Escuela de Ingeniería Civil en Bioinformática.

Identificación de genes reguladores involucrados en la respuesta a estrés salinos en dos portainjertos con respuestas contrastantes mediante el análisis de datos de secuenciación masiva.

Dámariz Gonzalez Zuñiga

Matrícula: 2016430014

Profesor informante: Dr. Mauricio Arenas

Tutor: Ariel Salvatierra

Co-Tutor: Patricio Mateluna

Memoria de título, Ingeniería civil en bioinformática

Semestre 2-2021

CONSTANCIA

La Dirección del Sistema de Bibliotecas a través de su unidad de procesos técnicos certifica que el autor del siguiente trabajo de titulación ha firmado su autorización para la reproducción en forma total o parcial e ilimitada del mismo.

Talca, 2022

Índice de Figuras

Índice de Tablas

1.Resumen.

La salinización de los suelos empleados para cultivo resulta ser un factor que influye negativamente en el crecimiento y desarrollo de organismos vegetales con valor comercial. Variadas herramientas se han testeado para plantear una solución válida a este problema, ya que la salinización de suelos tiende a afectar estructuralmente las células vegetales, además de incidir negativamente en el proceso de fotosíntesis. Una de las soluciones agronómicas más efectivas para hacer frente a esta problemática resulta en el desarrollo de organismos vegetales bimembres constituidos por un injerto con características productivas deseables y por un portainjerto (raíces) derivado de plantas tolerantes a concentraciones elevadas de sal en el suelo donde se ubican. El entender qué procesos biológicos subyacen a la tolerancia a estreses abióticos, tales como el estrés salino, es un factor clave para guiar el proceso de mejoramiento genético de portainjertos. Las técnicas de secuenciación masiva se han vuelto una de las metodologías usadas para descifrar estos procesos. En este trabajo se analizaron 48 librerías de secuenciación masiva RNA-seq de dos portainjertos derivados del género *Prunus* (*Prunus avium L. y Prunus cerasifera x munsoniana* con respuesta contrastante frente a estrés salino mediante el uso de herramientas bioinformáticas para la determinación de genes reguladores. Se encontraron genes codificadores de factores de transcripción a las seis horas en *Prunus cerasifera x munsoniana* que no se encuentran en *Prunus cerasifera,* los cuales interactúan directamente con proteínas membranales. También se encontraron genes codificadores de proteínas clave para la respuesta en primera fase ante estrés salino en *Prunus cerasifera x munsoniana* que no se encuentran en *Prunus avium.*

Abstract

The salinization of crop soils is a factor that negatively influences the growth and development of plant organisms with commercial value. Various tools have been tested to propose a valid solution to this problem, since the salinization of soils tends to structurally affect plant cells, in addition to negatively affecting the photosynthesis process. One of the most effective agronomic solutions to face this problem results in the development of two-membered plant organisms made up of a graft with desirable productive characteristics and a rootstock derived from plants that are tolerant to high concentrations of salt in the soil where they are located. Understanding which biological processes underlie tolerance to abiotic stresses, such as salt stress, is a key factor in guiding the rootstock breeding process. Massive sequencing techniques have become one of the methodologies used to

decode these processes. In this work, 48 RNA-seq libraries of two rootstocks derived from the genus Prunus (Prunus avium L and Prunus cerasifera x munsoniana) with contrasting response to salt stress were analyzed using bioinformatics tools with the aim of determinate regulatory genes. Genes encoding transcription factors were found at six hours in Prunus cerasifera x munsoniana that are not found in *Prunus avium*, which interact directly with membrane proteins. Key protein coding genes for the first-phase response to salt stress were also found in Prunus cerasifera x munsoniana that are not found in *Prunus avium*.

2. Introducción.

2.1. La respuesta de organismos vegetales ante estrés abiótico está mediada a partir de la regulación de genes específicos.

La expresión génica es el proceso que permite a todos los organismos vivos transformar la información codificada por los ácidos nucleicos en proteínas necesarias para el desarrollo, funcionamiento y reproducción de los individuos. El control de la expresión génica es un proceso esencial para la adaptación de cualquier organismo a las condiciones ambientales en las que se encuentra.

Muchos procesos vegetales dependen de la expresión génica diferencial de la información almacenada en los ácidos nucleicos. Esta diferencia suele estar controlada por proteínas complejas llamadas factores de transcripción (FT), los cuales aumentan o disminuyen las tasas de transcripción de otros genes. Los factores de transcripción son proteínas reguladoras que se unen a motivos específicos de ADN y, a través de interacciones específicas proteína-proteína, logran transmitir señales a la maquinaria transcripcional basal donde se encuentran las ARN polimerasas, lo que resulta en tasas particulares de expresión génica. El proceso comienza cuando las células, en este caso vegetales, perciben una señal de estrés, seguido de una respuesta rápida que transduce la señal externa en señales intracelulares a través de fitohormonas. De esta forma, las cascadas de señales que involucran moléculas o iones intracelulares se activan junto con cascadas de quinasas o fosfatasas, y los factores de transcripción responderán a la alza o baja de estas proteínas, uniéndose a los elementos reguladores relacionados con el estrés para inducir o suprimir la expresión. Si bien este mecanismo de regulación no es único y a nivel celular existe una variedad de procesos que regulan la expresión de genes, para fines del estudio nos enfocaremos en el recientemente mencionado *(Bianchi et al. 2015)*.

Aproximadamente el 5-7% del genoma de un organismo eucariota codifica para FT, los cuales se pueden agrupar en 50-60 grupos de familias.

En *Prunus persica,* se han identificado 1533 factores de transcripción que representan aproximadamente el 5,5% de los 27852 genes que codifican para proteínas en esa especie.

Estos FT se utilizan como referencia para el resto de especies *Prunus,* donde se han realizado variados estudios sobre el análisis de expresión génica diferencial en rasgos agronómicos incluyendo el control del proceso de floración, calidad del fruto, resistencia a estrés biótico y abiótico. La sequía, salinidad y bajas temperaturas se consideran los estreses abióticos que más limitan la producción y la calidad de la fruta en especies de *Prunus (Bianchi et al. 2015).* Existe una variedad de factores de transcripción caracterizados en la literatura para la especie *Prunus* que participan activamente en la respuesta ante estrés abiótico y algunos se nombran a continuación:

- NAC: estudios funcionales han demostrado que los genes NAC pueden ser inducidos por diferentes tipos de estreses abióticos. En la especie *Arabidopsis,* la inducción de la expresión del gen ANAC072 por sequía y ABA y ANAC019 por salinidad han sido demostrada en variados estudios. Además, la sobreexpresión de estos tres genes resultó en una tolerancia mejorada ante la sequía. Los genes NAC también regulan respuestas ante patógenos. Las proteínas NAC se caracterizan por un dominio NAC de unión a ADN altamente conservado de aproximadamente 150 aminoácidos en la región N-terminal. Este dominio es transmembranal y puede activar o reprimir la transcripción *(Baillo et al. 2019)*.
- Myb: Conocidos como factores de señalización activos del estrés abiótico, actúan posteriormente a la señal de estrés abiótico regulando genes y actúan tanto a nivel transcripcional como post-transcripcional. Se une al motivo 5'-AACGG-3' en promotores de genes *(Baillo et al. 2019)*.
- AP2/ERF : Se conocen como proteínas de unión a elementos sensibles a la deshidratación o factor en respuesta al etileno. La familia AP2/ERF incluye a factores inducibles por estrés y se sabe que muchos de estos genes se involucran en este tipo de respuesta. Estos factores de transcripción se unen al elemento sensible al etileno a través de una caja GCC (AGCCGCC) en el elemento regulatorio cis del gen diana *(Baillo et al. 2019)*.
- WRKY: Son una familia amplia de proteínas de unión a ADN los cuales se unen a W-box con motivo TTGACC/T en el promotor de genes que están río abajo en el ADN. Es uno de los dominios más conservados que se une a W-box y participa activamente en la respuesta a estrés abiótico en distintas especies *(Weixing Li et al. 2020)*.

2.2. Uso de portainjertos para aumentar la tolerancia de organismos vegetales ante estrés salino.

La respuesta en primera fase que se genera en organismos vegetales ante el estrés salino, incluyendo a los portainjertos Marianna 2624 y Mazzard F12/1, induce la captación de NA^+

hacia el medio intracelular, lo que resulta en una disrupción de la homeostasis del NA^+/K^+, razón por la cual la salinidad de los suelos influye en la productividad de las hortalizas y en el crecimiento de las plantas de interés comercial tales como cerezo y ciruela, variedades altamente exportadas desde el país.

La salinización de los suelos es un fenómeno que se debe principalmente, pero no exclusivamente, a dos factores: la existencia de trazas de cloruro de sodio en aguas utilizadas para actividades de regadío y la afluencia de agua salada hacia reservas y/o canales utilizados para el mismo fin.

Una de las soluciones que se han planteado para la problemática de tierras de cultivo afectadas por altas concentraciones de sal en los suelos es el uso de genotipos de plantas tolerantes a salinidad como portainjertos para incrementar la tolerancia de los cultivos a esta condición ambiental.

Estudios recientes han examinado la respuesta de cultivos injertados ante el estrés salino donde se ha demostrado que esta técnica resulta ser una herramienta útil para aumentar la tolerancia a estrés salino en organismos vegetales *(Yan et al. 2018).* Dichos estudios demuestran cómo organismos vegetales resultantes de la unión de un injerto y un portainjerto tolerante al estrés salino puede incidir en respuestas de reconfiguración de estructuras celulares, capacidad fotosintética y acumulación de peróxido de hidrógeno y otros radicales libres.

Un número importante de especies frutales y hortalizas se producen como plantas injertadas, las cuales se componen de dos plantas, una aportando el sistema de raíces y la otra la parte aérea. Un portainjerto aporta el sistema de raíces y sostiene al injerto. El injerto, la parte aérea, se define como un organismo vegetal donde una porción de tejido procedente de una planta —la variedad o injerto propiamente dicho— se une sobre otra ya asentada (portainjerto), de tal modo que el conjunto de ambos crezca como un solo organismo *(Goldschmidt 2014)* como se muestra en la Figura 1:

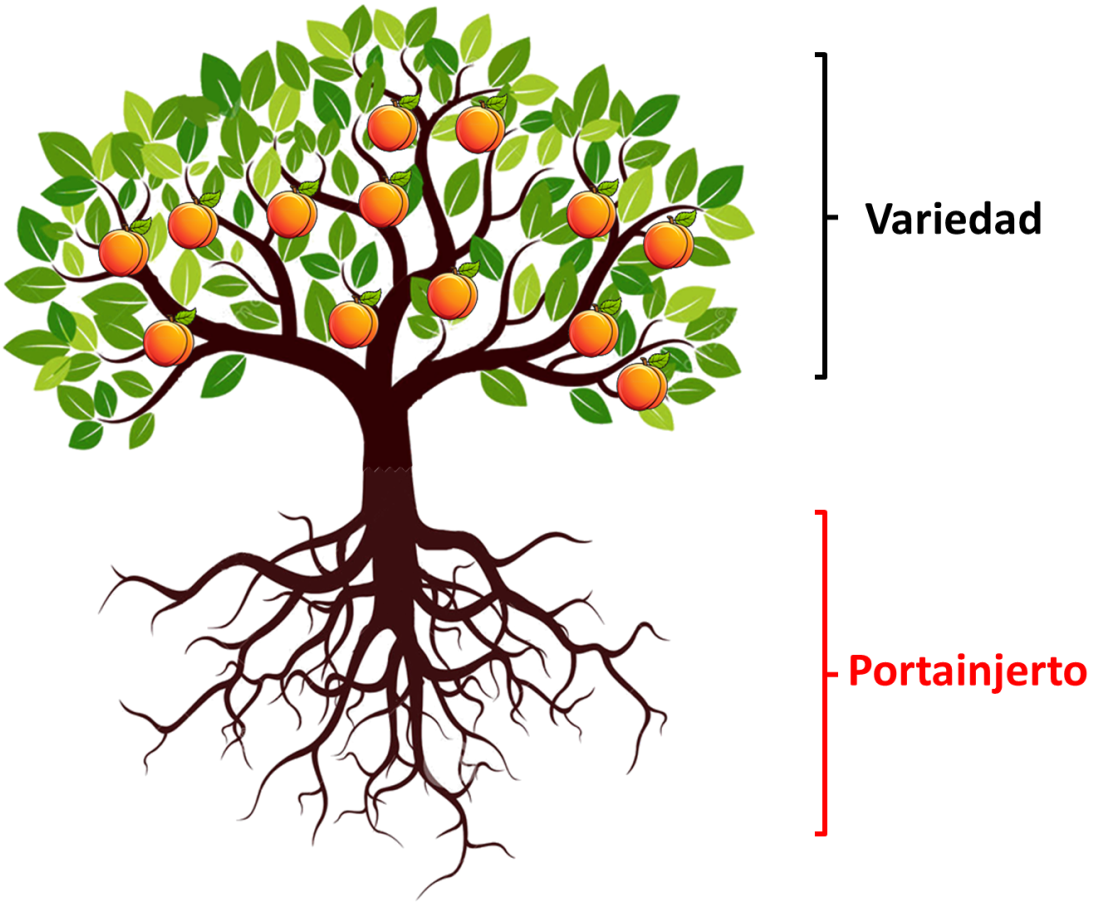


Figura 1: Esquema portainjerto. En la sección aérea se ubica el injerto soportado por el portainjerto correspondiente a la parte de la raíz.

En otras palabras, el portainjerto aporta la sección basal que incluye el sistema radical y al menos una porción de tallo. El injerto, constituido por una yema o por un pequeño esqueje con varias yemas de otra planta conformará la copa o parte superior del nuevo ejemplar, con sus ramas, hojas, flores y frutos. De la unión del injerto con el portainjerto se obtiene una planta compuesta de dos secciones provenientes de individuos distintos, que mostrará un comportamiento particular. En efecto, el portainjerto y el injerto mantienen su individualidad, sin que se produzca intercambio o mezcla de información genética; más aún, ambos miembros o secciones pueden ser bastante diferentes entre sí desde el punto de vista genético. Sin embargo, ambos componentes ejercerán una influencia recíproca, modulada a su vez por el ambiente *(Nawaz et al ., 2016)*

En específico, los portainjertos que se usarán en esta investigación corresponden a plantas de cerezo Mazzard F12/1 (*Prunus avium*), que genera árboles adultos de 6 metros o más, es compatible con la mayoría de las variedades de cerezo y resulta ser bastante resistente al frío aunque sensible a asfixia radical y estrés por déficit hídrico. Por otro lado, el estudio incluye al portainjerto Marianna 2624, que corresponde a un genotipo derivado de ciruelo,

híbrido entre las especies *Prunus cerasifera* y *Prunus munsoniana* la cual presenta gran vigor y una gran tolerancia a suelos pesados, húmedos y asfixia radical *(Foundation Plant Services s. f.*).

Estos 2 portainjertos tienen genomas con un porcentaje de similitud alto, aproximadamente de un 60%, pero existen diferencias en torno a la forma en la que estas se adaptan a distintos tipos de estrés ambiental.

En el año 2014, el Centro de Estudios Avanzados en Fruticultura (CEAF) realizó un análisis comparativo entre respuestas de distintas especies de *Prunus* ante estrés por hipoxia, donde se pudieron visualizar diferencias acentuadas entre algunas especies y su reacción ante este tipo de estrés abiótico, siendo las más contrastantes *Prunus avium L (*Mazzard F12/1, sensible) y *Prunus cerasifera x Prunus munsoniana* (Marianna 2624, tolerante) *(Pimentel et al. 2014)*. Las especies de *Prunus* son consideradas, en general, como sensibles ante el estrés por hipoxia, pero aún así existen grados distintos de tolerancia a este tipo de condición ambiental dentro de las especies de este género botánico, afirmación que podría extrapolarse para otro tipo de estrés abiótico, en este caso, estrés salino *(Toro et al., 2021)*, donde se ha observado que los dos portainjertos con respuestas contrastantes ante condiciones de hipoxia presentan igualmente respuestas contrastantes ante altas concentraciones de sal en el suelo de cultivo, por lo que se decidió utilizar estos dos genotipos para el objetivo de la investigación, donde se espera aplicar técnicas de análisis expresión diferencial de genes en conjunto con construcción de redes regulatorias de genes para descubrir los factores genéticos que inducen respuestas diferenciales ante estrés salino en los portainjertos Mazzard F12/1 (*Prunus avium L*.) y Marianna 2624 (*Prunus cerasifera x munsoniana W. Wight & Hedrick),* análisis que resulta clave para conocer los componentes que inciden en la respuesta diferencial que existe entre ambos.

2.3. La cantidad de ARN se puede cuantificar y representar gráficamente.

La abundancia de productos como ARN mensajeros (transcritos) generados por la síntesis de un gen específico se define como el nivel de expresión de un gen, indicador que luego puede ser constatado cuantificando la abundancia de productos funcionales resultantes de la transcripción de la molécula de ARN mensajero *(Park et al., 2012).*

Estudiar los distintos niveles de expresión de los genes que comprenden un organismo, se ha hecho esencial para estudios de distintas aplicaciones de interés comercial en sistemas biológicos. Estudios de factores genéticos que inducen a enfermedades congénitas, estudios poblacionales y evolutivos, conocer factores de tolerancia y resistencia de organismos vegetales ante estrés ambiental, se basan en estudios de niveles de expresión *(Costa-Silva et al., 2017).*

Los niveles de expresión de un gen en un organismo suelen ser plásticos y flexibles, ajustándose a los requerimientos ambientales que existan en el medio dentro del cual el individuo se desarrolla. Los estudios de expresión diferencial de genes se basan en la comparación de individuos bajo distintas condiciones ambientales (por lo general una condición control y la condición que se quiere estudiar). Los individuos que se utilizan para este análisis suelen ser de la misma especie, sexo y edad, desde donde se extraen muestras de ARN en distintos puntos del tiempo para conocer los niveles de expresión de genes que existen en distintas condiciones *(Censi et al. 2010)*. Para realizar el procedimiento de análisis de expresión diferencial de genes resulta conveniente construir la librería para el proceso de secuenciación a partir de secuencias de RNA. Esta técnica se conoce como RNA-seq la cual comprende cualquier herramienta de Next Generation Sequencing (NGS) asociada al trabajo con secuencias de ácido ribonucleico *(Chu y Corey 2012)*. Al utilizar esta técnica, se extraen muestras de secuencias de ARN las cuales resultan de la transcripción de genes codificados en el ADN, cuya cuantificación por locus resulta ser un indicador del nivel de expresión de un gen.

Una vez extraídas las moléculas de RNA de los individuos, estas son secuenciadas para transformar la información experimental a datos que reflejen las lecturas de bases nitrogenadas que existen en las secuencias de ARN de los individuos.

Este tipo de secuenciación se utiliza en muchos ensayos cuantitativos, donde las moléculas de ARN de un sistema biológico reflejan la actividad transcripcional de las moléculas de ADN. Las lecturas de ARN obtenidas en la secuenciación se asignan a un gen a través del proceso de mapeo, donde se ubican dichas lecturas en regiones específicas del genoma correspondiente al organismo objetivo *(Conesa et al. 2016).* Este procedimiento se realiza a través de alineamientos de secuencias dilucidando a qué gen corresponde el transcrito de ARN que refleja la lectura.

Una estadística de resumen importante del alineamiento de reads a genomas de referencias resulta ser el número de lecturas que existen para un gen, exón, secuencia codificante, etc. En estudios de RNA-seq, se ha encontrado que este conteo tiene una relación lineal con la abundancia de la transcripción en células diana y en procesos metabólicos específicos. La idea es comparar los recuentos de lectura entre diferentes condiciones biológicas *(Anders, Pyl, y Huber 2015)* .

De esta forma, un gen se declarará expresado diferencialmente si una diferencia o cambio observado en los recuentos de lectura o los niveles de expresión entre dos condiciones experimentales es estadísticamente significativo *(Anjum et al. 2016)*.

Una vez que se obtienen los conteos de las lecturas de ARN, se tiene una medida de la actividad de todos los genes conocidos de un organismo. El conjunto del conteo de genes y su respectiva función biológica se define como un perfil de expresión. A partir del perfil de

expresión se pueden realizar análisis más profundos sobre cómo los genes se modulan unos a otros dependiendo de los requerimientos ambientales en los cuales se encuentra el organismo biológico que funciona como muestra. Al agrupar transcritos con perfiles de expresión similares frente a una condición es posible establecer redes de regulación génica. Genes pertenecientes a una misma sub-red muy probablemente son parte de una o más vías metabólicas que responden coordinadamente a la condición en estudio.

Además, a partir de la construcción de Redes Regulatorias de Genes, es posible describir y predecir dependencias entre entidades moleculares (genes, proteínas, metabolitos o ARN) que están representados por nodos unidos entre sí de acuerdo con sus interacciones *(Emmert-Streib, Dehmer, y Haibe-Kains 2014)*.

Estas aproximaciones analíticas resultan ser herramientas muy útiles a la hora de investigar cómo los genes se coordinan e interactúan entre sí para modular su propia expresión y gatillar la adaptación biológica a una determinada condición ambiental.

2.4. Redes Regulatorias de Genes.

La regulación de genes consiste en un amplio rango de mecanismos usados por la célula para incrementar o disminuir la síntesis de productos génicos específicos (RNA o proteínas). Entre los mecanismos utilizados, la célula puede regular la tasa de transcripción de algún gen, regular el procesamiento de las moléculas de RNA a partir de splicing alternativo o en su defecto regular la tasa de traducción de ciertos transcritos *(Atkinson y Halfon 2014)*. Estos procesos que están acoplados a los requerimientos celulares se entienden como la regulación de genes y se puede representar gráficamente a partir de "redes" con herramientas computacionales.

Las redes de regulación génica funcionan a partir de abstracciones de los sistemas biológicos. Esta abstracción consiste en el proceso de aislar conceptualmente una propiedad o función concreta de un gen para luego poder relacionarla a otras propiedades, generando una representación gráfica de estas reacciones conjuntas. Este tipo de representaciones consisten en una colección de segmentos de ADN pertenecientes a una célula, segmentos que interactúan entre sí a través de ARN y proteínas *(Davidson y Levin 2005).*

El desarrollo de esta técnica nace de la necesidad que existe de medir y cuantificar las interacciones moleculares que existen en las células. Esto resulta ser un proceso muy difícil debido a todos los elementos que deben tomarse en cuenta cuando se estudia un organismo celular completo, por lo que medir los componentes celulares por separado resulta ser una tarea mucho más accesible, enfocando el estudio en la cuantificación y abundancia de los productos funcionales del ADN. Gracias a este enfoque en las últimas

décadas se han desarrollado distintos softwares para mediciones a gran escala de componentes celulares lo que ha llevado a la reconstrucción computacional de las estructuras de interacción a partir de patrones de expresión de genes *(Huynh-Thu y Sanguinetti 2019).*

Gráficamente, una red de regulación génica se visualiza como un sistema de nodos unidos por arcos, donde los nodos representan genes y los arcos representan las interacciones que existen entre estos. Esta representación gráfica se va configurando a partir de procesos de asociación de inteligencia artificial llevados a cabo por los programas computacionales que construyen estas redes. Las redes regulatorias tienen como elementos nodos y arcos. Una de las características más importantes del nodo, representante de un gen, corresponde al grado, concepto que representa la cantidad de arcos que van unidos al nodo *(Huynh-Thu y Sanguinetti 2019).* Los nodos estarán conectados entre sí sólo si existe alguna relación entre los genes que representan el nodo en los tejidos elegidos para el análisis *(Zhang y Horvath 2005).*

Para construir una red de regulación génica lo primero que se necesita es definir perfiles de expresión génica, donde se mide la cantidad de genes que están siendo expresados en condiciones específicas, como se mencionó anteriormente. Los perfiles de expresión de genes miden la actividad transcripcional de estos en un contexto específico y de esta forma el perfil puede construirse tanto para una célula completa como para vías metabólicas específicas, lo cual dependerá del objetivo de la investigación *(Microarrays Factsheet 2007)*.

Por lo general, el perfil de expresión génica corresponde a la cuantificación de transcritos que existen de un gen específico en la muestra de un organismo bajo condiciones también específicas.

El perfil puede ser construido a partir de un análisis de expresión diferencial para dilucidar qué genes cambian el nivel de expresión en productos proteicos(ya sea aumentando o disminuyendo) y qué genes se mantienen constantes cuando comparamos muestras de distintas condiciones ambientales y puntos en el tiempo. Una vez que se obtienen los perfiles, es necesario determinar una medida de similitud entre los perfiles de cada muestra, medida que refleja el nivel de concordancia que existe entre los perfiles de expresión. Cuando ya se selecciona esta medida de similitud, se construye una matriz de similitud para luego convertirla a una matriz de adyacencia, la que codificará para las conexiones que existirán entre los genes y qué tan fuertes o débiles resulten ser *(Zhang y Horvath 2005).*

2.5. Problemática.

La salinidad de los suelos es un fenómeno que afecta la producción de alimentos a escala mundial, ya que acarrea consigo procesos de degradación de los suelos *(Lamz Piedra y González Cepero 2013)*, perjudicando los rendimientos de los cultivos. Esta degradación de los suelos causa una disminución de la capacidad de los organismos vegetales para absorber agua y por ende, cuando el organismo vegetal absorbe trazas de cloruro de sodio en grandes cantidades por las raíces, el crecimiento de la planta se ve afectado negativamente debido a una alteración en los procesos metabólicos generando una disminución en la tasa fotosintética (*Tavakkoli, Rengasamy, y McDonald 2010*).

Además de afectar la actividad fotosintética, el estrés salino tiende a generar estrés osmótico en las células vegetales, lo cual puede llevar a pérdidas por turgencia debido a que las proteínas de membranas disminuyen su actividad o se desnaturalizan provocando desorganizaciones en las membranas biológicas *(Tavakkoli, Rengasamy, y McDonald 2010).*

La salinización de los suelos es un proceso que muestra una tendencia a aumentar, afectando en el presente a más de 50 millones de hectáreas alrededor del mundo *(Lamz Piedra y González Cepero 2013)*. Aproximadamente un 7% de las tierras de regadío son afectadas por la salinidad del suelo de un total del 20% de las zonas de cultivo alrededor del mundo *(Deinlein et al. 2014)*. Latinoamérica ocupa el 16% de los suelos con tendencia a salinización como se muestra en la Figura 2.

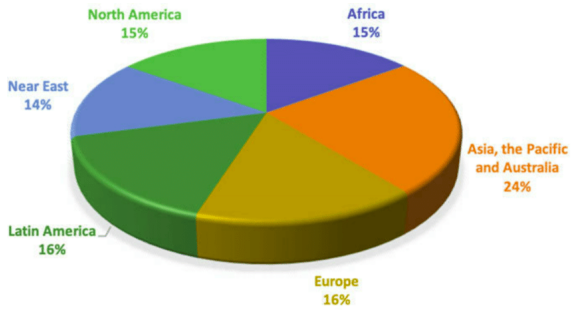


Figura 2: Distribución de suelos con tendencia a salinización alrededor del mundo *(Srivastava et al., 2019).*

En Chile existen 435.991 km2, desde la región del Bío-Bío al norte y equivalentes al 58% de la superficie nacional, bajo condición de déficit hídrico, por tanto susceptibles a procesos de salinización *(Belmar, 2016).*

El contexto actual se ve empeorado dadas las condiciones de cambio climático en la cual las altas temperaturas afectan directamente la capacidad que tiene la planta de tolerar altas concentraciones de sal. La mayoría de los cultivos pueden tolerar ciertos niveles de estrés a la salinidad si el clima es frío y húmedo en comparación con uno cálido y seco *(Pulido Madrigal, 2016)*. A partir de esta situación, cada vez más cultivos se encuentran expuestos a experimentar pérdidas a nivel productivo, incluyendo especies de cerezas altamente comercializadas a nivel país, las cuales figuran en el primer lugar de especies exportadas en Chile, como se muestra en la Figura 3.

El nivel de tolerancia de alguna especie de cultivo hacia concentraciones de sal se mide a partir de la *Conductividad Eléctrica medida en el extracto de saturación CEes (Tolerancia de los cultivos a la Salinidad, s. f.)*, a partir de la cual se establece que el cerezo, árbol proveedor de una de las principales frutas exportadas en el país, la cereza, como cultivo leñoso se caracteriza como sensible a concentraciones de sal con un CEes de 1,5 S*dS/m *(Maas & Hoffman, 1977)*.

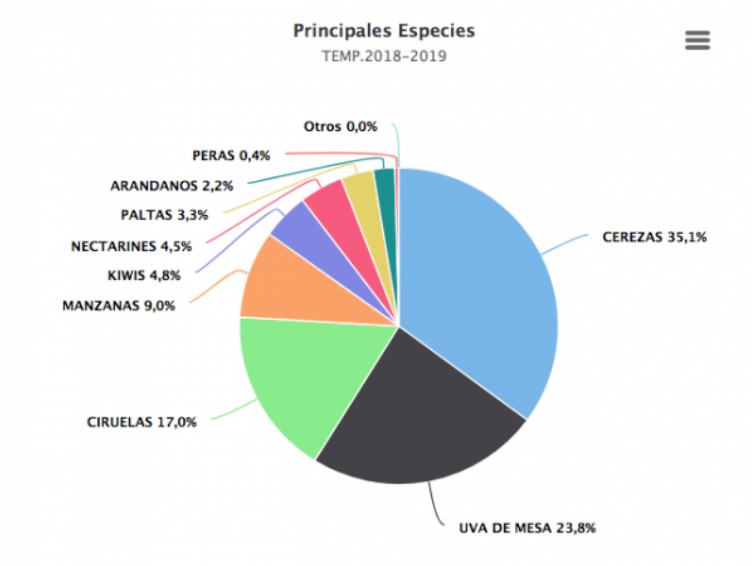


Figura 3: Principales especies exportadas desde Chile a China. Se puede visualizar que la cereza es la que más se exporta, siendo Chile el principal proveedor de frutas templadas,incluyendo ciruelas y cerezas *(Simfruit, 2020, s.f).*

Es por esto que se hace necesario la aplicación de técnicas utilizadas para la mejora de organismos vegetales con el fin de mantener los niveles productivos de cultivos con un interés comercial.

2.6. Estado del Arte.

En lo que concierne a información sobre la respuesta de plantas al estrés salino, hasta ahora se reconocen 2 estrategias principales que los organismos utilizan para poder sobrellevar el estrés salino: evasión del estrés y tolerancia al estrés. En la evasión del estrés las plantas se desarrollan y crecen en momentos ambientales favorables para evitar el estrés ambiental. Por otro lado, en la estrategia de tolerancia al estrés las plantas completan su ciclo vital incluso en condiciones ambientales adversas gracias a la activación de mecanismos de adaptación dados por la actividad transcripcional adaptativa que implica la sobre-expresión o represión de genes *(Krasensky y Jonak 2012)*.

En lo que respecta a la confección de redes regulatorias relacionadas a organismos vegetales y su respuesta a estrés ambiental, se realizó un análisis comparativo entre datasets de mRNA de los organismos modelo *Arabidopsis thaliana* y *Oryza sativa*, donde se pudo visualizar que existe una complejización de las redes regulatorias de genes en muestras de RNA correspondientes a muestras extraídas desde condiciones de estrés ambiental. Específicamente, para el estrés salino, tanto Arabidopsis como el arroz presentan activación de rutas y respuestas distintas tanto para regulaciones transcripcionales como post-transcripcionales, desde donde se puede deducir que las redes regulatorias de genes que se generan pueden variar de organismo en organismo. Por ejemplo, en el caso del arroz la respuesta es más lenta y se enfoca en desarrollar mecanismos de crecimiento y desarrollo que eviten el estrés salino, mientras que Arabidopsis activa genes para adaptar su metabolismo para defenderse de los efectos deletéreos de altas concentraciones de sal *(X. Wang et al. 2020).*

También se han hecho estudios de redes regulatorias de genes y su configuración en plantas frente al estrés por calor, debido a que las altas temperaturas ocasionadas por el calentamiento global tienden a tener un impacto negativo en el proceso fotosintético de los organismos vegetales, afectando la producción de ATP y energía. Las respuestas de los organismos vegetales ante cualquier estrés ambiental representan una reacción en cadena de muchos genes que se regulan y modulan unos a otros *(Krasensky y Jonak 2012).*

4. Hipótesis:

Se plantea que la respuesta del genotipo tolerante Marianna 2624 ante estrés salino está caracterizada por la expresión de una red de genes, la cual no se configura en el genotipo sensible Mazzard F12/1 sometido al mismo nivel de salinidad.

5.Objetivos.

5.1. General.

Estudiar los genes asociados a la respuesta a estrés salino en librerías RNA-Seq de raíces de Mazzard F12/1 (sensible) y Mariana 2624 (tolerante) usando redes regulatorias de genes.

5.2. Objetivos específicos.

- Establecer los transcriptomas asociados a la respuesta a estrés salino mediante método *de novo.*
- Identificar genes diferencialmente expresados y su anotación GO y KEGG.
- Analizar los patrones de reconfiguración transcriptómica mediante la construcción de redes regulatorias.

6.Materiales:

- Visualización de estadísticos con respecto a la calidad de los archivos fastqc.
  FASTQC: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
- Filtrar bases y reads de baja calidad.
  AfterQC: https://github.com/OpenGene/AfterQC
- Ensamble de novo para genomas de portainjertos utilizados.
  Trinity: https://github.com/trinityrnaseq/trinityrnaseq/wiki
- Disminución de redundancia de secuencias en el ensamble.
  TransDecoder: https://github.com/TransDecoder/TransDecoder/wiki
  CD-HIT: http://weizhongli-lab.org/cd-hit/
- Mapeo de *reads* a genoma y procesamiento de archivos de alineamiento.
  HISAT2: http://daehwankimlab.github.io/hisat2/
  Samtools: http://www.htslib.org/
- Preparación de archivos para análisis de expresión diferencial en R.
  Stringtie:https://ccb.jhu.edu/software/stringtie/
- Transformación de IDs de secuencias generados por Trinity a nombres generados por Stringtie.
  GFFREAD: https://github.com/gpertea/gffread
- Análisis de expresión Diferencial.
  R : https://www.r-project.org/
  DeSeq2: https://bioconductor.org/packages/release/bioc/html/DESeq2.html
- Anotación de genes expresados diferencialmente:
  Perl
  HMMER2GO: https://github.com/sestaton/HMMER2GO
- Construcción de redes regulatorias
  Cytoscape: https://cytoscape.org/
  Genie3: https://bioconductor.org/packages/release/bioc/html/GENIE3.html
  Blast2Go: https://www.blast2go.com/

7.Metodología.

Para comenzar a realizar los siguientes procedimientos , se utilizaron secuencias de RNA-seq extraídas de los portainjertos Mazzard F12/1 (*Prunus avium)* y Marianna 26 (*Prunus cerasifera x munsoniana*) por el equipo de trabajo del Centro de Estudios Avanzados en Fruticultura (CEAF). Las muestras de RNA-seq extraídas de cada portainjerto conformaban librerías paired-end para 4 puntos distintos en el tiempo, correspondientes a 0 horas, 6 horas, 3 días y 14 días. Cada librería perteneciente a un tiempo específico contiene 3 réplicas biológicas con su determinada muestra control para cada réplica. A continuación se muestra un esquema del flujo de trabajo realizado y por realizar del proyecto en la Figura 4:

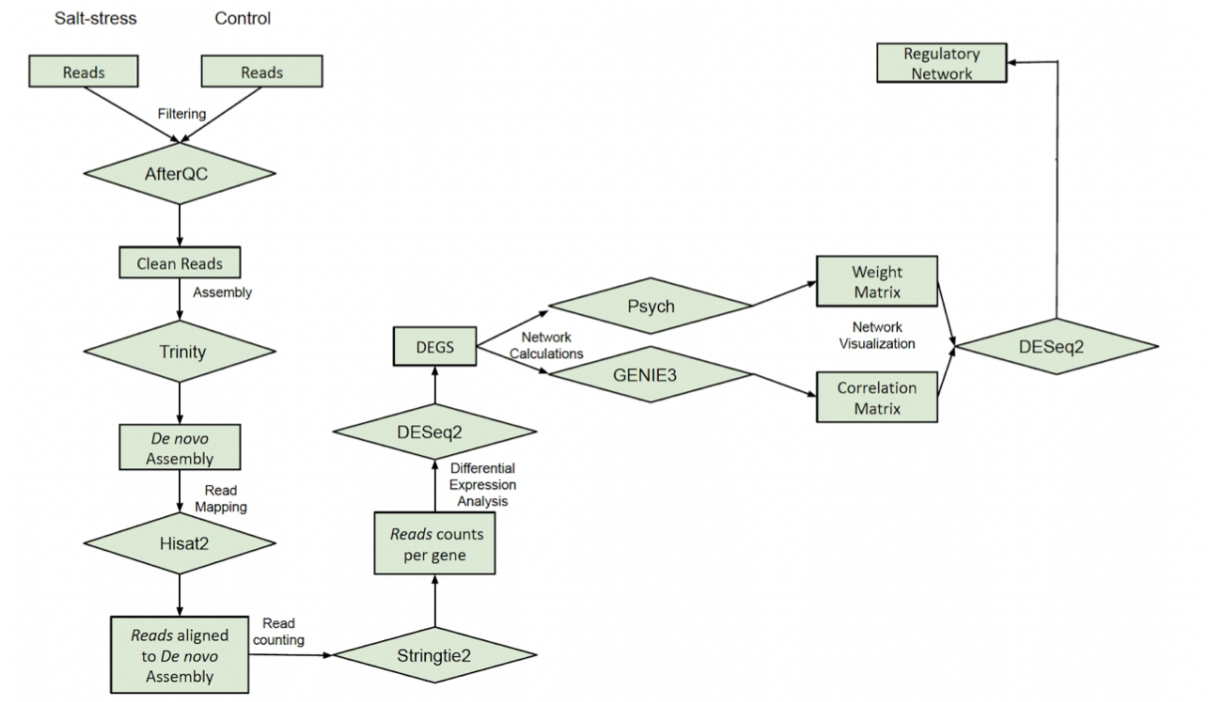


Figura 4: Workflow diseñado para el Proyecto.

7.1. Evaluación de calidad con software FASTQC.

Para evaluar la calidad de los *reads* contenidos en los archivos fastqc de las librerías de RNA-seq se utilizó el software FASTQC, el cual entrega estadísticos con respecto a calidad de composición de los *reads*. FastQC es una aplicación que permite a los usuarios realizar numerosas verificaciones de control de calidad en datos de secuencia sin procesar generados por *pipelines* de secuenciación de alto rendimiento como Illumina y ABI SOLiD

(Wingett y Andrews, 2018). Para este procedimiento es necesario instalar el programa *FastQC* y luego cargar los archivos.

Cuando hablamos de calidad nos referimos a la probabilidad de que una base nitrogenada haya sido llamada de forma incorrecta dentro del read *(Sequencing Quality Scores s. f.)*. Un puntaje Q30 indica que la base puede estar incorrecta con una probabilidad de 1/1000 *(Brown, Pirrung, y McCue 2017)*, por lo que, las bases que conforman a los reads analizados presentan un 99,9% de probabilidad de estar llamadas correctamente.Los transcriptomas fueron secuenciados por la metodología Illumina HiSeq. Bajo este método, las bases nitrogenadas que componen el read y que alcanzan el puntaje Q30, son bases con una buena calidad. Es por esto que se decidió utilizar el software AfterQC, que en este caso elimina regiones de las secuencias de los reads correspondientes a adaptadores y otro tipo de agregados a la secuencia de ADN utilizados para facilitar los procesos de síntesis que se llevan a cabo a la hora de secuenciar un genoma *(Per Base Sequence Quality s. f.).*

7.2. Limpieza de *reads* con AfterQC.

Para depurar los *reads* esta herramienta implementa funciones que perfilan errores de secuenciación y corrigen la mayoría de ellos, además de funciones de filtrado de datos y control de calidad altamente automatizadas. AfterQC analiza la superposición de secuencias emparejadas para obtener datos de secuenciación de los pares de bases de los extremos desde donde se puede detectar y cortar adaptadores. El algoritmo de AfterQC lo que hace es tomar el archivo fastqc y realizar un control de calidad para luego pasar al canal de filtrado. Este canal de filtrado lo primero que detecta son burbujas levantadas durante el proceso de secuenciación. Las burbujas corresponden a heteroduplexes compuestos por fragmentos de la biblioteca de RNA parcialmente homólogos. Estas moléculas heteroduplex contienen secuencias adaptadoras complementarias bicatenarias que flanquean inserciones no complementarias monocatenarias *(Bubble products in sequencing libraries: causes, identification, and workflow recommendations s. f.)*. El resultado es una conformación similar a una burbuja parcialmente abierta en el medio del fragmento. Una vez que se encuentran estas burbujas, se realiza un prefiltrado para perfilar datos con el contenido de bases nitrogenadas a partir de las curvas de calidad generadas. Luego, AfterQC realiza un corte basándose en estas curvas de calidad.

Los datos pueden pasar por los filtros de burbujas, filtro polyX , filtro de calidad y filtro de análisis de secuencias superpuestas. Por último, se aplica una corrección de errores para generar los informes HTML y los reads limpios *(Chen et al. 2017).*

Para esto se generó un script en bash de AfterQC de manera de automatizar el proceso para todas las secuencias de reads. No se agregaron parámetros adicionales ya que no resultó necesario cortar de manera específica los reads, sí no realizar una limpieza general para afinar resultados.

El comando que se automatizó fue el siguiente:

python2 AfterQC-master/after.py -1 Muestra_L1_1.fq.gz -2 Muestra_L1_2.fq.gz

Donde se le entrega al programa el nombre de ambos pares de archivos de *reads paired end* para la misma muestra. Se procesaron los reads pertenecientes a ambos portainjertos en scripts de automatización por separado.

7.3 Ensamblado con Trinity.

Si bien existen genomas de referencia de *Prunus avium* y otros miembros de la taxa disponibles en las bases de datos, no existe un genoma de referencia para *Prunus cerasifera x munsoniana.* Los genomas de las especies de estudio se alinean en promedio hasta un 60% versus organismos modelos o más utilizados de *Prunus* como por ejemplo el durazno. Sí se utilizara el genoma Prunus avium para ambos genotipos contrastantes, los resultados estarían sesgados hacia los patrones transcripcionales de F12, donde los reads alinearán más veces dentro del genoma que M2624. Sí se utilizara algún genoma disponible en bases de datos, se perdería información dado que restaría un 40% de bases nitrogenadas que no se alinearon con la referencia. Es por esto que se utiliza un *ensamblado de novo,* donde se este se puede entender como sí se armara un rompecabezas a partir de la concatenación de los fragmentos de ADN secuenciados, para cada uno de los genotipos por separado.

Trinity lo que hace es dividir los datos de secuenciación en muchos gráficos de Bruijn individuales, cada uno de los cuales representa distintas complejidades transcripcionales en un gen o locus específicos, para luego procesar cada uno de estos gráficos de forma independiente y extraer isoformas de empalme de longitud completa, y separar transcripciones derivadas de genes parálogos *(Grabherr et al. 2011).*

Para ejecutar *Trinity*, lo primero que se hizo fue concatenar todos los archivos de reads ya filtrados con AfterQC con el comando cat de unix shell. Los archivos fastqc se concatenan dado que *Trinity* necesita las piezas del rompecabezas, los *reads*, en un archivo único.

Una vez concatenados los archivos en formato fastq.gz para cada portainjerto, los archivos resultantes se le pasan a *Trinity* como input por separado a partir del siguiente comando, los parámetros utilizados se muestran en la Tabla 1:

Trinity --seqType fq --max_memory 100G --single reads_clean_salt.fq.gz --CPU 30 --no_normalize_readsF

Tabla 1: Parámetros utilizados para Trinity.

Parámetro utilizado	Valor	Utilidad
seqType	fq	Indica el tipo de archivo de entrada que se utilizará para el ensamble
max_memory	100G	Indica la cantidad de memoria del equipo que se utilizará para la tarea de ensamble
single	.	Indica que utilizaremos solo un archivo con todos los reads filtrados.
CPU	30	Indica las unidades de procesamiento del hardware que se utilizarán para la tarea
no_normalize_reads	.	Evita que se genere la normalización in silico por defecto que performa Trinity, donde se normalizan las lecturas con una cobertura de profundidad muy alta ya que con este proceso puede utilizar muchos recursos computacionales. Como en este caso se trabaja con el servidor de la empresa, se decide no utilizar dicha opción

Producto del comando se genera una carpeta con varios archivos de *Trinity*, donde podremos ver un archivo fasta correspondiente al genoma ensamblado de ambas variedades de *Prunus*.

7.4. Filtro de secuencias codificantes.

TransDecoder:

Dado que nuestro ensamble está construido a partir de secuencias de transcritos, es necesario identificar regiones codificantes para poder disminuir la redundancia de las secuencias. Para esto se decidió utilizar la herramienta TransDecoder. Con esto podemos generar un fasta con regiones importantes en torno a función metabólica y así reducir las regiones que no resultan del todo fundamentales para el estudio. TransDecoder identifica las posibles secuencias de codificación según los siguientes criterios *(TransDecoder/TransDecoder s. f.)*:

- Se encuentra un marco de lectura abierto (ORF) de longitud mínima en una secuencia de transcripción.
- Se calcula una probabilidad logarítmica la cual puntúa similar a la calculada por el software GeneID. El puntaje anterior es mayor cuando el ORF se califica en el

primer marco de lectura en comparación con los puntajes en los otros 2 marcos de lectura hacia adelante (TransDecoder/TransDecoder s. f.).

- Si un ORF candidato se encuentra completamente encapsulado por las coordenadas de otro ORF candidato, se informa sobre el más largo.
- Se construye / entrena / usa un PSSM para hacer más exacta la predicción del codón de inicio.

Opcionalmente, si el péptido putativo tiene una coincidencia con un dominio Pfam se incluye esta información en el resultado

En este trabajo se realizaron los siguientes procedimientos:

- Extraer marcos de lectura abierto largos

  ./TransDecoder.LongOrfs -t F12_denovo.fasta

  Donde F12_denovo.fasta resulta ser el archivo generado por *Trinity*

- Predecir las regiones codificantes más probables

./TransDecoder.Predict -t F12_denovo.fasta --retain_pfam_hits f12_sal_peptides.pfam.out

Donde se utiliza la opción retain_pfam_hits para conservar en las regiones codificantes los hits que se encuentren en Pfam.
El procedimiento se repitió para el portainjerto Marianna 26.

CD-HIT-EST:

TransDecoder genera un archivo de *coding sequences* el cual podemos utilizar luego en el programa CD-HIT-EST.

Una vez que tenemos las secuencias codificantes más probables, necesitamos reducir la redundancia de secuencias que podrían corresponder a una misma *coding sequence*. Para esto el programa CD-HIT-EST agrupa secuencias de nucleótidos que cumplen un umbral de similitud correspondiente generalmente a una identidad de secuencia (Weizhong Li y Godzik 2006). El comando utilizado es el siguiente:

cdhit-est -i F12sal_denovo.fasta.transdecoder.cds -o F12salnew_cdhit.fasta -t 1 -c 0.9 -n 9 -T 10 -M 20000

Donde se utiliza como *input* un archivo fasta pero de secuencias codificantes probables y se generan las salidas con secuencias codificantes no redundantes. El procedimiento se llevó

a cabo para Marianna 26 de la misma forma. Los parámetros utilizados se explican en la tabla continuación:

Tabla 2 : Parámetros utilizados para filtrar secuencias con CD-HIT-EST

Parámetro	Valor	Uso
-M	20000	Memoria a utilizar
-n	9	Tamaño de palabra para comenzar la búsqueda
-t	1	Tolerancia a redundancia
-T	10	Número de hebras
-c	0,9	Sequence Identity Threshold

7.5. Mapeo de reads a transcriptoma *de novo.*

Una vez que se tienen los transcriptomas *de novo* es necesario alinear los reads hacia este transcriptoma como para conocer las lecturas que tiene cada región de este y así conocer el nivel de transcripción. Para esto se utilizó el software hisat2.

Hisat2 es un software de mapeo de reads y es uno de los más rápidos actualmente.

Este programa utiliza el método de transformación Burrows-Wheler, donde se reordenan las cadenas de carácteres (en este caso A T G C) en series de carácteres similares para facilitar la comprensión del análisis (*Burrows y Wheeler s. f.*) y comprimir genomas de modo que se requiera poca memoria para almacenar. Hisat2 crea índices a partir de los genomas transformados utilizando un esquema especial de indexación FM. La indexación FM corresponde al proceso llevado a cabo por la transformación de Wheeler Borrow, donde se generan índices con sufijos específicos desde una cadena de texto. Hisat2 crea un índice global del genoma completo y decenas de miles de índices locales pequeños para hacer posible un alineamiento de empalme de las lecturas al genoma de referencia (*D. Kim, Langmead, y Salzberg 2015*), el comando se muestra a continuación:

hisat2-build F12sal_cdhit.fasta F12_salt

Y luego se llevó a cabo el siguiente comando de mapeo automatizado en un script para todas las muestras:

hisat2 -x index_F12salt -1 Muestra_1.good.fq.gz -2 Muestra_2.good.fq.gz -S Muestra.sam

Para seguir trabajando con los archivos, en el mismo script que se generó para automatizar el procedimiento de hisat2 se agregaron los comandos para samtools, programa que convierte los archivos de salida en formato .sam a formato .bam además de ordenarlos. Esto se hace debido a que los archivos en formato sam resultan utilizar mucha memoria RAM para trabajarlos, además de que los programas que utilizaremos a continuación necesitan de archivos en formato bam. Por otro lado, es necesario ordenar los archivos .bam a partir de samtools sort ya que cuando alinean archivos FASTQ con cualquier alineador de secuencia actual, los alineamientos producidos están en orden aleatorio con respecto a su posición en el genoma de referencia y los alineamientos deben ser ordenados de tal manera que los alineamientos ocurran en "orden genómico" *(samtools Tutorial s. f.)*

Los comandos incluidos fueron:

samtools view -b Muestra.sam > Muestra.bam

samtools sort -o Muestra.bam Muestra.bam

Los scripts de automatización se generaron por separado para Mazzard F12/1 y Marianna 2624.

7.6. Preparación de datos para R.

Una vez con los archivos de salida de mapeo adecuados y el genoma de referencia depurado, es necesario generar los archivos con los conteos de transcritos para poder realizar el análisis de expresión diferencial. Para esto utilizamos el software Stringtie, el cual corresponde a un ensamblador de alineamientos donde podemos generar archivos de anotación con la información contenida en los archivos de alineamiento .bam (*Pertea et al., 2020*).

Lo primero que se hizo fue generar un archivo gtf para cada una de las muestras extraídas en formato .bam:

stringtie -p 30 -o f12sal_0h_control1.gtf -l f12sal_14d_control2 ../0h-control1.bam

En este comando podemos ver que utilizamos los siguientes parámetros mostrados en la Tabla 3:

Tabla 3 : Parámetros utilizados en stringtie y su uso.

Parámetros utilizados	Valor	Uso
-p	30	Número de Hebras
-o	.	Archivo para escribir el output

-l	.	Etiqueta para el archivo

Una vez que se tuvieron los archivos gtf para todos los archivos .bam se generó un gtf que incluye todos los gtf generados anteriormente. Para esto, es necesario escribir un archivo llamado mergelist_f12sal.txt que contiene los nombres de los archivos gtf generados anteriormente.

Se utilizó la opción --merge que une todos los archivos indicados en el archivo mergelist_f12sal.txt y se escribe el resultado en el archivo stringtie_mergedf12.gtf.

stringtie --merge -p 32 -o stringtie_mergedf12.gtf mergelist_f12sal.txt

El último paso fue generar los archivos formato ball para trabajar con DeSeq2:

stringtie -e -B -p 30 -G stringtie_mergedf12.gtf -o ball_f12_0hrs_control1.gtf ../0hrs-control1.bam

Tabla 4 : Parámetros utilizados en Stringtie y su uso.

Parámetro	Valor	Uso
-e	.	Limita el procesamiento de alineamientos de reads para estimar y generar sólo las transcripciones ensambladas que coinciden con las transcripciones de referencia proporcionadas con la opción -G(M. Pertea et al. 2015)
-B	.	Se generan las salidas en formato ballgown el cual contiene los counts para los transcritos de referencia dada la opción -G. El formato ballgown corresponden a tablas en formato *ctab(M. Pertea et al. 2015).
-p	30	Número de Hebras(M. Pertea et al. 2015)
-G	.	Archivo de Anotación de Referencia(M. Pertea et al. 2015)
-o	.	Archivo para escribir el output(M. Pertea et al. 2015)

Convertir Identificadores del genoma de referencia a formato de Stringtie:

El programa gffread se puede utilizar para validar, filtrar, convertir y realizar otras operaciones en archivos GFF u otros formatos de archivos de anotación. Este programa se usa para verificar que estos programas comprenden correctamente un archivo GFF de una determinada fuente de anotaciones *(G. Pertea y Pertea 2020).*

Los transcriptomas generados tienen los identificadores de secuencia generados por *Trinity*, mientras que los identificadores generados por *stringtie* son distintos en nuestros archivos en formato *ball*, por lo que identificar las secuencias fasta de los DEGs desde el transcriptoma de referencia podría no ser posible debido a la diferencia de los identificadores de las secuencias. Para que esto no sucediera se utilizó el *software* gffread de la siguiente forma:

gffread -w f12_for_annotation.fasta -g F12salnew_cdhit.fasta stringtie_mergedf12.gtf

Donde la opción w genera archivo fasta con los exones para cada transcrito del GTF. Con la opción -g entregamos el archivo fasta que queremos utilizar y con el archivo strintie_mergedf12.gtf se entregan los nombres de los transcritos.

7.7. Expresión Diferencial con DESeq2.

El paquete DESeq2 proporciona métodos para probar la expresión diferencial mediante el uso de modelos lineales generalizados binomiales negativos; las estimaciones de la dispersión y los cambios de pliegue logarítmico incorporan distribuciones previas basadas en datos (*Love, Huber, y Anders 2014)*.

Una vez que tenemos los archivos en formato ball.gtf es necesario convertirlos a una matriz de expresión de genes que contiene los counts de cada *read*. Existe un *script* proporcionado por los desarrolladores de SringTie que convierte los archivos en formato *ball* a una matriz de counts *(M. Pertea et al. 2015)*. Para realizar este procedimiento el *script* toma como entrada un archivo con las etiquetas de los archivos ball.gtf junto con la ruta del archivo dentro del ordenador como se muestra en la siguiente imagen:

```
samplename_control1_ball /FULL/PATH/TO/samplename_control1_ball.gtf
samplename_control2_ball /FULL/PATH/TO/samplename_control2_ball.gtf
samplename_control3_ball /FULL/PATH/TO/samplename_control3_ball.gtf
samplename_1_ball /FULL/PATH/TO/samplename_1_ball.gtf
samplename_2_ball /FULL/PATH/TO/samplename_2_ball.gtf
samplename_3_ball /FULL/PATH/TO/samplename_3_ball.gtf
```

Figura 5: Formato para archivo de entrada para el script prepDE.py.

El procedimiento se hizo para cada uno de los tiempos en las librerías de RNA-seq, es decir, se agruparon todas las muestras correspondientes a un tiempo con su correspondiente muestra control. El análisis diferencial en DESeq se hizo para cada grupo generado.

Una vez con los archivos de matrices listos se procedió a abrir R. Se cargó la librería DESeq2.

Primero es necesario cargar la matriz de *counts* y etiquetar las columnas de nuestro archivo.

Para esto utilizamos los siguientes comandos:

>countData<- as.matrix(read.csv("gene_count_matrix.csv", row.names="gene_id"))

> colData <- read.csv(PHENO_DATA, sep="\t", row.names=1)

Donde PHENODATA corresponde a un archivo de texto que contiene información general sobre las muestras. En este caso contiene un identificador para la muestra, el tiempo en el que esta fue extraída, y un indicador correspondiente a la condición desde donde fue extraída la muestra,como se indica en la siguiente imagen:

```
ids,time,condition
samplename_control1_ball,0h,controla
samplename_control2_ball,0h,controla
samplename_control3_ball,0h,controla
samplename_1_ball,0h,controlb
samplename_2_ball,0h,controlb
samplename_3_ball,0h,controlb
```

` **Figura 6: Archivo PHENODATA utilizado.**

Con el archivo de características fenotípicas listas, se creó un objeto Deseq en formato de dataset desde nuestras matrices.

Para esto utilizamos el siguiente método:

dds <- DESeqDataSetFromMatrix(countData = countData, colData = colData, design = ~ CHOOSE_FEATURE)

Donde CHOOSE_FEATURE corresponde a la columna con la característica en la cual nos basaremos para realizar la expresión diferencial, en este caso, *condition*.

Luego, se realiza el análisis de expresión diferencial con:

> dds <- DESeq(dds)

Y se guardan los resultados en

> res <- results(dds)

Luego, se ordenan a partir de los p-values con:

> (resOrdered <- res[order(res$padj),]).

Los genes con un p-value ajustado < 0.01 y con un Fold Change mayor a 1.5 o menor a -1.5 fueron elegidos para seguir investigando.
Se decidió generar el corte de p-value en este valor dado que al utilizar el valor convencional de 0.05 más común en la literatura la cantidad de genes expresados diferencialmente aumentaba más allá de los límites para poder construir una red regulatoria con DEGs significativos. Se utilizaron varios valores de p-value para evaluar la cantidad de DEGs que se obtienen, en el caso 0,01 se limita en gran parte la probabilidad de observar una diferencia que sea un falso positivo pero al mismo tiempo se obtiene una cantidad de genes apropiada para construir las redes regulatorias.

7.8. Extracción de secuencias expresadas diferencialmente con Bioperl.

Se generó un *script* en perl utilizando las extensiones de Bioperl. Este *script* toma una lista de IDS de genes , en este caso, los IDs de genes expresados diferencialmente, identifica su presencia en el transcriptoma de referencia y extrae la secuencia, para obtener las secuencias en formato fasta expresadas diferencialmente para cada librería.

Los métodos utilizados fueron los objetos de id y next seq, donde se establece la condicional de que sí se encuentra el ID de la secuencia proveniente de la lista de DEGs en el transcriptoma de referencia, se escribe el id y la secuencia en un archivo de salida.

7.9. HMMER2GO y Blast2Go para anotación de genes.

HMMER2GO es una aplicación de línea de comandos para mapear secuencias de ADN a términos de Gene Ontology en función de la similitud de las secuencias de consulta con modelos HMM seleccionados para familias de proteínas representadas en Pfam (*Staton 2021*).

Primero se obtuvieron los marcos de lectura abiertos del fasta que se extrajo del genoma de referencia:

hmmer2go getorf -i genes.fasta -o genes_orfs.faa

Donde -i corresponde al *input*, que sería un archivo multifasta con las secuencias de los genes expresados diferencialmente y -o corresponde al archivo de salida. Luego, se buscaron los dominios codificantes y se escriben en un archivo de salida:

hmmer2go run -i genes_orfs.faa -d Pfam-A.hmm -o genes_orf_Pfam-A.tblout

Tabla 5 : Parámetros utilizados para HMMER2GO.

Parámetro utilizado.	Uso.
-i	archivo de entrada generado en hmmer2go getorfs
-d	La base de datos utilizada, en este caso se utilizaron los perfiles hmm de Pfam.
-o	output generado

Paralelamente se realizó una búsqueda con el software Blast2GO de donde obtenemos los GOTerms para cada una de las secuencias fasta de los DEGs. La idea es obtener la función metabólica y el proceso celular asociado a cada una de las secuencias de ADN. Esta información será utilizada después para realizar clusters en las redes regulatorias y de esta forma acceder de manera más directa a la información.

7.10. Determinar redes regulatorias de genes expresados diferencialmente.

Para generar las redes regulatorias se utilizó Genie3, el cual corresponde a un algoritmo de inferencia de redes regulatorias desde datos de expresión génica. En este algoritmo, lo que se hace es descomponer la predicción de redes regulatorias entre p genes en p diferentes problemas de regresión. En cada uno de estos problemas, la expresión del patrón de uno de los genes *target* es predicha desde los patrones de expresión de todo el resto de los genes a través de métodos de aprendizaje basados en árboles como Random Forest (*Huynh-Thu et al. 2010*)

Genie3 está implementado para R, por lo que se espera utilizar este paquete para construir la red regulatoria.
El paquete contiene la función GENIE3() , donde se utiliza como *input* una matriz del siguiente formato:

```
##       Sample1 Sample2 Sample3 Sample4 Sample5
## Gene1       2       6       6       4       8
## Gene2       3       8       4       3       4
## Gene3       7       2      10       1       3
## Gene4       4      10       9       6       5
## Gene5       9       8       2       4       2
## Gene6      10       1       4       9       6
```

Figura 7: Formato para matriz de entrada a GENIE3.

Este formato se puede extraer a partir de los *counts* generados por stringtie y ballgown en gene_count_matrix.cvs (*GENIE3 vignette s. f.*).

El resultado de la función GENIE3 utilizando el *input* de la matriz con el perfil de expresión corresponde a un archivo que contiene los pesos de las conexiones regulatorias putativas, donde aquellas interacciones con más alto peso corresponden a links regulatorios más probables (*GENIE3 vignette s. f.*). El archivo contiene todas las interacciones entre todos los genes que están siendo diferencialmente expresados, este archivo está ordenado de tal forma que las interacciones con más peso, las más "importantes", se encuentran al principio del archivo. Esto generará más de un millón de interacciones, cantidad que resulta excesiva para representar y analizar en Cytoscape, por lo que el output se cortó hasta tener las interacciones más importantes pero que la cantidad de DEGs no se viera disminuida. Así, las interacciones con peso muy bajo o no significativos resultan descartadas, pero no se pierde información de ninguno de los nodos que representan DEGs.

Para visualizar la red generada por GENIE3 hay que utilizar el programa Cytoscape . Además se utilizó la aplicación AutoAnnotate de Cytoscape para realizar una clusterización de los genes encontrados según la función o proceso celular asociado.Los procesos celulares y funciones metabólicas asociadas se extrajeron con el software Blast2Go a partir de términos GO de las secuencias fasta de DEGs extraídas del transcriptoma. Se utilizó la vista *Directed* donde se define la dirección en la que se genera la interacción, es decir, sí el gen está regulando a otro o está siendo regulado, o ambas. Se seleccionó este tipo de visualización con el fin de analizar cuáles son los posibles orígenes de las respuestas coordinadas de los organismos. Se configuraron los colores como *Discrete Mapping* para definir la tonalidad según el nivel de expresión representado por el Fold Change. Se utilizó el módulo Cytohubba para seleccionar subredes con los nodos más relevantes según el grado y el puntaje asociado a la importancia dentro de la red. Se calcularon los puntajes de los nodos con los métodos MCC (Maximal Click Centrality), Degree Method y Closeness. Las 3 visualizaciones coincidieron en los resultados por lo que se procedió a trabajar con MCC y se seleccionó la opción expandir subred con los 10 nodos más importantes. Según el paper de Cytohubba, el método MCC presenta la mejor performance que los otros 11 métodos de cálculo de puntajes, dado que captura las proteínas más esenciales en la lista de interacciones dado el flujo de información que pasa por el nodo, en proteínas con alto o poco grado.

8.Resultados.

8.1. Filtro de Secuencias.

En la visualización de FastQC se pudo apreciar que no existen *reads* con una calidad menor a Q30 en ninguna de las 48 librerías.

Al realizar el filtrado de los archivos .fastqc, se observó que la cantidad de *reads* que había en las muestras de F12 variaban entre los 12,000,000 hasta los 18,800,000 de reads, exceptuando la cantidad anómala de reads que se encuentra en la muestra de los tres días expuesta ante estrés salino donde encontramos 42,160,000 lecturas de RNA.

Por otro lado, en M2624 se puede observar que la cantidad de *reads* que se obtienen después del filtrado varió desde los 12,1000,1000 hasta los 19,128,000 de *reads*.

El porcentaje de reads filtrados más alto entre todas las muestras alcanzó solo un 0,04% correspondiente a una de las muestras en condiciones de estrés salino a los 3 días.

Tabla 6 : Reads filtrados por Software AfterQC por cada una de las muestras.

Sample	Reads Totales	Reads Filtrados	Sample	Reads Totales	Reads Filtrados
M26 0hrs_ctrl_1	13,841,000	34.002	M26 3d_ctrl_1	14,476,000	62.031
M26 0hrs_ctrl_2	13,787,000	33.693	M26 3d_ctrl_2	12,380,000	25.555
M26 0hrs_ctrl_3	14,383,000	33.444	M26 3d_ctrl_3	12,204,000	24.042
M26 0hrs_1	19,128,000	40.326	M26 3d_1	16,092,000	25.088
M26 0hrs_2	16,510,000	32.782	M26 3d_2	12,791,000	22.418
M26 0hrs_3	14,527,000	29.96	M26 3d_3	13,625,000	20.064
F12 0hrs_ctrl_1	12,810,000	26.624	F12 3d_ctrl_1	12,284,000	25.514
F12 0hrs_ctrl_2	14,257,000	32.148	F12 3d_ctrl_2	14,257,000	27.808
F12 0hrs_ctrl_3	14,740,000	30.417	F12 3d_ctrl_3	14,740,000	30.836
F12 0hrs_1	16,549,000	9.646	F12 3d_1	14,663,000	21.879
F12 0hrs_2	16,363,000	9.409	F12 3d_2	42,160,000	70.512
F12 0hrs_3	18,017,000	11.118	F12 3d_3	14,308,000	19.05
M26 6hrs_ctrl_1	18,787,000	38.489	M26 14d_ctrl_1	16,528,000	35.114
M26 6hrs_ctrl_2	13,824,000	30.793	M26 14d_ctrl_2	12,174,000	19.607
M26 6hrs_ctrl_3	15,535,000	35.732	M26 14d_ctrl_3	13,503,000	28.345
M26 6hrs_1	16,174,000	32.349	M26 14d_1	13,852,000	26.03
M26 6hrs_2	16,274,000	31.688	M26 14d_2	14,872,000	26.493
M26 6hrs_3	13,659,000	22.88	M26 14d_3	12,548,000	17.363
F12 6hrs_ctrl_1	16,931,000	33.335	F12 14d_ctrl_1	14,972,000	27.2
F12 6hrs_ctrl_2	17,707,000	36.748	F12 14d_ctrl_2	16,967,000	27.767
F12 6hrs_ctrl_3	15,878,000	35.578	F12 14d_ctrl_3	14,765,000	30.606
F12 6hrs_1	14,800,000	7.181	F12 14d_1	13,567,000	28.745
F12 6hrs_2	16,134,000	28.234	F12 14d_2	14,072,000	19.363
F12 6hrs_3	6,419,000	36.224	F12 14d_3	15,367,000	26.785

8.2. Ensamble.

Al realizar el ensamble se obtuvo un transcriptoma *de novo* para cada uno de los portainjertos, con un largo de 232,890,475 pares de bases para el caso de Mazzard F12/1 y 288,815,223 para el caso de M2624. El software *Trinity* encontró 406,666 secuencias para Mazzard F12/1 y 488,349 para Marianna 2624, que luego de ser filtradas con TransDecoder y CD-HIT resultaron en 115,556 y 130,870 respectivamente. En la Tabla 7 se muestran los valores resumen para cada genotipo:

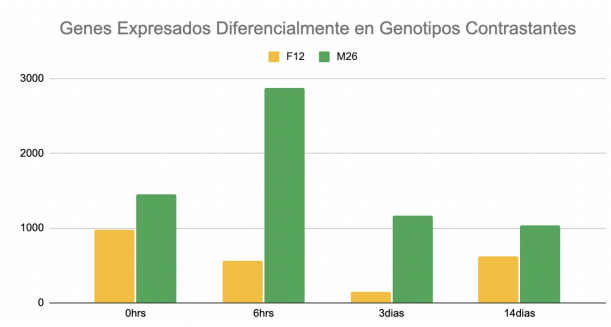
Tabla 7: Tabla resumen resultados de largo de ensamble

Genotipo	Mazzard F12/1	Marianna 2624
Largo de Ensamble	232,890,476 bp	288,815,223 bp
Cantidad de Secuencias	406,666	488,349
Secuencias Finales (cd-hit output)	115,556	130,870

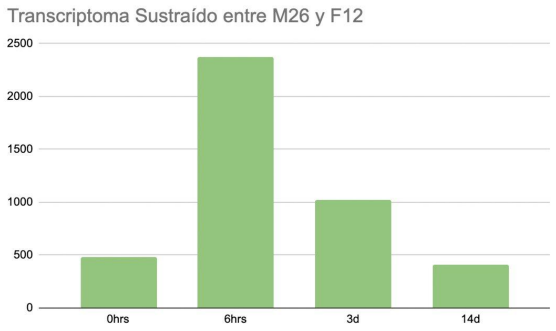
8.3. Expresión Diferencial.

Para el análisis de expresión diferencial se realizó la comparación entre las muestras control y las muestras expuestas a estrés salino correspondientes a un mismo tiempo y genotipo. A partir de este análisis se observó que, si bien la cantidad de Genes Expresados Diferencialmente(DEGs) de M2624 se mantiene considerablemente más alta que en F12/1 en los distintos puntos del tiempo. Aquí, es posible detectar un notorio contraste entre M2624 y F12/1 a las seis horas de tratamiento salino donde la diferencia en la cantidad de DEGs es muy amplia. La cantidad inicial de DEGs comienza a las 0 horas con 978 genes para F12/1 y 1454 genes para M2624. A las seis horas, mientras la cantidad del primer portainjerto subió hasta 2874 DEGs, la del segundo bajó a 563. A los 3 días ambos genotipos bajaron el nivel de DEGs hasta 145 y 1156 para F12 y M2624, respectivamente. A los 14 días en F12/1 los DEGs volvieron a subir a 626 genes mientras que los de M2624 bajaron a 1034 genes (**Figura 8A**).

A



B



C

Tiempo	F12/1	M2624
0 hrs	976	1454
6 hrs	563	2874
3 días	145	1166
14 días	626	1034

Figura 8: Cantidad de DEGs expresados diferencialmente. En A genotipos contrastantes en respuesta a estrés salino en distintos puntos del tiempo.En B diferencia de DEGs entre ambos genotipos en distintos puntos del tiempo. En C cantidad exacta de DEGs en cada genotipo en distintos puntos del tiempo.

Es así como la mayor diferencia de DEGs se alcanza a las seis horas. De los genes expresados diferencialmente existen genes que se sobre-expresan o se subexpresan.

A las 0 horas en ambos genotipos se observó que la mayor parte de DEGs están siendo inducidos en su expresión. En F12/1 solo 68 genes estuvieron sobre-expresados mientras que 908 genes estuvieron sub-expresados. Para M2624 a las mismas cero horas se sobre-expresaron solo 49 genes y se sub-expresaron 1405. A las seis horas se observó una respuesta contrastante entre F12/1 y M2624, aumentando F12 la cantidad de genes sobre-expresados hasta 415 y disminuyendo la cantidad de genes sub-expresados a -148, mientras que M2624 aumentó tanto genes sobre-expresados como sub-expresados a 373 y 2502, respectivamente. A los tres días F12/1 disminuyó la cantidad de DEGs en comparación con las cero horas y las seis horas, alcanzando 109 genes sobre-expresados y 36 genes sub-expresados. M2624 también disminuyó la cantidad de genes sub-expresados a 988 y la cantidad de genes sobre-expresados a 462. Por último, a los 14 días en F12/1 aumenta la cantidad de DEGs pero los genes sobre-expresados disminuyen a 163 mientras

los sub-expresados aumentan a 462. M2624 por otro lado a los 14 días aumentó la cantidad de genes sobre-expresados a 412 y disminuyó los genes sub-expresados a 622 (Figura 9).

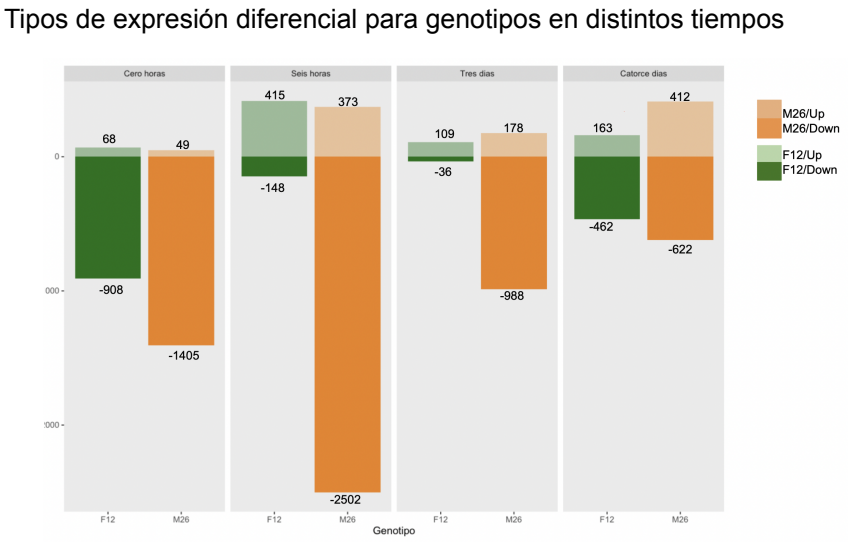


Figura 9: Cantidad de DEGs que aumentan o disminuyen la tasa de expresión en distintos puntos en el tiempo para ambos genotipos.

A B

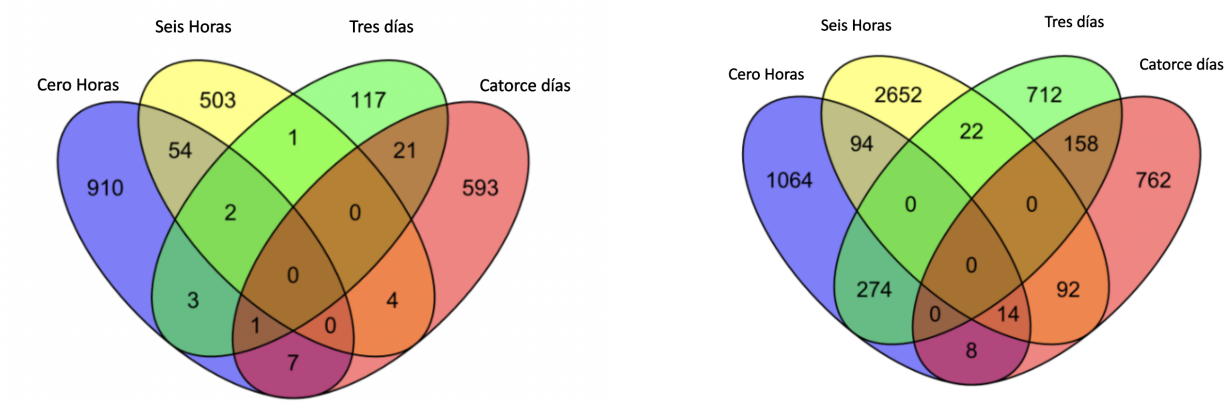


Figura 10: Intersección de genes expresados diferencialmente. En A, genes que coinciden en distintos puntos del tiempo para F12/1. En B, genes que coinciden en distintos puntos en el tiempo en M2624.

Podemos observar como la intersección con más alta cantidad de DEGs para F12/1 fue entre las cero y las seis horas con 54 genes que seguían siendo expresados

diferencialmente. Por el contrario, para M2624 la mayor intersección de genes se alcanzó entre los 3 y 14 días con 158 genes que se seguían expresando diferencialmente.

8.4. Anotación.

De los genes que fueron identificados como diferencialmente expresados, no todos se asociaron a funciones biológicas en bases de datos, a continuación se presentan las cantidades de DEGs por tiempo y por genotipo que pudieron anotarse:

Tabla 8 : Genes Diferencialmente Expresados donde se encontró función asociada con HMMER2GO.

Tiempo	Secuencias Anotadas M2624	%M2624 de DEGs totales	Secuencias Anotadas F12/1	%F12/1 de DEGs totales
0 horas	1125	77,37	846	86,68
6 horas	2367	82,35	446	79,21
3 días	1051	90,13	126	86,89
14 días	841	64,49	542	86,58

En la obtención de GOTerms con Blast2Go se destacan los siguientes procesos celulares, los cuales en el proceso de clusterización resultaron ser los grupos de procesos con más genes involucrados:

Tabla 9 : Procesos celulares con mayor cantidad de genes involucrados.

Tiempo	M2624	Genes Involucrados	F12/1	Genes Involucrados
0 horas	Procesos de traducción Ribosomal	14	Unión a ADN	2
6 horas	Procesos de traducción Ribosomal	8	Unión a ADN	5
3 días	Procesos de traducción Ribosomal	11	Unión a ADN	5
14 días	Actividad de oxido-reductasas	2	Actividad de óxido-reductasas	3

8.5. Redes Regulatorias.

La red generada para F12 a las 0 horas (**Figura 11**) consistió en 59 *nodos* con 89 genes donde solo 7 genes están siendo sobre-expresados y el resto sub-expresados. Los factores de transcripción DUF4407 y DUF3730 junto con B3 se vieron más sub-expresados que el resto de los genes de la red con valores Fold Change de -21,6 y -13,06 respectivamente. Las proteínas DUF contienen dominios de función desconocidos lo que podría indicar que existen proteínas no estudiadas que juegan un rol fundamental a la hora de mediar respuestas ante estrés abiótico. Los genes Med3, Histidinol_dh y ArsA_HSP20 resultaron ser los genes más sobre-expresados con valores Fold Change de 8.28, 12.03 y 15.7 respectivamente. En este tiempo la mayoría de los genes representados por los nodos disminuyeron su actividad transcripcional.

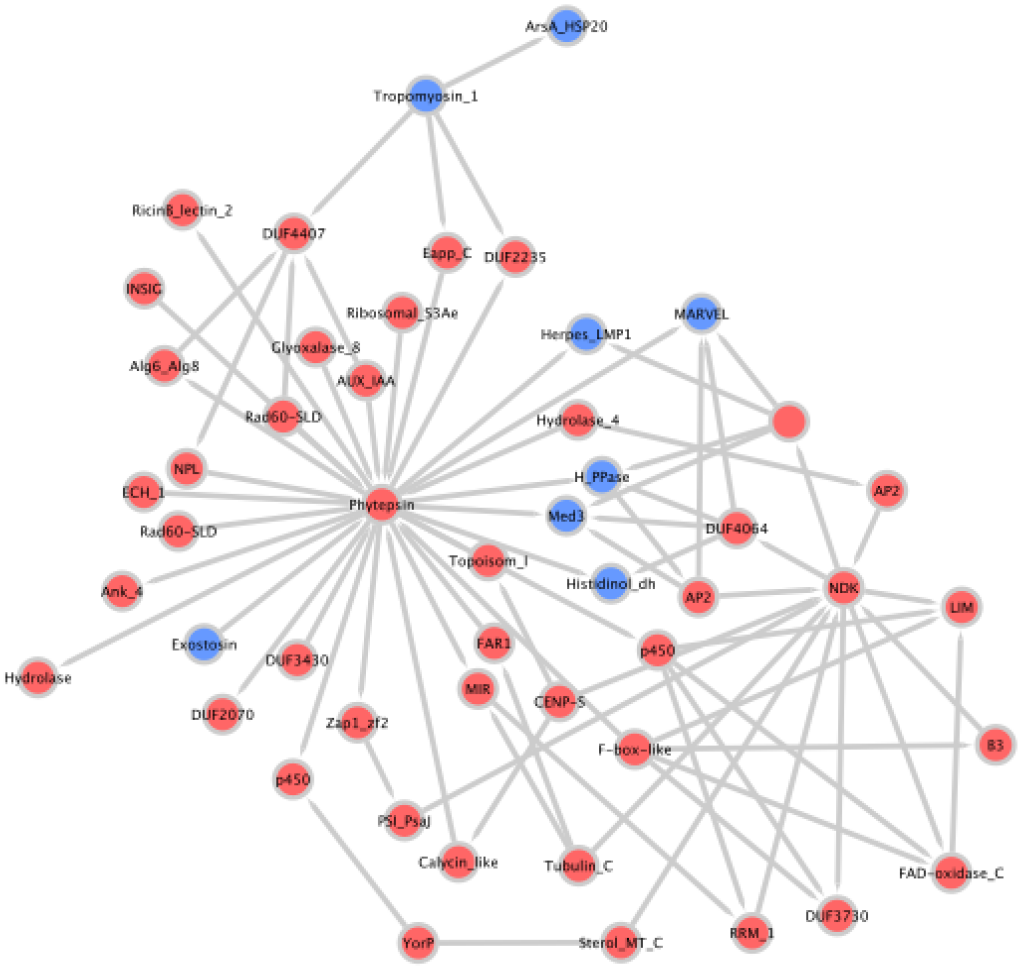


Figura 11: Red de Regulación génica para F12 a las 0 horas.

En M2624 a las 0 horas se obtuvo una red de 144 nodos con 334 interacciones donde solo 12 genes se fueron sobre-expresados (**Figura 12**). Se observó que existe una respuesta conjunta en proteínas ribosomales con un nivel de expresión que indica que se están sub-expresando. Los genes Crinkler, Dpy-30, RMM_1 reprimieron su expresión con valores Fold Change de -9.33, -8.16, -27.56 respectivamente. Los genes Peptidasa_C69 y ISTAND corresponden a los nodos con mayor grado en la red, lo que significa que estos genes interactúan con varios genes. Anoctaminas, CsbD, SPESP1 se están sobre-expresando con valores Fold Change de 5.7, 5.3 y 6.017 respectivamente. La mayoría de los genes respondieron en conjunto disminuyendo su expresión en comparación al control en este punto en el tiempo.

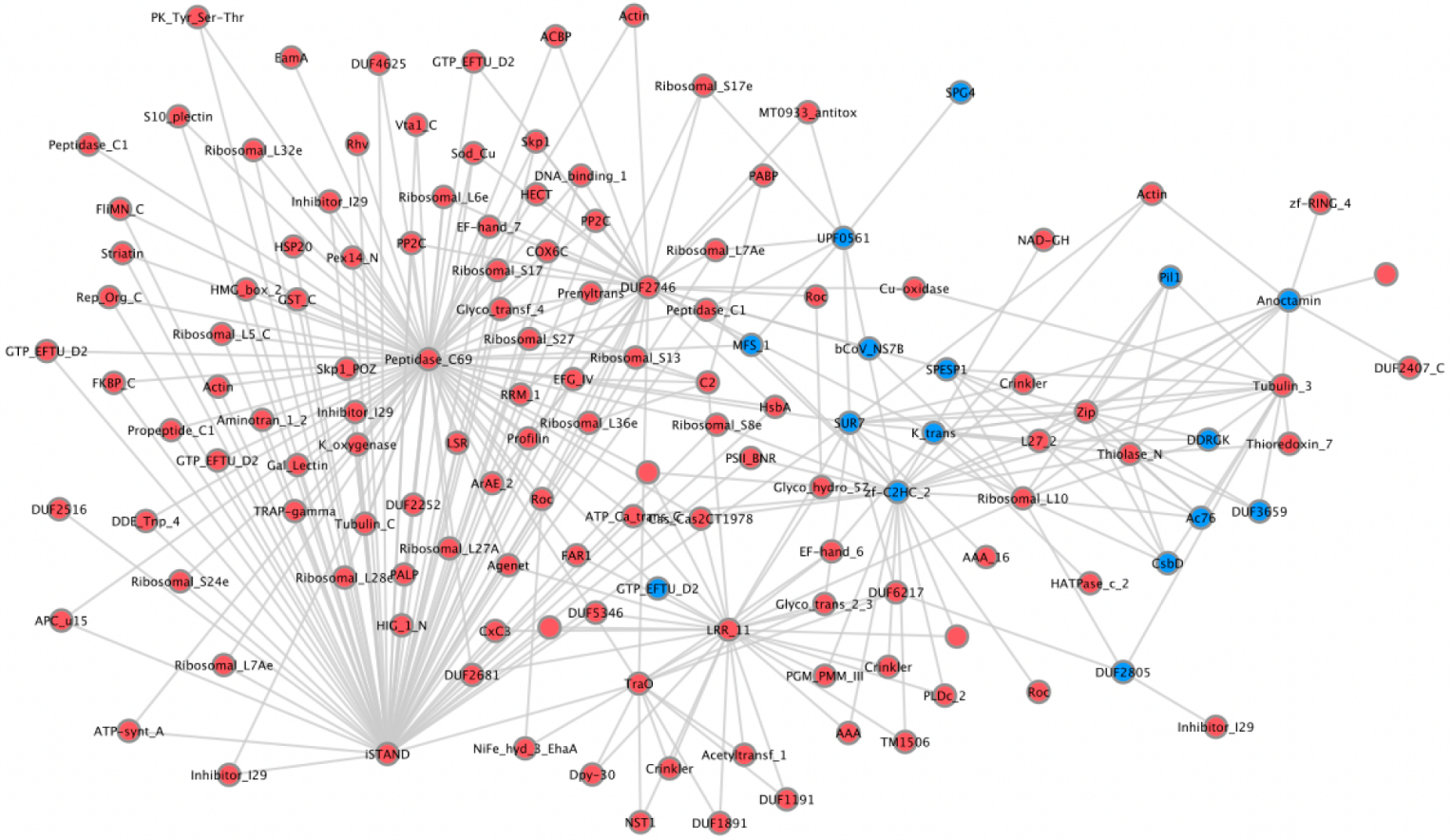


Figura 12: Red Regulatoria de genes para M2624 a las 0 horas.

A las 6 hrs en F12 se obtuvo una red de 154 DEGs o nodos con 325 interacciones (**Figura 13**). En estas condiciones los genes HSP90, genes asociados a actividad de kinasas, MSA-2c, NAD-binding,RHH-6, THDPS son los que más se subexpresan con respecto al resto de nodos de la red, presentando todos Fold Changes de -9, excepto el gen NAD-binding con un Fold Change de -8. A diferencia de la red anterior, en esta red la actividad de genes que fueron sobre-expresados y que se visualizan aumentan en 96 genes con respecto a las cero horas. Los factores de transcripción tipo DUFF se sobre-expresaron y sub-expresaron al igual que en la red anterior. Glioxalasas y genes PAS_3 representan los nodos con mayores grados de toda la red. Además, en la red con sin filtrar se puede apreciar cómo la reacción en cadena de cambios en las expresión en los genes comienza con el aumento de la actividad de genes Na_H_exchanger. Dentro de los genes más sobre-expresados se encontraron genes receptores de kinasas como CBM_4_9 con un Fold Change de 14.81, genes relacionados con funciones de desfosforilación como PLAT con un Fold Change de 14.43, genes codificadores de proteínas membranales como MIP con 30 de Fold Change y abidrolasas con 10 de Fold Change.

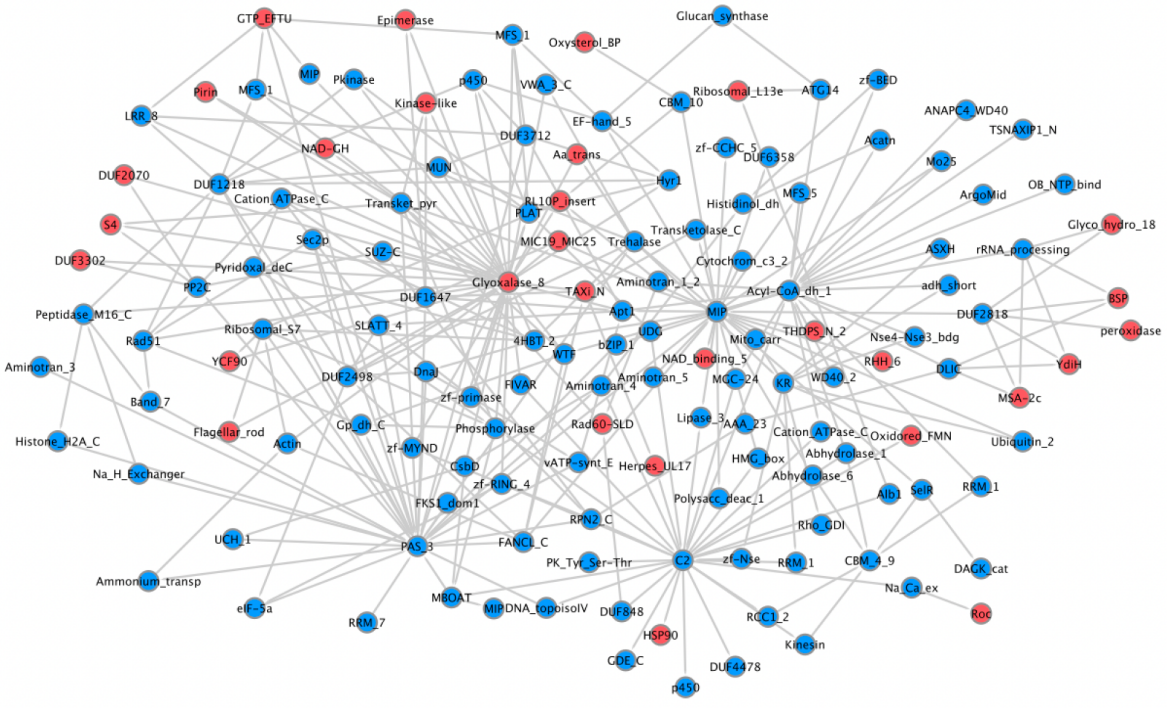


Figura 13: Red de Regulación Génica para F12 a las seis horas.

A las seis horas M2624 experimentó un cambio en lo que respecta a la actividad de proteínas ribosomales pasando de sub-expresar estos elementos a sobre-expresarlos (**Figura 14**). La red consiste en 131 nodos con 242 interacciones. Los anillos de zinc fueron sub-expresados junto con genes relacionados a fosforilación de kinasas. A estas alturas de la investigación se pudo visualizar la sobre-expresión de genes tipo NAC, AP2, MYB y WRKY. Los genes que están siendo más sobre-expresados corresponden a factores de transcripción TFIIA con valores Fold Change de 12 junto genes asociados a proteólisis aspartática como TAXI_C con valor Fold Change de 14.97. Metaloenzimas y genes SEO_C se sobre-expresaron aún más, con valores Fold Change de 28 y 27, respectivamente.

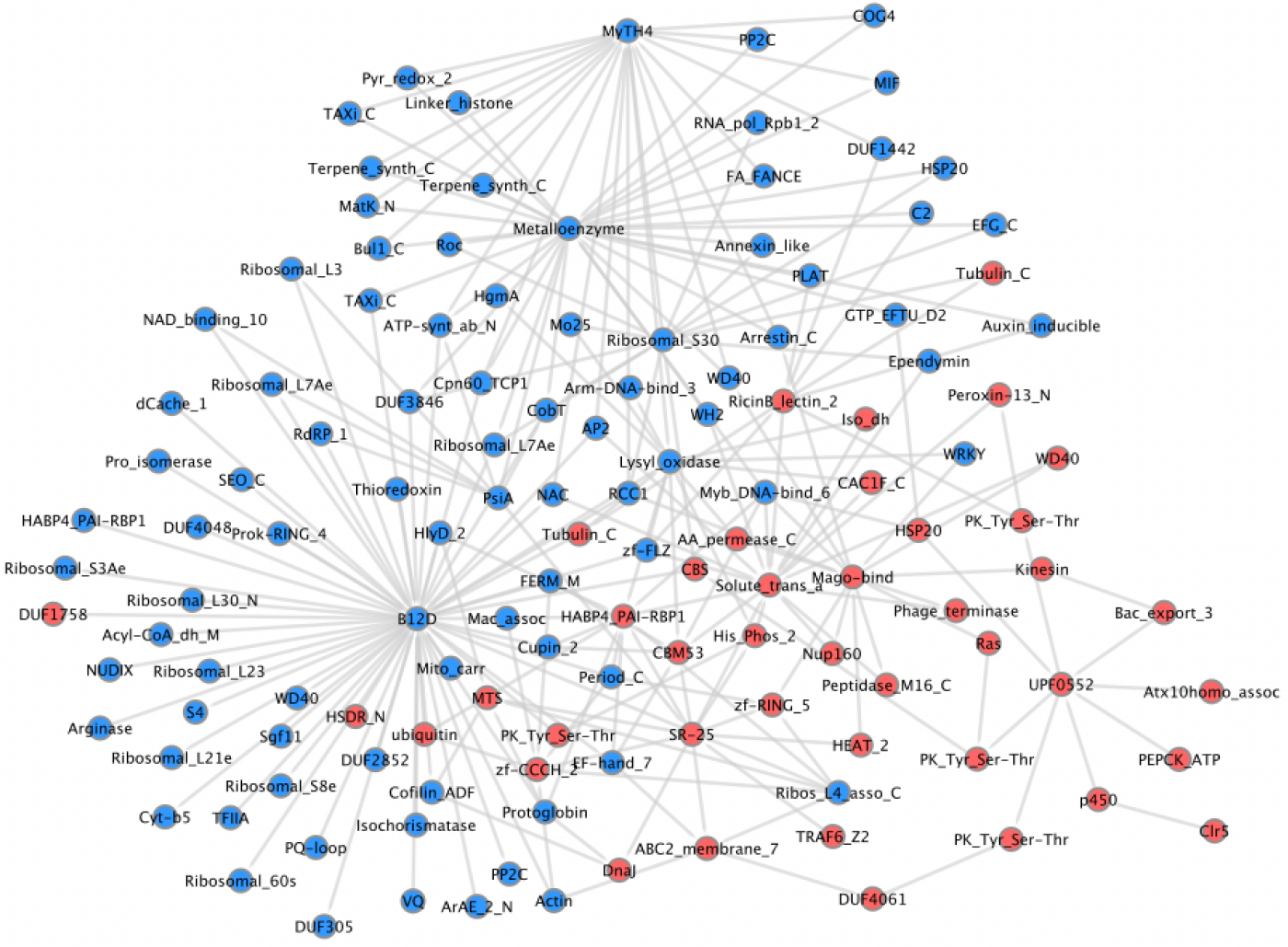


Figura 14: Red de Regulación Génica para M2624 a las seis horas.

A los tres días en F12 la red disminuyó notablemente la cantidad de nodos a 58 con 220 interacciones entre estos (**Figura 15**). El factor de transcripción bHLH se vió sobreexpresado e interactúa con factores de transcripción tipo MYB que también fueron sobre-expresados. Genes AP2 también están siendo sobreexpresados. Entre los genes más sobre-expresados que pueden reconocerse están los genes DIRP, RRM 1 y Cyt-b5 con Fold Changes de 20.45, 20.45 y 9.09, respectivamente. El factor de transcripción MYB es el gen que más se sobre-expresa con un Fold Change de 29.03. Los genes más sub-expresados en la red corresponden a genes tipo NAM con -28.7 de Fold Change, NAD_binding con -9.95, Rx_N con -12.6 , CofC con -10 , genes codificadores de Ubiquitinas con -15.83 y proteínas de dimerización con -9.02. Los nodos que representan los genes MYB y MBD tienen el mayor grado y por lo tanto la mayor cantidad de interacción entre genes.

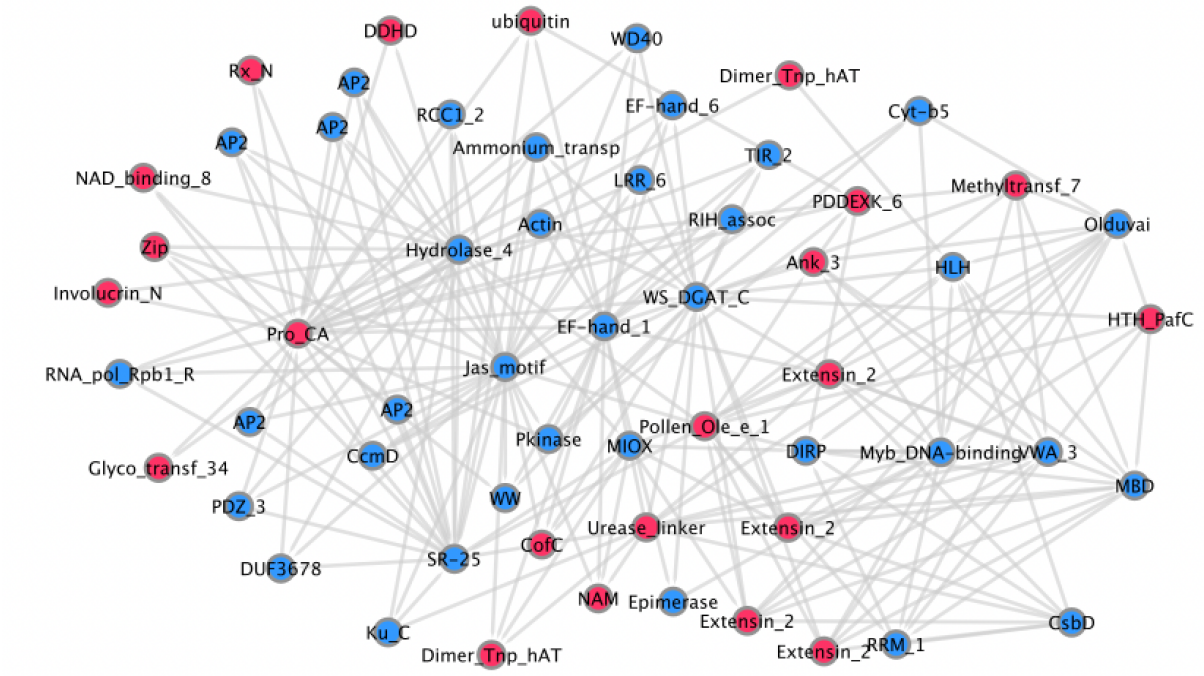
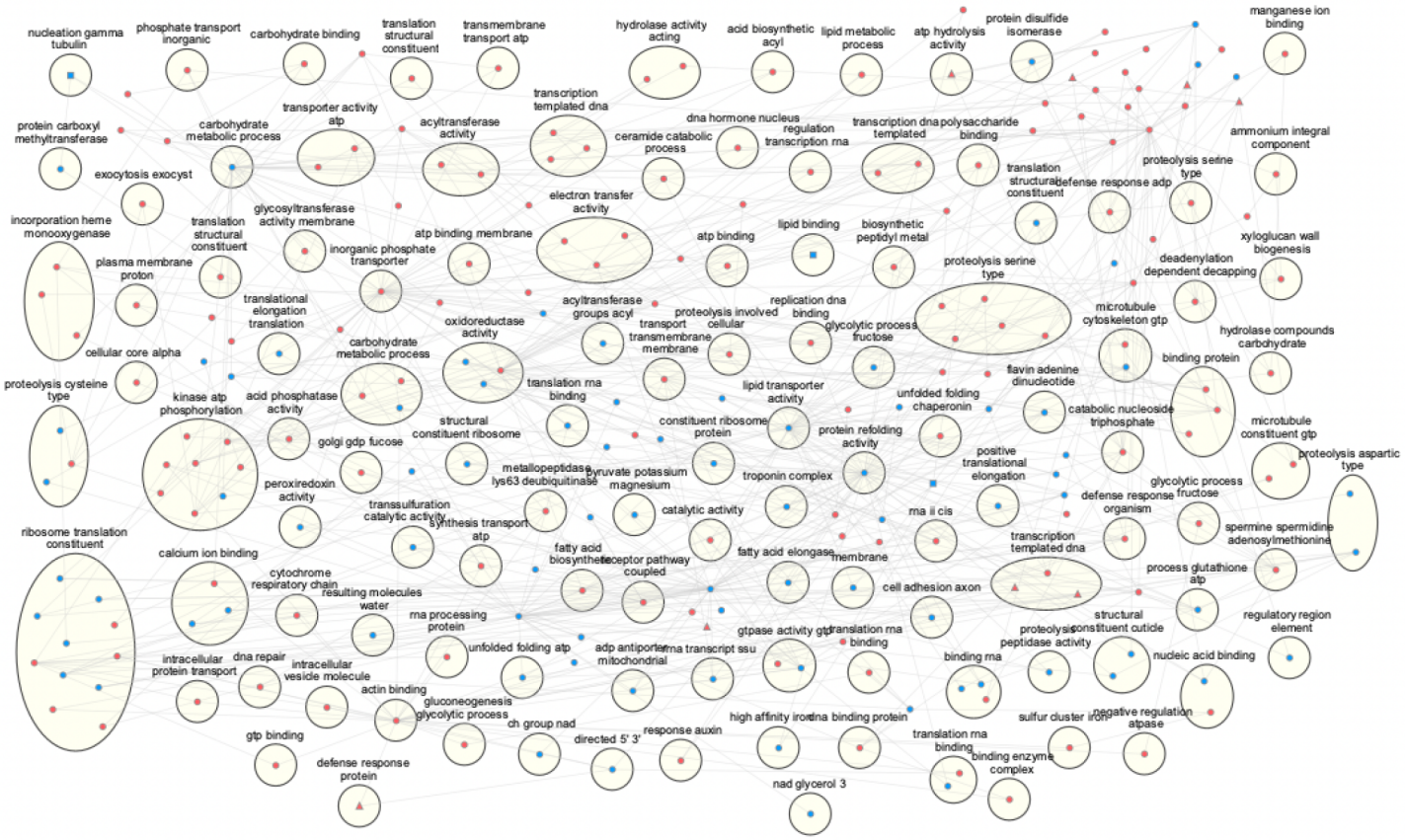


Figura 15: Red Reguladora de genes en F12 a los 3 días.

Para M2624 a los tres días entre los genes que fueron más subexpresados se encontraron genes FTR, EP400, C2 con Fold Changes de -8.97, -12.6 y -12.9, respectivamente. Entre los genes más sobre-expresados dentro de la red se encontraron HSP90 y UQ_con con Fold Changes de 29.8 y 27.13 respectivamente. Además, existen variaciones en los niveles de expresión de genes codificantes de proteínas ribosomales, donde a las seis horas este tipo de genes se sobre-expresada, ahora se sobre-expresan tanto como se sub-expresan. El gen TRMK asociado a actividad transferasa de grupos metilos tiene un alto grado en comparación al resto de los nodos. Genes asociados a plegado de proteínas como Cpn60TCP1 también representan una gran influencia en el flujo de información del nodo. En la Figura 16 se representa el formato de la red con la clusterización según función metabólica. Se muestra este formato dado que, por el tamaño de la red, se hace imposible para el programa visualizar el nombre individual de cada nodo.

Figura 16: Red Regulatoria de genes para M2624 a los 3 días.



A los 14 días en F12 se observó la red más pequeña de todas las condiciones en ambos genotipos, con 29 nodos y 46 interacciones (**Figura 17**). Existe una escasa respuesta

coordinada con respecto a factores de transcripción en esta red. Genes relacionados a síntesis de pared celular, metabolismo de carbohidratos como Glyco Hidrolasas y Glicosil transferasas se encuentran sobre-expresados mientras que genes como WS_DGAT_c y TAL_FSA están siendo sub-expresados. Genes relacionados a actividad oxidoreductoras como ADH_zinc fingers y adh_short_C2 se encuentran sub-expresadas también. El gen Bet_v_1 es el más sub-expresado en esta red con -19.85 de Fold Change junto con genes codificadores de Peptidasas con -12.67 de Fold Change. El gen más sobre-expresado en esta red corresponde a la secuencia de ADN que codifica para proteínas transportadoras de sodio, Na_Ca_ex, con un Fold Change de 12.5.

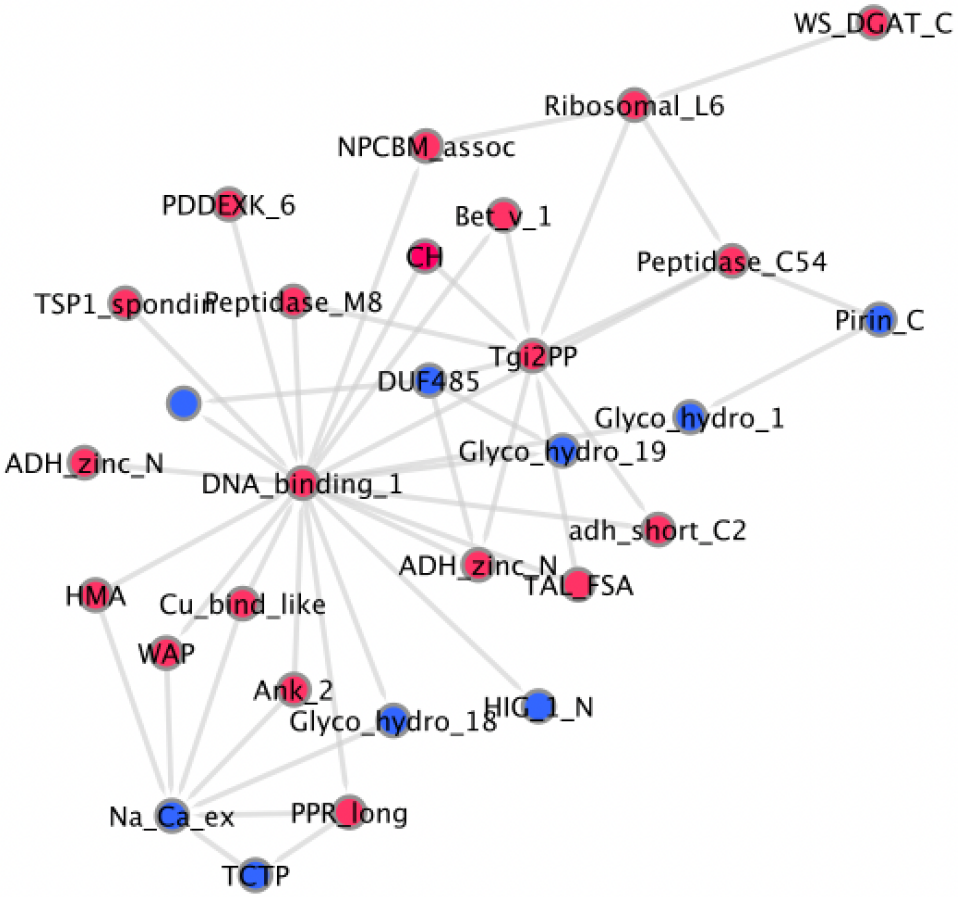


Figura 17: Red Regulatoria de genes para F12 a los 14 días.

La red de M2624 a los 14 días posee 66 nodos con 118 interacciones (**Figura 18**). El gen Staphylocoagulse es uno de los genes más sobreexpresados con un Fold Change de 8.6 y posee el mayor grado a lo largo de la red. Genes relacionados a dedos de zinc están siendo sobre-expresados a excepción de zf-Ring_9. Es aquí donde por primera vez comienzan a

aparecer factores de transcripción tipo GRAS. Factores de transcripción HSP20 y MYB se encuentran sub-expresados junto con genes relacionados con función de transferasas como TMEM230. En esta red, los genes representados que se sub-expresan no superan los -6.36, siendo el gen HMGbox el más sub-expresado, mientras que el que más se sobre-expresa es el gen Baculo_PEP_C con 9.99 de Fold Change.

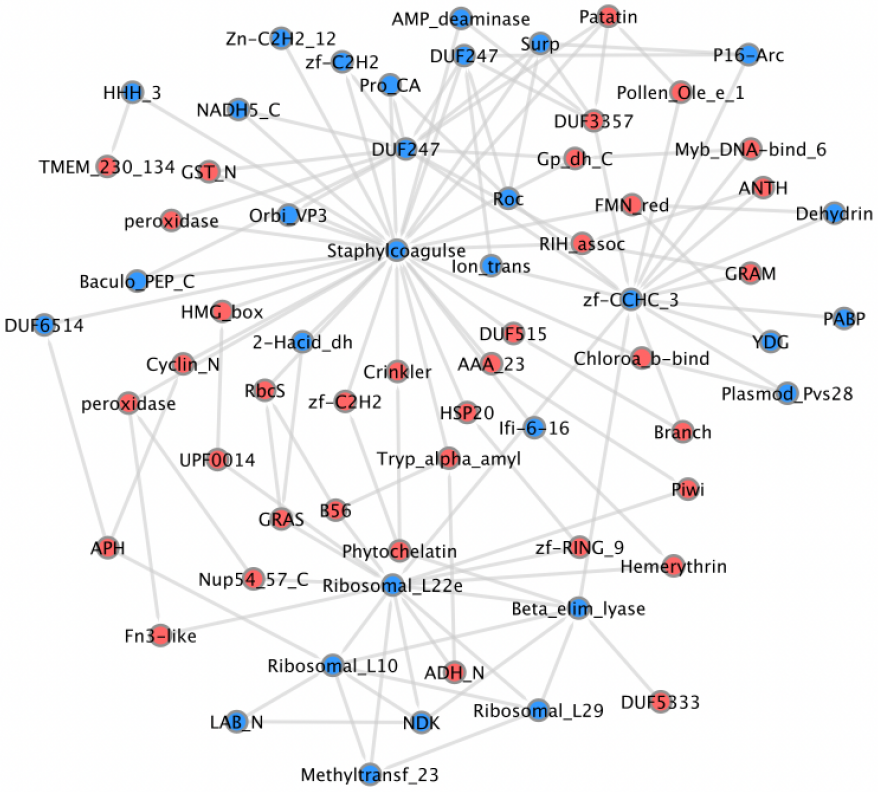


Figura 18: Red Regulatoria para M2624 a los 14 días.

9. Discusión.

Al realizarse el filtro de reads en ambos transcriptomas, se pudo observar como la cantidad de reads para una de las muestras correspondiente a F12/1 a los 3 días presentó un incremento en casi el doble con respecto al resto de librerías de RNA-seq para ambas especies, alcanzando los 42,160,000 millones de reads. Esta anomalía se relaciona a la cantidad de duplicados generados en procesos de amplificación de RNA durante PCR, duplicados que son secuenciados y se incluyen en los archivos fastq, los que al luego ser

alineados al transcriptoma *de novo*, pueden generar conteos aumentados con respecto a la cantidad real de transcritos para una secuencia *(Tin et al. 2015)*. Es por esto que al hacer el análisis de expresión diferencial uno de los pasos clave es la normalización de reads que se incluye automáticamente en DESeq2. Además, en el filtro del resultado del análisis de expresión diferencial, para seleccionar los genes expresados diferencialmente se utilizan valores de p-value y FDR acotados para así evitar la selección de falsos positivos.

A nivel de transcriptoma se observó cómo el de M2624 resulta ser más largo que el transcriptoma de F12/1. Las especies de *Prunus* tienen un porcentaje de identidad de un 60% cuando se alinean entre sí. Gran parte del 40% restante correspondiente a las bases nitrogenadas que no alinearon entre genotipos y esta cifra puede deberse principalmente a la diferencia de tamaños de transcriptoma que existe entre estos. Al calcular el tamaño de F12/1 con respecto a M2624 se puede establecer que *Prunus avium* representa aproximadamente un 80% del tamaño de *Prunus cerasifera x munsoniana,* cuando hablamos de bases nitrogenadas. Si se tienen transcriptomas que se alinean a un 60%, ese 20% restante corresponde a las variaciones transcripcionales propias de cada especie.

En el proceso de anotación podemos observar que existen distintas rutas de acción entre F12/1 y M2624 a las cero, seis y tres días. Mientras F12/1 sobre-expresa mayoritariamente rutas relacionadas proteínas de unión de ADN, M2624 sobre-expresa esas mismas rutas pero lo que más sobre-expresa son vías de traducción mediadas por ribosomas. La activación de síntesis proteica en gran cantidad podría ser el primero de los puntos de inflexión en la respuesta diferencial de los genotipos comparados. A los 14 días en ambas especies el cluster con mayor cantidad de genes involucrados corresponde a la actividad de óxido-reductasas. Es esperable que luego de haber pasado varios días de exposición a altas concentraciones de sal en el suelo de cultivo, los organismos comiencen a enfocar su actividad en procesos de oxidorreducción dado el aumento de especies reactivas de oxígeno en el organismo *(Kapoor et al. 2015)*

A las cero horas en F12/1 no se pudo apreciar una interacción de más de 3 genes que sigan la misma dirección en cadena o conjunta. Se logra visualizar la expresión diferencial de genes tipo DUF, pero estas secuencias de ADN se encuentran tanto sub-expresadas como sobre-expresadas a lo largo de los distintos tiempos en ambos genotipos. Como no se observa un patrón de expresión que sea significativamente distinto entre genotipos, este tipo de genes no se considera para análisis más profundos. Por otro lado, los genes que se mencionan como más sobre-expresados dentro de la red como Med3, Histidinol_dh y ArsA_HSP20 no interactúan con genes codificadores de proteínas membranales de señalización u otro factor de transcripción conocido en la literatura por funciones

fundamentales en la respuesta ante estrés abiótico. Esto podría representar que en ese tiempo aún no existe una respuesta de varios genes coordinados en su respuesta ante el estrés salino. Tampoco se observó la participación de genes codificadores de factores de transcripción coordinados en esta condición. Como se mencionó anteriormente, los factores de transcripción juegan un papel fundamental a la hora de mediar respuestas ante estreses medioambientales *(Baillo et al.,2019)*, por lo que la ausencia de estos dentro de la red indica que F12/1 no estaría presentando una regulación génica de genes relacionados a factores de transcripción que responden a estrés salino de la forma en que lo hace el genotipo con el que se compara. Por otro lado, en M2624 se reconoció la sobre-expresión de genes asociados a actividad ribosomal (proteínas ribosomales) y síntesis de ATP (ATP Sintasas). Estudios en organismos modelos han demostrado que la activación de proteínas ribosomales juega un rol importante en la biogénesis y ensamblado de proteínas ribosomales en su respuesta ante la disrupción de la homeostasis celular causadas por estrés abiótico *(Albert et al. 2019)* y cómo las moléculas de ATP cumplen un rol fundamental al servir como molécula central en la señalización de respuestas ante el mismo tipo de estrés *(Cao et al. 2014)*. Por otro lado, la sobre-expresión de fosfatasas como lo es el caso de genes PP2C y transferasas como los genes codificadores de Acetiltransferasas para M2624 pueden significar un punto clave en la respuesta diferencial de ambos genotipos, ya que son estas proteínas las que participan en las respuestas primarias ante estreses abióticos y la transducción de señales hacia factores de transcripción *(Cao et al. 2014).* En lo que respecta a los genes que fueron más diferencialmente expresados en M2624, la sub-expresión de genes Crinkler y RMM_1 corresponden a ADN de organismos ajenos a la planta, virus y pseudomonas respectivamente, los cuales pudieron interactuar en algún momento con el portainjerto o en su defecto ser contaminación dentro de la muestra. Por otro lado la sobre-expresión de genes codificadores de Anoctaminas, canales transportadores de cloro mediados por calcio, puede significar la activación y potenciación procesos relacionados a transporte de iones, los cuales se activan ante estreses por turgencia y pérdida de homeostasis celular *(Medrano-Soto et al. 2018)*. Los genes tipo CsbD, uno de los más sobre-expresados dentro de la red corresponden a genes bacterianos, fenómeno que puede deberse a altas similitudes de secuencia en el proceso de anotación donde se asocian secuencias de ADN vegetal a secuencias bacterianas, al igual que SPESP que corresponden a secuencias de ADN animal. Con respecto a los nodos que se presentan mayor grado dentro de la red, al gen ISTAND no se le encontró función asociada, pero el nodo correspondiente al gen codificador de Peptidasas representa un punto de partida para la señalización primaria de la respuesta ante estrés abiótico *(J. S. Kim, Jeon, y Kim 2021)*.

A las 6 horas se pudo observar cómo la cantidad de DEGs se comporta de manera contrastante en ambos portainjertos, aumentando en M2624 y disminuyendo en F12/1. Este comportamiento coincide con la idea de que, pasado un tiempo específico, en el genoma de M2624 existe una mayor cantidad de genes que responden ante el estímulo del estrés en comparación con F12/1, activando vías de modulación de la expresión génica, lo que se traduce en una mayor tolerancia al estrés salino en comparación a este último. Si bien F12/1 a estas alturas de la exposición ante estrés salino comenzó a sobre-expresar proteínas membranales como intercambiadores de sodio y calcio, esta sobre-expresión,sí bien desencadena la expresión diferencial de variados genes, no se logra observar la directa relación con factores de transcripción, al contrario de M2624, donde la sobre-expresión de genes codificadores de factores de transcripción como NAC *(Mohanta et al. 2020)* y WRKY *(Gao et al. 2020)*, ambos ya estudiados en su participación en la respuesta a estrés salino en otras especies, se vió aumentada dada la interacción con otro componente membranal B12D desde donde se perciben las señales bioquímicas. Para el caso de los genes más sub-expresados dentro de la red de F12/1 en este tiempo, el gen MSA_2c y RHH_6 no se encuentra caracterizado para plantas en las bases de datos, mientras que, genes relacionados al metabolismo de ATP como NAD-binding y THDPS (Succinil Transferasa), los cuales deberían estar sobre-expresados dada la cantidad de energía que se requiere para realizar transporte activo de iones y así mantener las concentraciones de sodio en equilibrio dentro de la célula, se encuentran sub-expresados. Por otro lado, M2624 tiene como genes más sobre-expresados secuencias de ADN que se relacionan directamente al desarrollo celular vegetal. El Factor de Transcripción Multifuncional (TFII), participa en procesos de señalización de crecimiento celular, lo cual podría ser uno de los factores de que M2624 pueda seguir desarrollándose a pesar de las condiciones ambientales desfavorables en las que se encuentra (*Tanikawa et al. 2011*). Por otro lado, otro de los genes que más se sobre-expresa para M2624 corresponde a TAXI_C, una proteína de tipo peptidasa de ácido aspártico, la cual no se relaciona directamente a actividades de respuesta ante estrés abiótico. Para el caso de las Metaloenzimas, las cuales están siendo sobre-expresadas para este genotipo en este tiempo y que además poseen uno de los mayores grados dentro de la red, se ha demostrado que estas cumplen un rol importante como marcadores bioquímicos ante variados estreses de tipo abiótico *(K. Berwal y Ram 2019)*. Por último, el gen SEO_c en M2624 a las seis horas es otro de los genes más sobre-expresados, no existe documentación de como este gen se relaciona con respuestas a estrés abiótico, pero sí como este participa activamente en el desarrollo del floema de organismos vegetales *(Srivastava y Tuteja 2014)*, desde donde se transportan los nutrientes e iones hacia la planta, proceso que podría estar relacionado con la absorción de iones de sodio.

A los tres días F12/1 presentó la menor cantidad de genes expresados diferencialmente entre todas las librerías analizadas. Genes que codifican factores de transcripción como AP2 y MYB fueron sobre-expresados, pero estos no interactúan con proteínas de membranas o de señalización dentro de la red. En la misma red para F12/1, los genes encontrados que más se expresan diferencialmente, DIRP, RRM 1 y Cyt-b5A , Rx_N y CofC no se encontraron con funciones asociadas a respuesta ante altas concentraciones de salinidad. Los genes NAM pertenecientes a la familia NAC se encuentran entre los más sub-expresados, cuando para aumentar la tolerancia a altas concentraciones de sal idealmente debería encontrarse sobre-expresados *(An et al. 2018)*. Los genes MBD tienen el mayor grado dentro de la red para este tiempo en F12, estos genes codifican para proteínas que participan el los procesos de metilación de la cromatina, donde cumplen un rol importante a nivel de expresión génica (*Grafi, Zemach, y Pitto 2007)* y se encuentran sobre-expresados. A los mismos 3 días M2624 presentó la red con mayor cantidad de nodos, pero no presentó una cantidad significativa de factores de transcripción que se estén sobre-expresando, por lo que M26 no requeriría una respuesta regulada que involucre la sobre-expresión entre varios factores de transcripción en este tiempo. Los genes que más se sub-expresan en esta red (FTR, EP400, C2) no se asociaron a procesos de metabolismo de sodio.

A los 14 días se pudo apreciar como la cantidad de genes expresados diferencialmente en M2624 comienza a estabilizarse y la red comienza a acotarse. A estas alturas, M2624 ya respondió ante el estrés salino y se encuentra generando procesos metabólicos de manera precisa para enfrentar el estrés iónico derivado de la exposición crónica a alta salinidad. Al visualizar la presencia de genes codificadores de Heat Shock Protein (HSP90) , podemos constatar cómo esta proteína cumple un rol importante en la regulación de los niveles de Na+ en el medio intracelular a partir de la regulación del compuesto Calcineurina *(Imai y Yahara 2000)* al igual que los genes tipo GRAS *(T.-T. Wang et al. 2020)*. F12/1 dado el mismo punto en el tiempo se encontró con una cantidad mínima de genes expresados diferencialmente y con la casi nula presencia de factores de transcripción en la red, la proteína de intercambio Na_Ca_ex desencadena expresiones diferenciales en variados genes, pero ninguno de éstos figura en la literatura como importante en respuestas ante estrés abiótico.

Una de las características interesantes a observar dentro de la red es el grado de los nodos. El grado corresponde al número de interacciones que nacen de los nodos, es decir, la cantidad de arcos. Biológicamente se puede interpretar como la cantidad de genes que interactúan con un gen específico, y esta interacción se interpreta como la relación de

expresión que existen entre ellos. En las redes, se encontraron varios genes que tienen el mayor grado dentro de la red y que resultan importantes en procesos homeostáticos del organismo vegetal en respuesta a estrés salino, transporte de iones, activación de vías de ATP, síntesis proteica, etc. Para el caso de M2624 a las seis horas uno de los nodos con mayor grado es aquel que representa a un gen codificador de proteínas tubulares encargadas del tráfico intramolecular de sustancias (MyTH), que junto con otra proteína de membrana (B12D), actúan como orígenes de la red regulatoria, desde donde se desencadenan reacciones bioquímicas que llevan hacia la respuesta de tipo tolerante a estrés salino.Además, cuando hablamos de valores de Fold Change los cuales reflejaron las tasas de cambio que existe entre la expresión de un gen en distintos tiempos, podemos encontrar cómo estos disminuyen dramáticamente su valor, llegando a extremos como lo es el caso de HSP20 para en M2624 a los 3 días, donde alcanza un Fold Change de 28,8. Resulta interesante analizar el por qué aún coincidiendo M26 en DEGs con F12/1, este último no alcanza a responder adecuadamente ante estrés abiótico. Esto puede deberse a que, para responder a este estímulo y sobrevivir, se hace necesario un cambio en las tasas de expresión mucho más dramático del que se observa en F12/1, donde a las seis horas se observa una sub-expresión que solo alcanza un Fold Change de -9 en genes de tipo Heat Shock.

10. Conclusiones.

Con respecto a los resultados discutidos en el presente documento, se exponen los siguientes puntos importantes con respecto a la investigación:

- Cuando se exponen genotipos de portainjertos Mazzard F12/1 y Marianna 2624 a condiciones de alta salinidad en el suelo, estos genotipos presentan respuestas transcriptómicas al estrés salino que contrastantes entre ambos genotipos de portainjertos de *Prunus sp*.
- La respuesta contrastante entre portainjertos F12/1 y M2624 puede estar influenciada por la participación de elementos génicos en M2624 que no ven potenciada su expresión en F12/1. En específico se destaca la sobre-expresión de genes que codifican para factores de transcripción como factores WRKY y NAC en la respuesta a las seis horas, los cuales no se encuentran en ninguna otra red en F12/1. Las interacciones de estos factores de transcripción con proteínas membranales como B12D y Myth_4 resultan en la activación de varios genes que participan de manera coordinada ante la respuesta al estrés salino. Además, se

enfatiza la sobre-expresión de genes que participan en actividades de intercambio iónico membranal, metabolismo de ATP, y fosforilación de aminoácidos. Todos estos genes tienen interacción directa con M2624, el genotipo tolerante a estrés salino, a las cero horas y seis horas, en los tiempos más tempranos de exposición a alta salinidad.

- El hecho de que de que M2624 presente estos factores de transcripción que ya han sido ampliamente estudiados en respuesta a estrés abiótico de tipo sequía y salinidad en otras especies, permite validar un primer acercamiento a entender las distintas configuraciones transcriptómicas que especies de interés comercial adoptan a enfrentarse a distintos tipos de condiciones ambientales que afectan su desarrollo.
- Sí bien ya se han realizado estudios enfocados a la construcción de redes regulatorias en plantas y su respuesta a estrés abiótico, estudios de este tipo en especies importantes a nivel país como *Prunus cerasifera* y *Prunus avium* no se habían realizado anteriormente. El procedimiento llevado a cabo permitió reconocer una red de co-expresión en *Prunus cerasifera x munsoniana* caracterizada por la respuesta coordinada de genes que codifican para proteínas de membrana y factores de transcripción que no se configura en *Prunus avium,* respuesta transcriptómica diferencial que podría explicar el por qué *Prunus avium* presenta tan poca tolerancia con respecto a *Prunus cerasifera,* siendo estos individuos de la misma especie.
- La integración de herramientas bioinformáticas como análisis de expresión diferencial hacia transcriptomas resultan útiles para conocer los niveles de expresión que existen en genes importantes a nivel metabólico. A pesar de la utilidad de esta información, se hace necesario conocer no solo los niveles de expresión de genes sí no cómo interactúan estos entre sí, de esta forma se pueden descubrir funciones y procesos celulares asociados que pueden pertenecer a rutas metabólicas esenciales para los tipos de estrés que se están estudiando.

11. Referencias.

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data.*Genome Biology*, *11*(10), R106. https://doi.org/10.1186/gb-2010-11-10-r106

Anjum, A., Jaggi, S., Varghese, E., Lall, S., Bhowmik, A., & Rai, A. (2016). Identification of Differentially Expressed Genes in RNA-seq Data of Arabidopsis thaliana: A Compound Distribution Approach. *Journal of Computational Biology*, *23*(4), 239-247. https://doi.org/10.1089/cmb.2015.0205

Atkinson, T. J., & Halfon, M. S. (2014). REGULATION OF GENE EXPRESSION IN THE GENOMIC CONTEXT. *Computational and Structural Biotechnology Journal*, *9*(13), e201401001. https://doi.org/10.5936/csbj.201401001

Baillo, Kimotho, Zhang, & Xu. (2019). Transcription Factors Associated with Abiotic and Biotic Stress Tolerance and Their Potential for Crops Improvement. *Genes*, *10*(10), 771. https://doi.org/10.3390/genes10100771

Batushansky, A., Toubiana, D., & Fait, A. (2016). Correlation-Based Network Generation, Visualization, and Analysis as a Powerful Tool in Biological Studies: A Case Study in Cancer Cell Metabolism. *BioMed Research International*, *2016*, 1-9. https://doi.org/10.1155/2016/8313272

Belmar, J. L. D. (2016). *Aplicación de dos ácidos orgánicos y su efecto en la dinámica de las sales en el suelo*. 69.

Bianchi, V. J., Rubio, M., Trainotti, L., Verde, I., Bonghi, C., & MartÃnez-GÃ³mez, P. (2015). Prunus transcription factors: Breeding perspectives. *Frontiers in Plant Science*, *6*. https://doi.org/10.3389/fpls.2015.00443

Bubble products in sequencing libraries: Causes, identification, and workflow recommendations. (s. f.). Recuperado 21 de junio de 2021, de https://support.illumina.com/bulletins/2019/10/bubble-products-in-sequencing-libraries--causes--identification-.html

Burrows, M., & Wheeler, D. J. (s. f.). *A block-sorting lossless data compression algorithm*. 24.

Censi, F., Calcagnini, G., Bartolini, P., & Giuliani, A. (2010). A Systems Biology Strategy on Differential Gene Expression Data Discloses Some Biological Features of Atrial Fibrillation. *PLoS ONE*, *5*(10), e13668. https://doi.org/10.1371/journal.pone.0013668

Chen, S., Huang, T., Zhou, Y., Han, Y., Xu, M., & Gu, J. (2017). AfterQC: Automatic filtering, trimming, error removing and quality control for fastq data. *BMC Bioinformatics*, *18*(S3), 80. https://doi.org/10.1186/s12859-017-1469-3

Chu, Y., & Corey, D. R. (2012). RNA Sequencing: Platform Selection, Experimental Design, and Data Interpretation. *Nucleic Acid Therapeutics*, *22*(4), 271-274. https://doi.org/10.1089/nat.2012.0367

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, *17*(1), 13. https://doi.org/10.1186/s13059-016-0881-8

Contreras-López, O., Moyano, T. C., Soto, D. C., & Gutiérrez, R. A. (2018). Step-by-Step Construction of Gene Co-expression Networks from High-Throughput Arabidopsis RNA Sequencing Data. En D. Ristova & E. Barbez (Eds.), *Root Development* (Vol. 1761, pp. 275-301). Springer New York. https://doi.org/10.1007/978-1-4939-7747-5_21

Correlation Coefficient: Simple Definition, Formula, Easy Steps. (s. f.). Statistics How To. Recuperado 21 de junio de 2021, de https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/

Costa-Silva, J., Domingues, D., & Lopes, F. M. (2017). RNA-Seq differential expression analysis: An extended review and a software tool. *PloS One*, *12*(12), e0190152. https://doi.org/10.1371/journal.pone.0190152

Davidson, E., & Levin, M. (2005). Gene regulatory networks. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(14), 4935. https://doi.org/10.1073/pnas.0502024102

Deinlein, U., Stephan, A. B., Horie, T., Luo, W., Xu, G., & Schroeder, J. I. (2014). Plant salt-tolerance mechanisms. *Trends in Plant Science*, *19*(6), 371-379. https://doi.org/10.1016/j.tplants.2014.02.001

Emmert-Streib, F., Dehmer, M., & Haibe-Kains, B. (2014). Untangling statistical and biological models to understand network inference: The need for a genomics network ontology. *Frontiers in Genetics*, *5*. https://doi.org/10.3389/fgene.2014.00299

Foundation Plant Services. (s. f.). Recuperado 21 de junio de 2021, de https://fps.ucdavis.edu/treedetails.cfm?v=939

Gene Co-Expression Network—An overview | ScienceDirect Topics. (s. f.). Recuperado 21 de junio de 2021, de https://www.sciencedirect.com/topics/medicine-and-dentistry/gene-co-expression-network

GENIE3 vignette. (s. f.). Recuperado 21 de junio de 2021, de https://bioconductor.org/packages/devel/bioc/vignettes/GENIE3/inst/doc/GENIE3.html

Goldschmidt, E. E. (2014). Plant grafting: New mechanisms, evolutionary implications. *Frontiers in Plant Science*, *5*, 727. https://doi.org/10.3389/fpls.2014.00727

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644-652. https://doi.org/10.1038/nbt.1883

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., & Geurts, P. (2010). Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *PLoS ONE*, *5*(9), e12776. https://doi.org/10.1371/journal.pone.0012776

Huynh-Thu, V. A., & Sanguinetti, G. (2019). Gene Regulatory Network Inference: An Introductory Survey. *Methods in Molecular Biology (Clifton, N.J.)*, *1883*, 1-23. https://doi.org/10.1007/978-1-4939-8882-2_1

Interpret the key results for Correlation. (s. f.). [Mtbconcept]. Recuperado 21 de junio de 2021, de https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/how-to/correlation/interpret-the-results/

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: A fast spliced aligner with low memory requirements. *Nature Methods*, *12*(4), 357-360. https://doi.org/10.1038/nmeth.3317

Kovaka, S., Zimin, A. V., Pertea, G. M., Razaghi, R., Salzberg, S. L., & Pertea, M. (2019). Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biology*, *20*(1), 278. https://doi.org/10.1186/s13059-019-1910-1

Krasensky, J., & Jonak, C. (2012). Drought, salt, and temperature stress-induced metabolic rearrangements and regulatory networks. *Journal of Experimental Botany*, *63*(4), 1593-1608. https://doi.org/10.1093/jxb/err460

Li, B., & Dewey, C. N. (2011). RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*(1), 323. https://doi.org/10.1186/1471-2105-12-323

Li, W. (2015). Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences. En K. E. Nelson (Ed.), *Encyclopedia of Metagenomics* (pp. 173-177). Springer US. https://doi.org/10.1007/978-1-4899-7478-5_221

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Maas, E. V., & Hoffman, G. J. (1977). Crop Salt Tolerance—Current Assessment. *Journal of*

the Irrigation and Drainage Division, *103*(2), 115-134. https://doi.org/10.1061/JRCEA4.0001137

Microarrays Factsheet. (2007, octubre 29). https://web.archive.org/web/20071029145451/http://www.ncbi.nlm.nih.gov/About/primer/microarrays.html

Nawaz, M. A., Imtiaz, M., Kong, Q., Cheng, F., Ahmed, W., Huang, Y., & Bie, Z. (2016). Grafting: A Technique to Modify Ion Accumulation in Horticultural Crops. *Frontiers in Plant Science*, *7*. https://doi.org/10.3389/fpls.2016.01457

Park, J., Xu, K., Park, T., & Yi, S. V. (2012). What are the determinants of gene expression levels and breadths in the human genome? *Human Molecular Genetics*, *21*(1), 46-56. https://doi.org/10.1093/hmg/ddr436

Per Base Sequence Quality. (s. f.). Recuperado 21 de junio de 2021, de https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/2%20Per%20Base%20Sequence%20Quality.html

Pertea, G., & Pertea, M. (2020). GFF Utilities: GffRead and GffCompare. *F1000Research*, *9*, ISCB Comm J-304. https://doi.org/10.12688/f1000research.23297.2

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, *33*(3), 290-295. https://doi.org/10.1038/nbt.3122

Pimentel, P., Almada, R. D., Salvatierra, A., Toro, G., Arismendi, M. J., Pino, M. T., Sagredo, B., & Pinto, M. (2014). Physiological and morphological responses of Prunus species with different degree of tolerance to long-term root hypoxia. *Scientia Horticulturae*, *180*, 14-23. https://doi.org/10.1016/j.scienta.2014.09.055

Pulido Madrigal, L., & Pulido Madrigal, L. (2016). Cambio climático, ensalitramiento de suelos y producción agrícola en áreas de riego. *Terra Latinoamericana*, *34*(2), 207-218.

Samtools-sort(1) manual page. (s. f.). Recuperado 21 de junio de 2021, de http://www.htslib.org/doc/samtools-sort.html

Sequencing Quality Scores. (s. f.). Recuperado 21 de junio de 2021, de https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/quality-scores.html

Srivastava, P., Wu, Q.-S., & Giri, B. (2019). Salinity: An Overview. En B. Giri & A. Varma (Eds.), *Microorganisms in Saline Environments: Strategies and Functions* (Vol. 56, pp. 3-18). Springer International Publishing. https://doi.org/10.1007/978-3-030-18975-4_1

Staton, E. (2021). *Sestaton/HMMER2GO* [Perl]. https://github.com/sestaton/HMMER2GO (Original work published 2014)

Tan, G., Opitz, L., Schlapbach, R., & Rehrauer, H. (2019). Long fragments achieve lower base quality in Illumina paired-end sequencing. *Scientific Reports*, *9*(1), 2856. https://doi.org/10.1038/s41598-019-39076-7

Tavakkoli, E., Rengasamy, P., & McDonald, G. K. (2010). High concentrations of Na+ and Cl– ions in soil solution have simultaneous detrimental effects on growth of faba bean under salinity stress. *Journal of Experimental Botany*, *61*(15), 4449-4459. https://doi.org/10.1093/jxb/erq251

Tolerancia de los cultivos a la Salinidad. (s. f.). Recuperado 22 de junio de 2021, de http://agrosal.ivia.es/tolerancia.html

TransDecoder/TransDecoder. (2021). [Perl]. TransDecoder. https://github.com/TransDecoder/TransDecoder (Original work published 2015)

Tin, M.M.Y., Rheindt, F.E., Cros, E. and Mikheyev, A.S., 2015. Degenerate adaptor sequences for detecting PCR duplicates in reduced representation sequencing data improve genotype calling accuracy. *Molecular ecology resources*, *15*(2), pp.329-336. doi:10.1111/1755-0998.12314

Wang, R., Cheng, Y., Ke, X., Zhang, X., Zhang, H., & Huang, J. (2020). Comparative analysis of salt responsive gene regulatory networks in rice and Arabidopsis. *Computational Biology and Chemistry*, *85*, 107188. https://doi.org/10.1016/j.compbiolchem.2019.107188

Wingett, S. W., & Andrews, S. (2018). FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Research*, *7*, 1338. https://doi.org/10.12688/f1000research.15931.2

Yan, Y., Wang, S., Wei, M., Gong, B., & Shi, Q. (2018). Effect of Different Rootstocks on the Salt Stress Tolerance in Watermelon Seedlings. *Horticultural Plant Journal*, *4*(6), 239-249. https://doi.org/10.1016/j.hpj.2018.08.003

Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, *4*, Article17. https://doi.org/10.2202/1544-6115.1128