de un par en múltiples condiciones experimentales, es decir en la distribución de razones de expresión. Existen

diferentes formas de evaluar una distribución, en nuestro caso, lo que nos interesa es conocer cómo se comporta esta relación y describirla de forma objetiva. Luego de ensayar algunas medidas estadísticas típicas para evaluar la forma de la distribución como la moda y la curtosis, nos dimos cuenta de que podían existir diferentes tipos y formas de distribuciones. Para detectar posibles restricciones en el comportamiento de 2 genes basándose en la distribución, se decidió utilizar una medida utilizada en la teoría de la información conocida como la entropía de Shannon. Para esta medida, no es relevante la forma de la distribución, pero si la probabilidad de que un evento ocurra, considerando como evento cada intervalo del histograma de la distribución de razones. La entropía de Shannon muestra el grado de incerteza en diferentes tipos de datos, si la entropía de Shannon da un valor bajo, indica que tiene un bajo nivel de incerteza en la información, lo que sugiere que las 2 variables están fuertemente relacionadas. Es decir, la información de una de las variables describe el comportamiento de la otra variable. Biológicamente esto sugiere que existe algo, no descrito específicamente, que provoca que el comportamiento entre este par de genes tenga una restricción. Como se observa en los resultados, casos extremos de estos valores aparecen en genes relacionados con el complejo ribosomal, donde se puede suponer que existen fuertes presiones evolutivas para que los miembros de este complejo actúen coordinados, más aún, se ha reportado que desbalance por mutación de algunos de estos genes afecta el crecimiento y división celular en diferentes organismos (revisado en(Barakat et al., 2001).) Por otro lado, genes de alto GPE, es decir, con un alto grado de incerteza, no se puede conocer el comportamiento de una variable respecto a la otra. Biológicamente puede sugerir que son genes de comportamiento más libre entre ellos, que no están en un proceso relacionado, o que permiten explorar diferentes estados para hacer frente a cambios en las condiciones ambientales.

Como se mencionó anteriormente, existen diferentes métodos que se han utilizado para evaluar relaciones entre genes basados en su expresión, el más utilizado es la correlación seguido por la información mutua. Si bien estos métodos son ampliamente utilizados, no toda la información que se encuentra en la expresión puede ser detectada. Un muy buen ejemplo gráfico del uso de métodos para evaluar el comportamiento entre 2 variables es el cuarteto de Anscombe (Hanna et al., 2010; Nielsen, 2016). Estos son 4 conjuntos de datos, los cuales comparten idénticos valores de media, varianza, regresión lineal y correlación, pero que visualmente son muy diferentes (Nielsen, 2016). Al evaluar la información mutua y la GPE de estos datos, se puede observar que los valores son diferentes, demostrando que métodos alternativos permiten capturar otro componente utilizando los mismos datos. Si bien el orden encontrado en información mutua y entropía es similar, la magnitud de cambio es diferente, lo que nos indica que se está detectando un componente diferente.

Cuando se seleccionan los "mejores" pares utilizando el método propuesto y los 2 más utilizados, es decir la correlación y la información mutua, se define un número fijo de pares para los 3 métodos evaluados, para así facilitar la comparación. Se puede observar que existe una alta intersección entre los pares de genes con la más alta información mutua y alta correlación. En literatura se ha reportado que si bien las métricas y formas de cálculo son diferentes, en la práctica se han observado resultados similares (Steuer et al., 2002; Bansal et al., 2007). El cálculo de la información mutua puede realizarse a través de diferentes métodos, en este caso utilizamos el propuesto por Qiu 2009. Recientemente en el trabajo de Ish-Horowicz (Ish-Horowicz y Reid, 2017), el cual está disponible pero no revisado aún en Peer review, pone a disposición 6 variantes metodologías para calcular información mutua. Estas entregan

resultados similares a lo obtenido con el método utilizado en la sección de resultados. De igual forma podemos observar que los pares de genes encontrados como de baja GPE, en general poseen poca intersección con los otros 2 métodos. Al revisar los valores de correlación e información mutua, de estos pares de baja GPE, podemos ver que poseen valores medianamente altos de correlación e información mutua. Esto nos puede mostrar que, si bien tienen relación, estamos detectando algún componente diferente del comportamiento de los pares de genes. Algo similar ocurre con la información de levadura, lo que es un indicativo que esta métrica puede servir para diferentes organismos. La principal diferencia en los métodos utilizados es el espacio en que se están describiendo las relaciones, si bien los datos utilizados son los mismo, cuando se calcula el valor de GPE se está evaluando el espacio de las diferencias entre los genes, lo cual puede otorgar información relacionada, pero diferente a lo obtenido por los otros métodos.

Para evaluar si este componente detectado tiene sentido biológico se intersectaron las diferentes listas de pares con información de bases de datos de interacciones. Para Arabidopsis se utilizó la base de datos ANAP (Wang et al., 2012), la cual recopila información de 11 bases de datos de interacción de genes de Arabidopsis. La intersección fue realizada con las interacciones definidas por medio de métodos experimentales o inferidas por el curador. Para levadura, se utilizó la información de la base de datos BIOGRID, la cual es una de las más completas para levadura. En ambos organismos observamos que el nivel de validación obtenidos por los diferentes métodos es comparable. Como esta intersección puede ser debida al azar, se hicieron 10000 comparaciones con el mismo número de pares de genes obtenidos al azar. Lo detectado por los diferentes métodos en todos los casos fue más que lo esperado por azar, lo que nos indica que el método propuesto entrega información complementaria a los métodos

tradicionales. De los pares reportados en las bases de datos se excluyeron las interacciones depositadas como de correlación y correlación de ortólogos ya que pueden distorsionar los resultados al ser igual a una de las métricas que estamos analizando.

Como estos pares de baja GPE poseen la característica que existen fuertes restricciones para que se expresen en una relación similar en la mayoría de las condiciones, hace pensar que pueden poseer algún componente de genes esenciales para el desarrollo del organismo. Para descartar esto, se analizó si es que estaban enriquecidos en el grupo de genes CEGMA. Este último es un conjunto de genes que está presente en la mayoría de los eucariotas, el resultado del análisis fue que estaba enriquecido con un $p < 1*10^{-6}$.

Al hacer la red con los pares de baja GPE, podemos observar que se tiene un gran componente conectado. Si cambiamos el umbral de corte del valor de GPE, aparecen diferentes pares sin conexión. Utilizando el mismo número de pares que en el análisis anterior, podemos agrupar los genes por topología de la red y ver qué tipo de procesos están relacionados a cada clúster. Podemos observar que la mayoría de los procesos sobrerrepresentados tienen relación a procesos necesarios para el funcionamiento del organismo. Esto nos muestra que pares de baja GPE están restringidos para la mantención de organismo. Lo expuesto anteriormente, nos sugiere que el valor de GPE es un valor útil y con sentido biológico. Al análizar como se distribuye la entropía promedio de los pares anotados dentro de un proceso de la base de datos KEGG nos muestra que hay procesos que tienen una baja entropía, es decir que están posicionados en un percentil bajo de la distribución, los cuales están asociados al funcionamiento básico de la célula, los cuales en general están compuestos por genes que forman complejos proteicos, incluyendo el proteosoma, la maquinaria de transcripción basal,

ribosomas entre otros. Los genes participantes en estos procesos tienen un comportamiento en que las relaciones se mantienen independiente de la condición experimental en que se encuentren, lo que indica que su expresión posee restricciones. Procesos con valor alto de GPE, es decir, pares de genes que fluctúan con más flexibilidad, están involucrados en vías relacionadas con condiciones ambientales cambiantes. Estas vías incluyen un precursor del ácido jasmónico como es el ácido Alfa-Linolénico y el metabolismo del nitrógeno que pueden ser dependientes de la condicicón ambiental en que se evalúa el transcriptoma. El número de genes relacionado a cada proceso no correlaciona significativamente con el valor medio de GPE, lo cual sugiere que no hay sesgo entre el número de genes involucrados y el valor medio de GPE. Al considerar los valores promedio de GPE de los diferentes procesos y compararlos con la simulación de genes al azar (considerando el mismo número de genes), se obtiene que 26 y 16 procesos de baja y alta entropía respectivemente tienen un valor significativo considerando 10000 iteraciones de muestreos aleatorios. Estos grupos de genes con promedio de entropía significativamente baja pueden ser potenciales clasificadores de funciones de genes sin función conocida.

Debido a que cuando las condiciones ambientales cambian o el organismo tiene algún cambio de estado del desarrollo, puede cambiar el nivel de expresión de ciertos genes. Considerando que en general, desde el punto de vista de la expresión diferencial, solo un 1.5% de los genes tiene cambios estadísticamente significativos entre 2 condiciones (Aceituno et al., 2008), las relaciones entre genes se verán afectada. El organismo al estar en una condición diferente, el perfil de expresión para algunos genes se va a modificar, por lo tanto, las relaciones entre ellos cambiarán. Esta información es la que capturamos con el valor de ΔGPE, el cual nos

indica cuanto se perturba el sistema al cambiar el perfil de expresión. Como vemos en la Figura 10, el cuanto se perturba una relación va a estar determinado por el comportamiento o el nivel de restricciones que posee cada par de genes. Junto con esto se puede cuantificar cuanto se perturba cada gen tomando en cuenta todas las posibles relaciones.

Como vimos anteriormente hay diferentes fuentes ya sea experimentales o predichas que nos entregan información de cómo se relacionan los genes, sin embargo, esto solo cubre una pequeña fracción del número total de posibles interacciones biológicas. En nuestro caso, para calcular cuánto se perturba un gen, estamos tomando en cuenta todas las posibles combinaciones de relaciones de genes. Si bien en muchos casos no hay evidencia de que un gen está relacionado con otro, la expresión de un gen puede afectar de muchas maneras al resto. Por ejemplo, la alta expresión de un gen puede saturar los ribosomas y así modificar el patrón de traducción de otro gen (Brockmann et al., 2007; Liu et al., 2016). Existen otros casos en que un transcrito al expresarse puede actuar como "esponja" de microRNAs, donde la alta expresión de un blanco de microRNA provoca un secuestro que libera al resto de los blancos (Bak y Mikkelsen, 2014). A pesar de que nombramos solo 2 ejemplos, esto indica que hay relaciones intrínsecas entre genes que no podemos observar, pero que actúan e influyen en el comportamiento del organismo.

Un trabajo que podría tener relación con nuestra propuesta es el publicado por Wang 2014, en este utilizan la entropía de Shannon diferencial y análisis de coeficientes de variación para detectar genes claves en enfermedades que no son detectados por métodos de expresión diferencial por si solos (Wang et al., 2014), sin embargo tiene 2 grandes diferencias con nuestro

análisis: El análisis que proponemos se basa en las relaciones entre todos los pares de genes y en el análisis del comportamiento del gen en múltiples condiciones experimentales.

Al hacer el test estadístico tanto para los valores de expresión como para los valores de EPI utilizamos RankProd (Hong et al., 2006), un método no paramétrico ampliamente utilizado para identificar genes que se expresan diferencialmente en 2 condiciones. Al comparar los valores de expresión se identifican como regulados tanto los genes inducidos y reprimidos en la nueva condición. Para los valores de EPI, se identifican como regulados solamente los que tienen una perturbación "inducida" por el tratamiento, descartando los genes que se "reprime" su perturbación debido a que, según nuestra propuesta, no estaría identificando un cambio. Dentro de los genes DEPI existen genes que su expresión está inducida o reprimida por el tratamiento. En estos casos estamos observando genes que su expresión provoca un cambio en las relaciones con el resto de los genes que debido a una topología no observada provoca que el transcriptoma se encuentre perturbado. La condición de este gen la podríamos describir como que la expresión de este gen cambia y tiene las relaciones perturbadas. Existen otros casos, en que un gen es diferencialmente expresado, pero no es DEPI. Lo cual quiere decir que el cambio en la expresión de este gen posee menos restricciones, por lo que en lo global no se afectan de forma significativas sus relaciones. Por último, está el caso en que un gen no está diferencialmente expresado, pero si es DEPI, esto quiere decir que a pesar de que el gen no cambie o cambie su expresión de forma no significativa, el entorno de este gen está alterado, lo que puede afectar su funcionamiento, lo cual es el principal aporte de esta propuesta metodológica ya que no se puede observar por la simple expresión génica, siendo la estrategia planteada un buen complemento.

Como vimos en la sección de resultados se analizaron los datos para identificar listas de genes DE y DEPI en un experimento de plantas tratadas con KNO3 o KCl como control realizado en nuestro laboratorio y en experimentos de la serie de ATGEenxpress. En estos experimentos se identificaron los diferentes comportamientos de la respuesta de un gen a un cambio en las condiciones experimentales. Para validarlo se utilizó información de la base de datos de GO, la cual, entre otras informaciones, relaciona los diferentes genes con posibles funciones o procesos biológicos, estas relaciones son obtenidas por diferentes métodos experimentales o inferidas de forma computacional. Se intersectaron las listas obtenidas anteriormente con esta información, considerando para cada condición experimental, los procesos biológicos más relacionados. Si bien la mayoría de los genes anotados en una función son identificados gracias a cambios en la expresión génica, podemos observar que las proporciones encontradas por DE o por DEPI son similares y complementarias, lo que le otorga un nuevo valor a esta métrica. Si se hace un análisis de precisión y sensibilidad, o la mezcla de ellos en el estadístico F, podemos observar que los valores para las listas DE y DEPI no son tan diferentes, lo que confirma que son complementarias.

Si nos centramos en el experimento de nitrato y en las listas de genes anotadas en GO como de respuesta a nitrato, hay genes que se detectan por uno u otro método. Dentro de los genes que se detectan solo por DE está el gen AT5G65210 (TGA1) y BT2, los cuales se han descrito responden a múltiples estímulos (Aceituno et al., 2008), esto confirma lo expresado anteriormente en cuanto a que estos genes tienen un comportamiento con menos restricciones, por lo tanto no los detectamos como DEPI. Dentro de este mismo experimento, es detectado como DEPI el gen NLP7, el cual es un importante coordinador de la respuesta a nitrato, el cual

fue anotado como de respuesta a nitrato gracias a un análisis de mutantes, el cual se expresa constitutivamente y es un gen clave en la nodulación en leguminosas y orquesta la respuesta temprana a nitrato en plantas (Castaings et al., 2009; Marchive et al., 2013), este gen no cambia significativamente su expresión en respuesta a nitrato, lo que indica que estamos detectando este componente de la respuesta no identificado con los anteriores métodos, confirmando nuestra metodología como un análisis complementario utilizando la misma información. Otro caso importante en los genes de respuesta a nitrato es la aparición del gen AFB3 (AUXIN SIGNALING F-BOX 3), el cual aparece como regulado en DEPI y no en DE en los datos evaluados en este trabajo, en estudios previos se ha reportado como regulado por nitrato, pero en un tiempo más temprano al del tratamiento evaluado en este trabajo (Vidal et al., 2010a; Vidal et al., 2014). Esto último nos sugiere que este gen detectado puede encontrarse en una condición perturbada en el punto de la medición.

Este trabajo se realizó utilizando los datos de transcriptómica de la base de datos Nascarray para Arabidopsis (Craigon et al., 2004) y para levadura se descargaron mediante la búsqueda en la base de datos Arrayexpress (Brazma et al., 2003). Por motivos de estabilizar el set de datos, una vez creada la propia base de datos no se han añadido nuevas condiciones para el cálculo de la GPE. Sin embargo, este método puede ser utilizado con un mayor número de condiciones experimentales, así como también con otros organismos o métodos de medición. Es de suponer que con el tiempo este método puede ir perfeccionándose al agregar nuevas condiciones gracias al rápido crecimiento en la cantidad de condiciones experimentales que se han analizado. Utilizando el mismo esquema de análisis, se puede aplicar a otros tipos de datos, por ejemplo datos RNA-seq, que ha ido aumentando cada vez más rápido el número de trabajos depositado

en las bases de datos, de 2011 a 2016, se ha incrementado 10 veces el número de experimentos realizados sólo en Arabidopsis (Kolesnikov et al., 2015), lo que nos permitirá aplicar el modelo a otros organismos.

Si bien la propuesta del método de análisis de texto para describir en que está involucrado un gen puede ser considerada como una herramienta adicional que puede no estar directamente involucrada en el tema central de la tesis, es una herramienta útil para analizar y comprender listas de genes. Para buscar información de la función de un gen, una de las bases de datos más completas es la de Gene Ongology. Para Arabidopsis, si bien el porcentaje de anotación en Gene Ontology es alto solo el 20,8% de los genes anotados tiene evidencia experimental, es decir, la mayoría de las anotaciones son obtenidas de forma electrónica sin revisión de curadores. Por otro lado, dependiendo de la profundidad de la búsqueda generalmente es posible encontrar información relacionada en alguna publicación. Sin embargo, esta búsqueda puede ser utilizada en diferente profundidad, lo que va a influir en el resultado. Por estas 2 razones, esta es una herramienta útil, sencilla y objetiva para en nuestro caso entregar información contenida en literatura. Para hacer este análisis, es necesario que algún curador de la base de datos haya agregado la vinculación de un gen a un trabajo de Pubmed en particular, lo que puede no ser tan objetivo como se desea. Una elección arbitraria para el proceso de análisis de tópicos es la selección del número de tópicos posibles en el modelo. En nuestro caso seleccionamos 2,4,8,16,64,128,256 y 512 tópicos, con lo que se puede obaservar información a diferente profundidad. La asignación de genes a tópicos nos permite analizar listas de genes ya sea lo encontrado por DE o por DEPI entregando como resultado que lo seleccionado por DEPI puede

entregar información relacionada con la respuesta al estímulo complementaria a lo encontrado por DE ya que no son los mismos genes detectados.

9. CONCLUSIONES

Las células o tejidos presentan perfiles de expresión característicos o estados, que determinan su comportamiento. En esta tesis, investigamos el espacio en que se encuentran todos los estados posibles y estudiamos sus restricciones. Aprendimos que existen estados poco probables y otros de alta probabilidad, a los cuales asignamos de forma indirecta un costo dependiente del comportamiento habitual de cada gen. Se propuso que los diferentes estados operan bajo condiciones ambientales consideradas normales y perturbaciones ambientales producen cambios de estado que requiere de un ajuste del sistema, modificando el comportamiento y las relaciones de los genes del organismo. Con la información entregada por los estados se pudo determinar las relaciones probables entre genes, entregando información complementaria a la correlación e información mutua. Se propuso un nuevo método en el cual se puede identificar un componente de la respuesta del organismo frente a perturbaciones basado en las relaciones entre genes que no se identifica con el análisis de expresión diferencial típico. Adicionalmente, se diseño una herramienta basada en análisis semántico de textos que entrega información acerca de en que temas o tópcios puede estar involucrado un gen o una lista de genes.

10. REFERENCIAS

Aceituno FF, Moseyko N, Rhee SY, Gutiérrez RA (2008) The rules of gene expression in plants: organ identity and gene body methylation are key factors for regulation of gene expression in Arabidopsis thaliana. BMC Genomics **9**: 438

Alon U (2007) Network motifs: theory and experimental approaches. Nat Rev Genet **8**: 450–461

Alvarez JM, Riveras E, Vidal EA, Gras DE, Contreras-López O, Tamayo KP, Aceituno F, Gómez I, Ruffel S, Lejay L, et al. (2014) Systems approach identifies TGA1 and TGA4 transcription factors as important regulatory components of the nitrate response of Arabidopsis thaliana roots. Plant J **80**: 1–13

Araus V, Vidal EA, Puelma T, Alamos S, Mieulet D, Guiderdoni E, Gutiérrez RA (2016) Members of BTB Gene Family of Scaffold Proteins Suppress Nitrate Uptake and Nitrogen Use Efficiency. Plant Physiol **171**: 1523–32

Audic S, Claverie JM (1997) The significance of digital gene expression profiles. Genome Res **7**: 986–95

Bak RO, Mikkelsen JG (2014) miRNA sponges: Soaking up miRNAs for regulation of gene expression. Wiley Interdiscip Rev RNA **5**: 317–333

Bansal M, Belcastro V, Ambesi-impiombato A, Bernardo D, di Bernardo D (2007) How to infer gene networks from expression profiles. Mol Syst Biol **3**: 78

Barabasi A, Oltvai Z (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet **69**: 572–6

Barakat A, Szick-Miranda K, Chang IF, Guyot R, Blanc G, Cooke R, Delseny M, Bailey-Serres J (2001) The organization of cytoplasmic ribosomal protein genes in the Arabidopsis genome. Plant Physiol **127**: 398–415

Berckmans B, Vassileva V, Schmid SP, Maes S, Parizot B, Naramoto S, Magyar Z, Alvim Kamei CL, Koncz C, Bogre L, et al. (2011) Auxin-dependent cell cycle reactivation through transcriptional regulation of Arabidopsis E2Fa by lateral organ boundary proteins. Plant Cell **23**: 3671–3683

Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, Fridman WH,

Pagès F, Trajanoski Z, Galon J (2009) ClueGO: A Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics **25**: 1091–1093

Birkenbihl RP, Diezel C, Somssich IE (2012) Arabidopsis WRKY33 Is a Key Transcriptional Regulator of Hormonal and Metabolic Responses toward Botrytis cinerea Infection. Plant Physiol **159**: 266–285

Blei DM, Ng AY, Jordan MI (2012) Latent Dirichlet Allocation. J Mach Learn Res **3**: 993–1022

De Bodt S, Carvajal D, Hollunder J, Van den Cruyce J, Movahedi S, Inzé D (2010) CORNET: a user-friendly tool for data mining and integration. Plant Physiol **152**: 1167–79

De Bodt S, Hollunder J, Nelissen H, Meulemeester N, Inzé D (2012) CORNET 2.0: integrating plant coexpression, protein-protein interactions, regulatory interactions, gene associations and functional annotations. New Phytol **195**: 707–20

Bracha-Drori K, Shichrur K, Katz A, Oliva M, Angelovici R, Yalovsky S, Ohad N (2004) Detection of protein-protein interactions in plants using bimolecular fluorescence complementation. Plant J **40**: 419–27

Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG, et al. (2003) ArrayExpress - A public repository for microarray gene expression data at the EBI. Nucleic Acids Res **31**: 68–71

Brockmann R, Beyer A, Heinisch JJ, Wilhelm T (2007) Posttranscriptional expression regulation: What determines translation rates? PLoS Comput Biol **3**: 0531–0539

Canales J, Moyano TC, Villarroel E, Gutiérrez RA (2014) Systems analysis of transcriptome data provides new hypotheses about Arabidopsis root response to nitrate treatments. Front Plant Sci **5**: 22

Castaings L, Camargo A, Pocholle D, Gaudon V, Texier Y, Boutet-Mercey S, Taconnat L, Renou J-P, Daniel-Vedele F, Fernandez E, et al. (2009) The nodule inception-like protein 7 modulates nitrate sensing and metabolism in Arabidopsis. Plant J **57**: 426–35

Causton HC, Ren B, Koh SS, Harbison CT, Kanin E, Jennings EG, Lee TI, True HL, Lander ES, Young RA (2001) Remodeling of yeast genome expression in response to environmental changes. Mol Biol Cell **12**: 323–37

Chatr-aryamontri A, Breitkreutz B-J, Oughtred R, Boucher L, Heinicke S, Chen D, Stark C, Breitkreutz A, Kolas N, O'Donnell L, et al. (2015) The BioGRID interaction database: 2015 update. Nucleic Acids Res **43**: D470–D478

Chen L, Song Y, Li S, Zhang L, Zou C, Yu D (2012) The role of WRKY transcription factors in plant abiotic stresses. Biochim Biophys Acta - Gene Regul Mech **1819**: 120–128

Chen S-C, Tsai T-H, Chung C-H, Li W-H (2015) Dynamic association rules for gene expression data analysis. BMC Genomics **16**: 786

Cheng C, Yan K-K, Hwang W, Qian J, Bhardwaj N, Rozowsky J, Lu ZJ, Niu W, Alves P, Kato M, et al. (2011) Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data. PLoS Comput Biol **7**: e1002190

Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, et al. (2012) Saccharomyces Genome Database: The genomics resource of budding yeast. Nucleic Acids Res. doi: 10.1093/nar/gkr1029

Childs KL, Davidson RM, Buell CR (2011) Gene coexpression network analysis as a source of functional annotation for rice genes. PLoS One. doi: 10.1371/journal.pone.0022196

Ciruela F (2008) Fluorescence-based methods in the study of protein-protein interactions in living cells. Curr Opin Biotechnol **19**: 338–43

Cossio P, Granata D, Laio A, Seno F, Trovato A (2012) A simple and efficient statistical potential for scoring ensembles of protein structures. Sci Rep **2**: 1–8

Craigon DJ, James N, Okyere J, Higgins J, Jotham J, May S (2004) NASCArrays: a repository for microarray data generated by NASC's transcriptomics service. Nucleic Acids Res **32**: D575–D577

Creighton C, Hanash S (2003) Mining gene expression databases for association rules. Bioinformatics **19**: 79–86

Cunningham MJ, Liang S, Fuhrman S, Seilhamer JJ, Somogyi R (2000) Gene expression microarray data analysis for toxicology profiling. Ann N Y Acad Sci **919**: 52–67

D'haeseleer P, Liang S, Somogyi R (2000) Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics **16**: 707–26

DiLeo M V., Strahan GD, den Bakker M, Hoekenga OA (2011) Weighted correlation network analysis (WGCNA) applied to the tomato fruit metabolome. PLoS One. doi: 10.1371/journal.pone.0026683

Dong M a, Farré EM, Thomashow MF (2011) Circadian clock-associated 1 and late elongated hypocotyl regulate expression of the C-repeat binding factor (CBF) pathway in Arabidopsis. Proc Natl Acad Sci U S A **108**: 7241–6

Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A **95**: 14863–8

Ferro A, Giugno R, Mongiovi M, Pigola G, Pulvirenti A (2006) Distributed antipole clustering for efficient data search and management in Euclidean and metric spaces. Proc. 20th IEEE Int. Parallel Distrib. Process. Symp. IEEE, p 12 pp.

Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C, et al. (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res **41**: D808-15

Friedel S, Usadel B, von Wirén N, Sreenivasulu N (2012) Reverse engineering: a key component of systems biology to unravel global abiotic stress cross-talk. Front Plant Sci **3**: 294

Fuhrman S, Cunningham MJ, Wen X, Zweiger G, Seilhamer JJ, Somogyi R (2000) The application of Shannon entropy in the identification of putative drug targets. BioSystems. pp 5–14

Gautier L, Cope L, Bolstad BM, Irizarry RA (2004) Affy - Analysis of Affymetrix GeneChip data at the probe level. Bioinformatics **20**: 307–315

Gene Ontology Consortium (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res **32**: 258D–261

Guo Y, Feng Y, Trivedi NS, Huang S (2011) Medusa structure of the gene regulatory network: dominance of transcription factors in cancer subtype classification. Exp Biol Med (Maywood) **236**: 628–36

Gutiérrez RA (2012) Systems biology for enhanced plant nitrogen nutrition. Science **336**: 1673–5

Gutiérrez RA, Gifford ML, Poultney C, Wang R, Shasha DE, Coruzzi GM, Crawford NM (2007a) Insights into the genomic nitrate response using genetics and the Sungear Software System. J Exp Bot **58**: 2359–67

Gutiérrez RA, Lejay L V, Dean A, Chiaromonte F, Shasha DE, Coruzzi GM (2007b)

Qualitative network models and genome-wide expression data define carbon/nitrogen-responsive molecular machines in Arabidopsis. Genome Biol **8**: R7

Gutiérrez RA, Shasha DE, Coruzzi GM (2005) Systems biology for the virtual plant. Plant Physiol **138**: 550–4

Gutiérrez RA, Stokes TL, Thum K, Xu X, Obertello M, Katari MS, Tanurdzic M, Dean A, Nero DC, McClung CR, et al. (2008) Systems approach identifies an organic nitrogen-responsive gene network that is regulated by the master clock control gene CCA1. Proc Natl Acad Sci U S A **105**: 4939–44

Guyon I, Elisseeff A (2003) An Introduction to Variable and Feature Selection. J Mach Learn Res **3**: 1157–1182

Hahn A, Kilian J, Mohrholz A, Ladwig F, Peschke F, Dautel R, Harter K, Berendzen KW, Wanke D (2013) Plant core environmental stress response genes are systemically coordinated during abiotic stresses. Int J Mol Sci **14**: 7617–7641

Hanna ARG, Rao C, Athanasiou T, Anscombe FJ, Hanna ARG, Rao C, Athanasiou T (2010) Graphs in statistical analysis. Key Top Surg Res Methodol **27**: 441–475

Hong F, Breitling R, McEntee CW, Wittner BS, Nemhauser JL, Chory J (2006) RankProd: A bioconductor package for detecting differentially expressed genes in meta-analysis. Bioinformatics **22**: 2825–2827

Huang S (1999) Gene expression profiling, genetic networks, and cellular states: an integrating concept for tumorigenesis and drug discovery. J Mol Med (Berl) **77**: 469–80

Huang S (2012) The molecular and mathematical basis of Waddington's epigenetic landscape: a framework for post-Darwinian biology? Bioessays **34**: 149–57

Huang S (2010) Cell lineage determination in state space: a systems view brings flexibility to dogmatic canonical rules. PLoS Biol **8**: e1000380

Huang S, Eichler G, Bar-Yam Y, Ingber DE (2005) Cell Fates as High-Dimensional Attractor States of a Complex Gene Regulatory Network. Phys Rev Lett **94**: 1–4

Huang S, Guo Y-P, May G, Enver T (2007) Bifurcation dynamics in lineage-commitment in bipotent progenitor cells. Dev Biol **305**: 695–713

Ideker T, Galitski T, Hood L (2001) A new approach to decoding life: systems biology. Annu Rev Genomics Hum Genet **2**: 343–72

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics **4**: 249–264

Ish-Horowicz J, Reid J (2017) Mutual information estimation for transcriptional regulatory network inference. doi: 10.1101/132647

Joyce AR, Palsson BØ (2006) The model organism as a system: integrating "omics" data sets. Nat Rev Mol Cell Biol **7**: 198–210

Kanehisa M (2002) The KEGG database. Novartis Found Symp **247**: 91-101–3, 119–28, 244–52

Karlebach G, Shamir R (2008) Modelling and analysis of gene regulatory networks. Nat Rev Mol Cell Biol **9**: 770–80

Kashtan N, Alon U (2005) Spontaneous evolution of modularity and network motifs. Proc Natl Acad Sci U S A **102**: 13773–8

Katari MS, Nowicki SD, Aceituno FF, Nero D, Kelfer J, Thompson LP, Cabello JM, Davidson RS, Goldberg AP, Shasha DE, et al. (2010) VirtualPlant: A Software Platform to Support Systems Biology Research. Plant Physiol **152**: 500–515

Kilian J, Whitehead D, Horak J, Wanke D, Weinl S, Batistic O, D'Angelo C, Bornberg-Bauer E, Kudla J, Harter K (2007) The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. Plant J **50**: 347–63

Kitano H (2002) Systems biology: a brief overview. Science **295**: 1662–4

Kohl M, Wiese S, Warscheid B (2011) Cytoscape: software for visualization and analysis of biological networks. Methods Mol Biol **696**: 291–303

Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, Dylag M, Kurbatova N, Brandizi M, Burdett T, et al. (2015) ArrayExpress update-simplifying data submissions. Nucleic Acids Res **43**: D1113-6

Krishnakumar V, Contrino S, Cheng CY, Belyaeva I, Ferlanti ES, Miller JR, Vaughn MW, Micklem G, Town CD, Chan AP (2017) Thalemine: A warehouse for Arabidopsis data integration and discovery. Plant Cell Physiol **58**: e4

Krishnakumar V, Hanlon MR, Contrino S, Ferlanti ES, Karamycheva S, Kim M, Rosen BD, Cheng CY, Moreira W, Mock SA, et al. (2015) Araport: The Arabidopsis Information Portal. Nucleic Acids Res **43**: D1003–D1009

Krouk G, Lacombe B, Bielach A, Perrine-Walker F, Malinska K, Mounier E, Hoyerova K, Tillard P, Leon S, Ljung K, et al. (2010a) Nitrate-regulated auxin transport by NRT1.1 defines a mechanism for nutrient sensing in plants. Dev Cell **18**: 927–37

Krouk G, Mirowski P, LeCun Y, Shasha DE, Coruzzi GM (2010b) Predictive network modeling of the high-resolution dynamic plant transcriptome in response to nitrate. Genome Biol **11**: R123

Krouk G, Tranchina D, Lejay L, Cruikshank A a, Shasha D, Coruzzi GM, Gutiérrez R a (2009) A systems approach uncovers restrictions for signal interactions regulating genome-wide responses to nutritional cues in Arabidopsis. PLoS Comput Biol **5**: e1000326

Langfelder P, Horvath S (2008) WGCNA: An R package for weighted correlation network analysis. BMC Bioinformatics. doi: 10.1186/1471-2105-9-559

Li JJ, Bickel PJ, Biggin MD (2014) System wide analyses have underestimated protein abundances and the importance of transcription in mammals. PeerJ **2**: e270

Li K-C (2002) Genome-wide coexpression dynamics: theory and application. Proc Natl Acad Sci U S A **99**: 16875–16880

Li Y, Pearl SA., Jackson S (2015) Gene Networks in Plant Biology: Approaches in Reconstruction and Analysis. Trends Plant Sci **20**: 664–675

Lister R, Gregory B, Ecker J (2009) Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond. Curr Opin Plant Biol **12**: 19157957

Liu Y, Beyer A, Aebersold R (2016) On the Dependency of Cellular Protein Levels on mRNA Abundance. Cell **165**: 535–50

Maere S, Heymans K, Kuiper M (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics **21**: 3448–9

Malik VS (2016) RNA sequencing as a tool for understanding biological complexity of abiotic stress in plants. J Plant Biochem Biotechnol **25**: 1–2

Marchive C, Roudier F, Castaings L, Bréhaut V, Blondet E, Colot V, Meyer C, Krapp A (2013) Nuclear retention of the transcription factor NLP7 orchestrates the early response to nitrate in plants. Nat Commun **4**: 1713

Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano

A (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics **7 Suppl 1**: S7

Miller C, Schwalb B, Maier K, Schulz D, Dümcke S, Zacher B, Mayer A, Sydow J, Marcinowski L, Dölken L, et al. (2011) Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. Mol Syst Biol **7**: 458

Moustafa K, Cross JM (2016) Genetic Approaches to Study Plant Responses to Environmental Stresses: An Overview. Biology (Basel) **5**: 1–18

Nielsen CB (2016) Visualization: A Mind-Machine Interface for Discovery. Trends Genet **32**: 73–75

O'Malley RC, Huang S-SC, Song L, Lewsey MG, Bartlett A, Nery JR, Galli M, Gallavotti A, Ecker JR (2016) Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. Cell **165**: 1280–1292

Obayashi T, Okamura Y, Ito S, Tadaka S, Aoki Y, Shirota M, Kinoshita K (2014) ATTED-II in 2014: evaluation of gene coexpression in agriculturally important plants. Plant Cell Physiol **55**: e6

Opitz L, Salinas-Riester G, Grade M, Jung K, Jo P, Emons G, Ghadimi BM, Beissbarth T, Gaedcke J (2010) Impact of RNA degradation on gene expression profiling. BMC Med Genomics **3**: 36

Parra G, Bradnam K, Korf I (2007) CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics **23**: 1061–1067

Parthiban V, Gromiha MM, Abhinandan M, Schomburg D (2007) Computational modeling of protein mutant stability: analysis and optimization of statistical potentials and structural features reveal insights into prediction model development. BMC Struct Biol **7**: 54

Petricka JJ, Schauer MA, Megraw M, Breakfield NW, Thompson JW, Georgiev S, Soderblom EJ, Ohler U, Moseley MA, Grossniklaus U, et al. (2012) The protein expression landscape of the Arabidopsis root. Proc Natl Acad Sci U S A **109**: 6811–8

Ponzoni I, Nueda M, Tarazona S, Götz S, Montaner D, Dussaut J, Dopazo J, Conesa A (2014) Pathway network inference from gene expression data. BMC Syst Biol **8 Suppl 2**: S7

Puelma T, Araus V, Canales J, Vidal EA, Cabello JM, Soto A, Gutiérrez RA (2017) GENIUS: web server to predict local gene networks and key genes for biological functions. Bioinformatics **33**: 760–761

Puelma T, Gutierrez RA, Soto A, Gutiérrez RA, Soto A (2012) Discriminative local subspaces in gene expression data for effective gene function prediction. Bioinformatics **28**: 2256–2264

Qiu P, Gentles AJ, Plevritis SK (2009) Fast calculation of pairwise mutual information for gene regulatory network reconstruction. Comput Methods Programs Biomed **94**: 177–180

R Core Team (2015) R: A language and environment for statistical computing. R A Lang. Environ. Stat. Comput.

Rosikiewicz M, Robinson-rechavi M (2011) IQRray , a new method for Affymetrix microarray quality control , and the homologous organ conservation score , a new benchmark method for quality control metrics. Bioinformatics **30**: 1392–1399

Rost HL, Malmstrom L, Aebersold R, Röst HL, Malmström L, Aebersold R (2015) Reproducible quantitative proteotype data matrices for systems biology. Mol Biol Cell **26**: 3926–31

Ruffel S, Krouk G, Coruzzi GM (2010) A Systems View of Responses to Nutritional Cues in

Arabidopsis: Toward a Paradigm Shift for Predictive Network Modeling. Plant Physiol **152**: 445–452

Schubert OT, Röst HL, Collins BC, Rosenberger G, Aebersold R (2017) Quantitative proteomics: Challenges and opportunities in basic and applied research. Nat Protoc **12**: 1289–1294

Schug J, Schuller W-P, Kappen C, Salbaum JM, Bucan M, Stoeckert CJ (2005) Promoter features related to tissue specificity as measured by Shannon entropy. Genome Biol **6**: R33

Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M (2011) Global quantification of mammalian gene expression control. Nature **473**: 337–342

Shalem O, Groisman B, Choder M, Dahan O, Pilpel Y (2011) Transcriptome kinetics is governed by a genome-wide coupling of mRNA production and degradation: a role for RNA Pol II. PLoS Genet **7**: e1002273

Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J **27**: 379–423

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: A software Environment for integrated models of biomolecular interaction networks. Genome Res **13**: 2498–2504

Song L, Langfelder P, Horvath S (2012) Comparison of co-expression measures: mutual information, correlation, and model based indices. BMC Bioinformatics **13**: 328

Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A, Tyers M (2006) BioGRID: a general repository for interaction datasets. Nucleic Acids Res **34**: D535–D539

Steuer R, Kurths J, Daub CO, Weise J, Selbig J (2002) The mutual information: Detecting and evaluating dependencies between variables. Bioinformatics **18**: S231–S240

Tegge AN, Caldwell CW, Xu D (2012) Pathway correlation profile of gene-gene co-expression for identifying pathway perturbation. PLoS One **7**: e52127

Usadel B, Fernie AR (2013) The plant transcriptome-from integrating observations to models. Front Plant Sci **4**: 48

Vidal EA, Álvarez JM, Gutiérrez RA (2014) Nitrate regulation of AFB3 and NAC4 gene expression in Arabidopsis roots depends on NRT1.1 nitrate transport function. Plant Signal Behav **9**: e28501

Vidal EA, Álvarez JM, Moyano TC, Gutiérrez RA (2015) Transcriptional networks in the nitrate response of Arabidopsis thaliana. Curr Opin Plant Biol **27**: 125–132

Vidal EA, Araus V, Lu C, Parry G, Green PJ, Coruzzi GM, Gutiérrez RA (2010a) Nitrate-responsive miR393/AFB3 regulatory module controls root system architecture in Arabidopsis thaliana. Proc Natl Acad Sci U S A **107**: 4477–82

Vidal EA, Moyano TC, Riveras E, Contreras-López O, Gutiérrez RA (2013) Systems approaches map regulatory networks downstream of the auxin receptor AFB3 in the nitrate response of Arabidopsis thaliana roots. Proc Natl Acad Sci U S A **110**: 12840–5

Vidal EA, Tamayo KP, Gutierrez RA (2010b) Gene networks for nitrogen sensing, signaling, and response in Arabidopsis thaliana. Wiley Interdiscip Rev Syst Biol Med **2**: 683–93

Wang C, Marshall A, Zhang D, Wilson ZA (2012) ANAP: an integrated knowledge base for Arabidopsis protein interaction network analysis. Plant Physiol **158**: 1523–33

Wang J, Xu L, Wang E, Huang S (2010) The potential landscape of genetic circuits imposes the arrow of time in stem cell differentiation. Biophys J **99**: 29–39

Wang K, Phillips CA, Rogers GL, Barrenas F, Benson M, Langston MA (2014) Differential Shannon entropy and differential coefficient of variation: alternatives and augmentations to differential expression in the search for disease-related genes. Int J Comput Biol Drug

Des **7**: 183

Wang R, Tischner R, Gutiérrez RA, Hoffman M, Xing X, Chen M, Coruzzi G, Crawford NM (2004) Genomic Analysis of the Nitrate Response Using a Nitrate Reductase-Null Mutant of Arabidopsis. Plant Physiol **136**: 2512–2522

Wang YXR, Huang H (2014) Review on statistical methods for gene network reconstruction using expression data. J Theor Biol **362**: 53–61

Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, et al. (2014) Determination and inference of eukaryotic transcription factor sequence specificity. Cell **158**: 1431–1443

Wetterstrand KA (2016) DNA Sequencing Costs: Data from the NHGRI Large-Scale Genome Sequencing Program. www.genome.gov/sequencingcostsdata Accessed [4 September 2016]

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S, et al. (2008) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res **36**: 13–21

Wu Y, Liu M, Zheng WJ, Zhao Z, Xu H (2012) Ranking gene-drug relationships in biomedical literature using Latent Dirichlet Allocation. Pac Symp Biocomput 422–33

Yan J, Risacher SL, Shen L, Saykin AJ (2017) Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. Brief Bioinform 1–12

Yi X, Du Z, Su Z (2013) PlantGSEA: a gene set enrichment analysis toolkit for plant community. Nucleic Acids Res **41**: 98–103

Yon Rhee S, Wood V, Dolinski K, Draghici S (2008) Use and misuse of the gene ontology annotations. Nat Rev Genet **9**: 509–515

Zanin M, Alcazar JM, Carbajosa JV, Paez MG, Papo D, Sousa P, Menasalvas E, Boccaletti S (2014) Parenclitic networks: uncovering new functions in biological data. Sci Rep **4**: 5112

Zuberi K, Franz M, Rodriguez H, Montojo J, Lopes CT, Bader GD, Morris Q (2013) GeneMANIA prediction server 2013 update. Nucleic Acids Res **41**: W115-22