**UNIVERSIDAD NACIONAL DE ROSARIO**

**Facultad de Ciencias Bioquímicas y Farmacéuticas**

**Facultad de Ciencias Agrarias**

CONICET

UNIVERSIDAD NACIONAL DE ROSARIO

C E F O B I

**Centro de Estudios Fotosintéticos y Bioquímicos**

# "Estudios *in silico* de la expresión génica relativa a factores protectores frente al daño por frío en duraznos"

*Especialización en Bioinformática*

***Lic. en Biotecnología*** **Mauro Gismondi**

*Tutor:* ***Dr. Lucas Daurelio***
*Co-tutor:* ***Dr. Luis Esteban***

2016

## *Agradecimientos*

A las doctoras Claudia Bustamante y Ma. Fabiana Drincovich por confiar en mí y permitirme explorar el mundo de la Bioinformática durante mi tesis doctoral, para que hoy sea parte de mis conocimientos. A los doctores Lucas Daurelio y Luis Esteban por acompañarme en el desarrollo del trabajo final, por su dedicación y sus consejos.

A todos los que formaron parte de la Especialización en Bioinformática, por seguir hoy en dia esforzándose y sumando a la investigación en Biología mediante su integración con la Informática y la Estadística. En especial, a Iva, Mauri y Flo, y a Estefi por su ayuda y por compartir excelentes momentos.

A Mai, por acompañarme donde sea y apoyarme en todo lo que me propongo…

A los chicos del subsuelo del CEFOBI y a todos los que escucharon alguna vez una excusa relacionada con el trabajo de Bioinformática…

# *Índice*

## *Abreviaturas y términos especiales*

| | |
|---|---|
| 2D-DIGE-MS/MS | Bidimensional Difference gel electrophoresis seguida de Tandem Mass Spectrometry. Técnica proteómica para detectar e identificar PEDs |
| bs | Binding site. Sitio de unión de factor de transcripción proteico en una molécula de ADN |
| ADNc | ADN codificante generado por retrotranscripción |
| CDS | ADNc sin regiones 5' y 3'UTR |
| Da (Dalton) | Unidad estándar de masa atómico o molecular |
| TED (del inglés, DET) | Transcripto de expresión diferencial |
| Differential Display (DD) | Técnica transcriptómica para detectar TEDs |
| EC ID | Enzyme Commission Number |
| Match | Aparición de un motivo buscado en una secuencia target |
| OE | Overexpressing transgenic plant |
| pb (del inglés, bp) | Pares de bases de una secuencia nucleotídica |
| pI | Punto isoeléctrico |
| PED | Péptido de expresión diferencial |
| PM (o MW) | Peso molecular |
| RT-qPCR | Retrotranscripción seguida de PCR en tiempo real. Técnica que permite evaluar niveles transcripcionales de un dado transcripto |
| TT (HAT o HWT) | Tratamiento térmico por aire (HAT) o agua (HWT) |
| Western blot | Técnica proteómica para inmunodetección específica de polipéptidos |
| Accession | Código de acceso relativo a genes en bases de datos genómicas |
| UTR | Región no traducible del ARNm (del inglés, *UnTranslated Region*) |

# *1. Introducción*

## 1.1. Poscosecha en duraznos

Una de las especies pertenecientes a la familia *Rosaceae* es el duraznero, *Prunus persica* (L.) Batsch. El consumo de su fruto es mundialmente popular debido a sus atractivos organolépticos y su alto valor nutricional. El durazno es un fruto climatérico, por lo que es capaz de seguir madurando incluso separado de la planta. A pesar de que esta especie ha evolucionado a la abscisión natural, la cosecha implica un estrés que deja a la fruta a expensas de sus propias reservas. Sumados a los procesos biológicos internos, factores externos imperantes durante cada etapa productiva (incluidas las condiciones ambientales y las prácticas relativas a cultivo, cosecha y poscosecha), determinan la vida de estantería y la calidad organoléptica del fruto como alimento.

Debido al rápido deterioro de los frutos a temperatura ambiente, la **refrigeración** se ha vuelto indispensable para lentificar la maduración manteniendo las propiedades alimenticias y organolépticas deseables, y a la vez extender la vida útil de la fruta. El almacenamiento refrigerado se realiza a en condiciones de 0 °C de temperatura y 85 a 90 % de humedad relativa, lo que permite preservar la calidad y el valor nutritivo por 2 a 4 semanas.

Sin embargo, dependiendo de la temperatura y duración del almacenamiento refrigerado, así como también de las condiciones precosecha, algunas variedades de duraznero se ven afectadas más que otras por una serie de desórdenes fisiológicos genéricamente englobados en la denominación **daño por frío** (chilling injury o CI). Estas afecciones comprometen la calidad de los duraznos como alimento y, agravando la problemática, son observables al momento del consumo y difícilmente predecibles en las etapas previas de la cadena de comercialización (Lurie y Crisosto, 2005).

La harinosidad, por ejemplo, es un trastorno textural en la pulpa de la fruta que se manifiesta principalmente como pérdida de la jugosidad. No es posible detectarla en los frutos desde el exterior sin abrirlos y, al hacerlo, la falta de jugosidad produce una importante pérdida en el sabor debido a que vuelve a la pulpa pastosa y seca, generando insatisfacción en el consumidor (Brummel y col., 2004). La figura 1.1 nos permite apreciar los efectos de este desorden.

**Figura 1.1. Manifestación de harinosidad en duraznos (INTA-Alto Valle, Revista: Rompecabezas Tecnológico Nº 42).**

Han surgido numerosas estrategias, previas o simultáneas a la refrigeración, exitosas en la protección de la fruta frente al daño por frío. Se han estudiado almacenamientos en atmósferas controladas, tratamientos térmicos y adiciones de diferentes compuestos químicos al entorno de la fruta (Lurie y Crisosto, 2005). En particular, el **tratamiento térmico** a altas temperaturas previo a la refrigeración se ha demostrado útil para contrarrestar al daño por frío (y otras afecciones poscosecha) en frutos de variadas especies vegetales, modulando la velocidad de maduración, y preservando apariencia y propiedades nutricionales del alimento (Lurie, 1998). Pedreschi y Lurie (2015) revisan los diversos procesos biológicos de la fruta posiblemente afectados frente a estos cambios de temperatura; sin embargo, hasta el momento, se desconocen los mecanismos moleculares subyacentes de la protección así como del mismo daño por frío.

### 1.2. Antecedentes de investigación y presente trabajo

Con el fin de lograr una solución de mayor alcance al problema del daño por frío en diferentes especies frutales, es de especial interés entender las bases genéticas y bioquímicas de este grupo de trastornos y de los tratamientos que los contrarrestan. Diversos enfoques se han utilizado para comprender la fisiología de este tipo de estreses de poscosecha y su compleja regulación en frutos de importancia alimenticia (Pedreschi y Lurie, 2015).

En nuestro laboratorio y tomando la variedad Dixiland (Estación Experimental (EEA) San Pedro del INTA) como modelo, se han analizado las modificaciones bioquímicas y proteómicas (Lara y col., 2009), transcriptómicas (Lauxmann y col., 2012), metabolómicas (Lauxmann y col., 2014) y de la matriz extracelular (Bustamante y col., 2012), luego de aplicar tratamiento térmico (3 días a 39 °C) a frutos de durazno cosechados. De esta manera, lograron identificarse diversos genes que contribuirían a la protección de la fruta frente al daño por frío.

En este contexto se enmarca el presente trabajo final de la Especialización en Bioinformática. Considerando que los efectos benéficos del tratamiento térmico frente al daño por frío podrían deberse (entre otros) a la modificación de las tasas de transcripción de los genes identificados, es de interés ahondar en el conocimiento de los requisitos moleculares que cada uno de ellos posee para su expresión génica a este nivel. Para ello, este trabajo propone la caracterización *in silico* de las secuencias upstream (promotores y 5' UTRs) de genes con expresión génica diferencial frente al tratamiento térmico en duraznos. Estos estudios comprenden la detección y el análisis de la presencia de motivos lineales en las mencionadas secuencias, que podrían suponer sitios de unión funcionales para factores de transcripción de durazno, con roles importantes en los procesos biológicos de interés.

### 1.3. Inicio transcripcional, pattern matching y cisAnalyzer

El dogma central de la Biología Molecular como concepto, plantea los mecanismos de transmisión y expresión génica que involucran a las moléculas ADN, ARN y proteína (Crick, 1958). El mismo postula que solo la molécula de ADN es sometida al proceso de replicación y que la misma contiene (en unidades conocidas como genes) la información que codifica para la transcripción de diferentes especies de ARN. Una de ellas, el ARNm (ARN mensajero) es materia prima para producción de polipeptidos o proteínas mediante el proceso de traducción. Estos procesos moleculares, ilustrados en la figura 1.2 y sumados a modificaciones hoy descubiertas, son los principales que rigen formación, desarrollo y funcionamiento de toda célula viva.

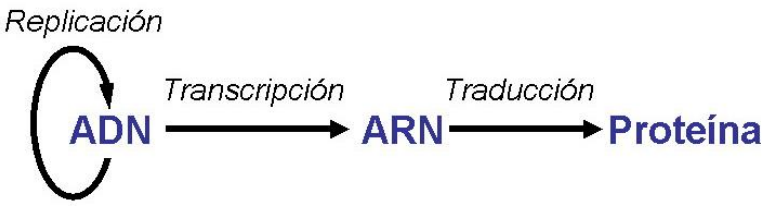


**Figura 1.2. Dogma central de la Biología Molecular (Crick, 1958)**

La expresión génica comienza entonces con el proceso de transcripción que consiste en la síntesis de ARN utilizando como molde la secuencia nucleotídica de un gen (un segmento conciso de ADN con su secuencia particular). Un gran desafío en investigación en biología molecular es entender los mecanimos regulatorios de la expresión génica.

En la región upstream o corriente arriba de un gen, adyacente a su inicio transcripcional, son reclutados y ensamblados variados factores moleculares y el supercomplejo enzimático ARN polimerasa (**elementos *trans***). Este reclutamiento es dependiente de ciertos requisitos estructurales y químicos de la cromatina, y uno de ellos es la presencia y/o el arreglo espacial de motivos específicos en zonas proximales y/o distales de la molécula de ADN (**elementos *cis***). Los términos *cis* y *trans* provienen del latín: son las características químicas de los motivos presentes en *cis* ("en el mismo lado" del ADN, en el mismo gen), las que permiten la unión de ciertos factores moleculares y de la ARN polimerasa en *trans* (provenientes de "otro lado" por estar codificados por otros genes) al sitio de inicio transcripcional para permitir su regulación. Un dado factor de transcripción es una proteína que puede reconocer e interaccionar con un grupo de elementos de ADN que comparten una secuencia consenso, el motivo o elemento *cis*.

Por todo esto, **el inicio de la transcripción es el primer punto de control de la expresión génica** y las regiones upstream de un gen, mediante el conjunto y el arreglo de elementos *cis* funcionales que contienen, suelen denominarse interruptores génicos.

La figura 1.3 muestra un esquema del inicio transcripcional del gen X, y la interacción de elementos *cis* y *trans* que lo permiten.

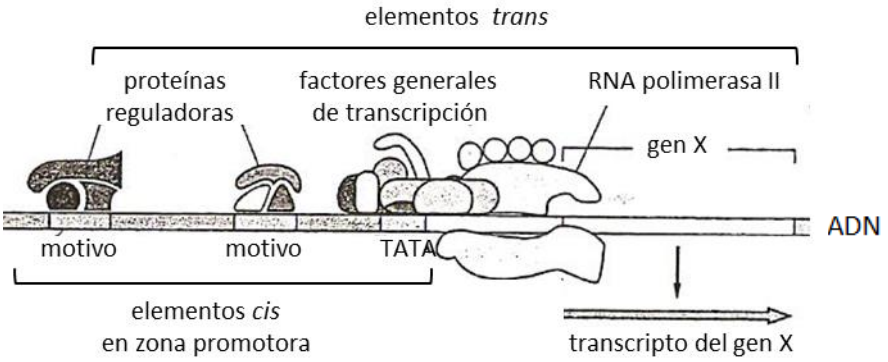


**Figura 1.3. Esquema simplificado del inicio transcripcional del gen X.** Se observan elementos *cis* y *trans* en un promotor ejemplo**.** La caja TATA es un elemento *cis* general, común a muchos promotores eucariotas; la misma dicta la unión de factores generales de transcripción y de la ARN polimerasa II (complejo mutiproteico encargado de iniciar la transcripción). Otros motivos en la zona promotora actúan como reclutadores de proteínas reguladoras (incluidos otros factores de transcripción) que, en conjunto, modulan la regulación del inicio transcripcional. Adaptado de *Molecular Biology of the Cell* (Alberts y col., 2011).

La identificación de los elementos *cis* y *trans* cuyas interacciones determinan la transcripción de un gen dado, son críticas para entender la regulación de este proceso biológico. Variados algoritmos y servicios web han surgido para identificar motivos en *cis*, usando como input a las secuencias promotoras (Hertz y col., 1999; Lawrence y col., 1993; Thompson y col., 1993; Peng y col., 2006). El pattern matching (identificación de patrones), fácilmente aplicable a través de diferentes comandos parte de muchos lenguajes de programación (incluido Perl), es extremadamente útil para el reconocimiento de motivos lineales en cadenas de caracteres de interés. Este tipo de búsqueda posee mucho potencial para la caracterización de secuencias de ADN promotoras en busca de motivos conocidos o desconocidos, y para el análisis de enriquecimiento y coocurrencia de los mismos.

Tomando como iniciativa el objetivo de caracterizar *in silico* las secuencias upstream de los genes identificados de respuesta al tratamiento térmico, el presente trabajo permitió el desarrollo de **cisAnalyzer**: un programa bioinformático desarrollado en lenguaje Perl, que asiste a un usuario sin conocimientos en programación en la detección y el análisis de la presencia de motivos lineales en grupos de secuencias biológicas. En particular, cabe destacarse que el investigador en biología podrá, a través de este programa, caracterizar sets de secuencias a través de la presencia, el enriquecimiento y la coocurrencia de uno o más motivos.

## 1.4. Objetivos

### 1.4.1. Objetivo general

Realizar un análisis *in silico* de la regulación de la expresión génica de factores protectores del daño por frío en duraznos, frente a tratamiento térmico. ADN

### 1.4.2. Objetivos específicos

Desarrollar un programa para analizar motivos lineales en secuencias promotoras de genes con expresión transcripcional y proteica modificada en respuesta a tratamiento térmico.

Integrar la información obtenida a partir de las secuencias promotoras con los datos de expresión génica previos, con el fin de dilucidar elementos *cis* y factores de transcripción implicados en la protección frente al daño por frío.

# *2. Materiales y métodos*

## 2.1. Construcción de una base de datos de determinantes moleculares posiblemente relacionados a la protección frente al daño por frío en duraznos

Los artículos de Lara y col. (2009), Lauxmann y col. (2012 y 2014), y Bustamante y col. (2012), aplicaron diversas técnicas ómicas que permitieron la identificación de variados determinantes moleculares posiblemente implicados en la protección frente al daño por frío en duraznos. Los resultados de todos ellos fueron agrupados y organizados en cuanto a la información de expresión génica que proveyeron acerca de los genes estudiados. Por otro lado, resultó interesante incluir dos trabajos adicionales sobre tratamiento térmico en frutos de durazno, ambos pertenecientes a otros grupos de investigación: Zhang y col. (2011) aplican la técnica proteómica (2D-DIGE-MS/MS) en duraznos cv. Huiyulu sumergidos en agua caliente, y Jin y col., (2009) realizaron medidas de diversas actividades interesantes en frutos de *P.persica* cv. Baifeng mantenidos bajo aire caliente.

Las muestras utilizadas en los trabajos de nuestra línea de investigación fueron frutos pertenecientes a la especie *Prunus persica* (L.) Batsch y a la variedad Dixiland (DX), cosechados durante las temporadas 2007-2008, y 2008-2009. Las mismas provinieron de un lote experimental de la Estación Experimental (EEA) San Pedro del INTA (33° 44´ 12,1´´ de latitud Sur y 59° 47´48,0´´ de longitud Oeste). Se incluyeron variadas condiciones para cada artículo publicado, aprovechándose las estrategias en tecnología de poscosecha (tratamiento térmico y vida de estantería) y dependiendo de su finalidad. En particular, para la identificación de determinantes posiblemente protectores frente al daño por frío, se emplearon técnicas ómicas luego de la aplicación de una estrategia poscosecha preventiva como lo es el tratamiento térmico.

La tabla 2.1 describe, a modo general, las condiciones poscosecha utilizadas por los diferentes trabajos mencionados (incluyendo los trabajos de otros grupos de investigación).

**Tabla 2.1. Condiciones poscosecha muestreadas**

| Variedad(es) | Temporada(s) | Condición(es) poscosecha | Descripción |
|---|---|---|---|
| **Dixiland** | 2007-2008<br>2008-2009 | **H** | Harvest. Duraznos cosechados alcanzada la madurez fisiológica |
| | | **SL1, 3, 5, 7** | Shelf-Life. Duraznos cosechados y almacenados durante 1 / 3 / 5 / 7 días a 20 °C |
| | | **HAT** | Hot Air Treatment. Duraznos cosechados y almacenados durante 3 días a 39 °C. Tratamiento térmico vía aire. |
| | | **HAT+SL3, 7** | Duraznos cosechados, almacenados durante 3 / 7 días a 39 °C y luego mantenidos por 3 días a 20 °C |
| | | **HAT+CS2** | Duraznos cosechados, almacenados durante 3 días a 39 °C y luego refrigerados por 2 días a 0 °C. Combinación. |
| | | **CS3, 5** | Cold Storage. Duraznos cosechados y almacenados durante 3/5 días a 0 °C. Refrigeración corta |
| | | **CS5+SL2** | Duraznos cosechados, almacenados durante 5 días a 0 °C y luego mantenidos por 2 días a 20 °C. |
| **Huiyulu** | 2010 | **H** | Duraznos cosechados alcanzada la madurez fisiológica (Harvest) |
| | | **SL1, 3, 5** | Duraznos cosechados y almacenados durante 1 / 3 / 5 / 7 días a 20-25 °C (Shelf-Life) |
| | | **HWT+SL1, 3, 5** | Duraznos cosechados, sumergidos por 10 min en agua a 48 °C (Hot Water Treatment), secados mediante flujo de aire y mantenidos por 1 / 3 / 5 días a 20-25 °C |
| **Baifeng** | 2008 | **SL1+CS21, 35** | Duraznos cosechados alcanzada la madurez fisiológica, mantenidos a 20 °C por 1 día y refrigerados por 3 ó 5 semanas a 0 °C |
| | | **MJ+CS21, 35** | Duraznos cosechados, mantenidos en cámara con 1 µmol / L de vapor de MJ (Metil Jasmonato) a 20 °C por 1 día y refrigerados por 3 ó 5 semanas a 0 °C |
| | | **HAT+CS21, 35** | Duraznos cosechados, almacenados durante 12 horas a 38 °C en cámara y refrigerados por 3 ó 5 semanas a 0 °C |
| | | **HMJ+CS21, 35** | Duraznos cosechados, mantenidos en cámara con 1 umol / L de vapor de MJ a 38 °C por 12 horas y refrigerados por 3 ó 5 semanas a 0 °C |
| | | **HAT+MJ+CS21, 35** | Duraznos cosechados, almacenados durante 12 horas a 38 °C, luego mantenidos en cámara con 1 umol / L de vapor de MJ a 20 °C por 1 día y refrigerados por 3 ó 5 semanas a 0 °C |

A modo de resumen, la tabla 2.2 describe los artículos incluidos para la construcción de la base de datos de interés y el tipo de determinantes identificados en cada uno de ellos.

**Tabla 2.2. Artículos elegidos para construir la base de determinantes moleculares de durazno**

| Referencia | Moléculas y Métodos | Variedad \| Condiciones poscosecha | Principales decubrimientos |
|---|---|---|---|
| **Lara y col., 2009** | Transcriptos (RT-qPCR) Metabolitos (Actividades y GC-MS) Péptidos (Western blot, 2D-DIGE-MS/MS) Actividades enzimáticas (Espectrofotometrías) | DX \| H, SL3, SL5, SL7, HAT, HAT+SL3 | Aumento de proteínas de estrés/defensa y disminución de PPO luego de HAT |
| **Lauxmann y col., 2012** | Transcriptos (Differential Display y RT-qPCR) | DX \| H, SL3, SL7, HAT, HAT+SL3, HAT+SL7, CS3, CS5, CS5+SL2 | 127 transcriptos diferencialmente expresados (47% inducidos y 36% reprimidos) luego de HAT |
| **Bustamante y col., 2012** | Transcriptos (RT-qPCR) Péptidos (Western blot, 2D-DIGE-MS/MS) | DX \| H, SL3, HAT | Disminución de transcriptos de enzimas de pared celular y aumento de PpDUF642 y PpGAPDH luego de HAT |
| **Lauxmann y col., 2014** | Transcriptos (RT-qPCR) Metabolitos (GC-MS) | DX \| H, SL3, SL7, HAT, HAT+SL3, HAT+SL7, CS3, CS5, CS5+SL2 | Aumento de azúcares y azúcares-alcoholes, y modificación de aminoácidos precursores de la vía de los fenilpropanoides luego de HAT |
| **Zhang y col., 2011** | Péptidos (2D-DIGE-MS/MS) | Huiyulu \| H, SL1, SL3, SL5, HWT+SL1, HWT+SL3, HWT+SL5 | Aumento de proteínas de estrés/defensa y de estructura celular luego de HWT |
| **Jin y col., 2009** | Actividades enzimáticas (Espectrofotometrías) | Baifeng \| SL1+CS21/35, MJ+CS21/35, HAT+CS21/35, HMJ+CS21/35, HAT+MJ+CS21/35 | Ambas combinaciones de MJ y HAT resultaron en la prevención de l daño por frío. Aumentan actividades PAL, SOD, PG y disminuyen PPO y POD |

Reunidos los determinantes moleculares (péptidos, polipéptidos, fragmentos de ADNc, actividades enzimáticas) identificados, los datos de expresión génica y las muestras en las que sus patrones y niveles moleculares fueron estudiados, se procedió a completar la base de datos realizando la asignación de los mismos a códigos de acceso génicos y transcripcionales específicos, pertenecientes al genoma de duraznero (The International Peach Genome Initiative, 2013; versión 1.0). El último es accesible a través de las bases de datos GDR (Genome Database of Rosaceae,

http://www.rosaceae.org/; Jung y col., 2014) y Phytozome (http://www.phytozome.net/; Goodstein y col., 2011).

Para la mencionada asignación se aplicaron diversas estrategias visualizables en la tabla 2.3, que dependieron exclusivamente de la naturaleza molecular de los datos obtenidos por cada método experimental. En los casos en que se utilizó el algoritmo blast de NCBI, el mismo se ejecutó de manera autónoma desde consola Linux (Tao, 2010). Las bases de datos de trabajo fueron obtenidas a través de la plataforma Biomart (accesible desde la web de Phytozome) y formateadas para su uso mediante el comando makeblastdb (Tao, 2010).

**Tabla 2.3. Estrategias utilizadas para la asignación de códigos de acceso**

| Molécula | Método/s experimental/es | Estrategia de asignación a códigos de acceso génico/s y transcripcional/es |
|---|---|---|
| **Fragmentos de ADNc** | Differential Display RT-qPCR | ***blastn* contra bases de datos de ADNc, CDS, 5'UTR y 3'UTR**<br>***blastx* contra bases de datos de péptidos** |
| **Péptidos** | 2D-DIGE-MS/MS | ***blastp* contra bases de datos de péptidos**<br>**Verificación de PMs y pIs teóricos y experimentales** |
| **Polipéptidos** | Western blot | ***blastp* contra bases de datos de péptidos de GDR**<br>**Verificación de PMs teóricos y experimentales** |
| **Actividades enzimáticas** | Ensayos espectrofotométricos | **Filtrado de genes con EC ID anotado a través de Biomart (Phytozome)** |

Una vez completa y extensivamente revisada, la base de datos se complementó con anotaciones funcionales de los genes asignados, mediante herramientas provenientes de Biomart (Phytozome). Por otro lado, se decidió un criterio para la clasificación de los genes de duraznero estudiados, que contempló naturaleza molecular del determinante (transcripto, polipéptido, actividad enzimática) y modificación de sus niveles frente a maduración organoléptica y tratamiento térmico.

## 2.2. Construcción de una base de datos de sitios de unión a factores de transcripción conocidos en diversas especies vegetales: PASF db

A partir de las bases de datos de los sitios web PlantPAN (Plant Promoter Analysis Navigator, http://plantpan.mbc.nctu.edu.tw/; Chang y col., 2008) y footprintDB (http://floresta.eead.csic.es/footprintdb/; Sebastian y col., 2013), como también a partir de bibliografía, se realizó una selección de elementos *cis* conocidos y caracterizados de ser sitios de unión de factores de transcripción vegetales,

funcionales bajo diferentes estímulos que las especies vegetales enfrentan. Cabe destacar que la base de elementos *cis* de PlantPAN nuclea otras bases particulares, algunas de ellas tradicionalmente utilizadas para análisis de promotores génicos de plantas como: PLACE (Higo y col., 1999), AGRIS (Davuluri y col., 2003), TRANSFAC (Wingender y col., 2000), JASPAR (Sandelin y col., 2004), PlantCARE (Lescot y col., 2002).

En particular, se consideraron elementos *cis* de variadas especies vegetales que hayan sido comprobados experimentalmente (*in vivo* e/o *in vitro*) de ser targets de factores de transcripción. Se verificó que su longitud sea de al menos 5 pb, aumentando las probabilidades de que la unión de proteína/s verificada no sea al azar. Las referencias asociadas a cada motivo lineal de interés fueron revisadas exhaustivamente para recopilar la mayor información experimental posible acerca de:

los elementos *cis* y su caracterización experimental,

los factores de trascripción que interaccionarían con ellos en *trans* y su estudio,

los procesos biológicos bajo los que cumplirían sus roles.

Teniendo en cuenta bajo qué estímulo/s (estrés abiótico o señalización de fitohormona) desempeñarían sus roles cada par de elementos *cis* y *trans,* se construyeron manualmente 18 grupos de motivos lineales, cada uno relacionado a un dado proceso biológico.

### 2.3. Creación de un programa de Perl para análisis de motivos lineales en secuencias biológicas: cisAnalyzer

En vista de cumplir con los objetivos planteados, se trabajó en el diseño del programa deseado, utilizando el lenguaje de programación Perl (Practical Extraction and Report Language; Perl 5.18.2), con la asistencia de la bibliografía (Wall y col., 2000; Tisdall, 2001) y los sitios web oficiales https://www.perl.org/, http://perldoc.perl.org/, http://www.cpan.org/. Perl es un reconocido lenguaje para el procesamiento de texto, por lo que es común su elección para el manejo de secuencias biológicas. Una de las razones más importantes por los que Perl es útil en este sentido es la posibilidad de trabajar con expresiones regulares, que permiten altas flexibilidad y eficiencia por parte de este lenguaje.

Diferentes estrategias, elementos y comandos propios de Perl fueron empleados para las metas particulares que cada etapa del programa se propone. Se utilizaron subrutinas y módulos que permiten diseñar el código de una manera simple; se hizo necesario incluir diversas acciones de input

por parte del usuario que van seteando sus decisiones de análisis y la generación de archivos output que el programa va reutilizando o que son parte de los resultados finales; se aplicaron diferentes comandos para llamadas al sistema desde Perl y manejo de secuencias como sustitución, transliteración, reversión, pattern matching (exacto y con variantes) y se usaron diversos elementos como escalares, listas (arrays) y hashes. La programación realizada buscó aprovechar las características de los archivos necesarios como input debido a que siempre es requerido insertar un archivo con las secuencias target de pattern matching y, según las decisiones del usuario, también puede requerirse incluir archivos con UTRs y/o motivos propios. De esta manera, se consideraron ciertos rasgos de los archivos fasta de secuencias targets del pattern matching, directamente provenientes de la plataforma Biomart, Phytozome. Relacionado a esto, se debió tener en cuenta el posible uso de genomas con limitado acceso en la actualidad (se verificó el permiso para cada genoma incluido en Phytozome hasta el mes de diciembre de 2015, aunque el programa lo deja en manos del usuario).

### 2.4. Selección de grupos de secuencias upstream de duraznos y de motivos lineales de interés para llevar a cabo análisis mediante cisAnalyzer

Con el fin de realizar los estudios propuestos, se procedió a obtener las regiones promotoras de los genes de duraznero de interés. Se definió tomar secuencias de 1500 pb upstream a partir del codón de inicio traduccional ATG (incluyendo las regiones 5' UTR) para caracterizar el sitio de inicio transcripcional por completo. Para obtener resultados correspondientes a cada grupo de determinantes moleculares con características comunes en cuanto a su expresión génica frente al tratamiento térmico, se descargaron los promotores y 5'UTRs pertenecientes a:

- transcriptos **inducidos** por TT (RT-qPCR y/o DD) **DET-I**
- transcriptos **reprimidos** por TT (RT-qPCR y/o DD) **DET-R**
- transcriptos inducidos o reprimidos (**afectados**) por TT (RT-qPCR y/o DD) **DET-A**
- péptidos o proteínas **inducidos** por TT (Western blot o 2D-DIGE-MS/MS) **DEP-I**
- péptidos o proteínas **reprimidos** por TT (Western blot o 2D-DIGE-MS/MS) **DEP-R**
- transcriptos y péptidos o proteínas **inducidos** por TT (DET-I + DEP-I) **DETyP-I**

Por separado (y por requerimientos del programa), se descargaron las secuencias 5' UTR de los mismos genes. Para aquellos genes con variantes transcripcionales anotadas, se trabajó solamente con las secuencias promotoras tomadas de las variantes transcripcionales primarias..

Adicionalmente, se tomaron secuencias promotoras + 5'UTRs y 5'UTRs para sets de genes que actuarán como controles. Se descargaron las secuencias correspondientes al set completo de transcriptos de duraznero (47089 secuencias en los archivos *AllPpe1500.txt* y *AllPpeUTRs.txt*) y luego se construyeron grupos de secuencias tomando tantos códigos de acceso transcripcionales al azar, como tenga el set a analizar (ejemplo: el set VsDET-I posee tantas secuencias como DET-I sólo que fueron elegidas aleatoriamente y no tendrían relación biológica). El set de promotores completo de duraznero permitirá comparar el enriquecimiento de cada motivo encontrado en los grupos de secuencias de interés con su ocurrencia en el genoma mientras que los sets particulares construidos al azar actuarán como controles de que el enriquecimiento no ocurra en un grupo equivalente de secuencias y solo se deba a la relación biológica que existe entre promotores de los transcriptos de interés.

Los multifasta de secuencias promotoras+5'UTR así como sus equivalentes de 5'UTR, para todos los sets de genes de trabajo (construidos a partir de los determinantes identificados y controles), se encuentran en el archivo comprimido *cisA_Ppe_FastasAndResults.zip.* Las denominaciones antes mencionadas, también presentes en el archivo *cisA_Ppe_Analysis_Names.xls*, permiten identificarlos.

Finalmente, se procedió como se describe en el tutorial asociado a cisAnalyzer, resumido en los siguientes pasos: se colocaron los archivos fasta con extensión .txt de secuencias target de búsqueda y de sus correspondientes secuencias 5'UTR en las carpetas indicadas, se seleccionó el grupo de motivos relacionado a estímulo calor o frío de PASF db, y se realizaron los análisis en sí.

# *3. Resultados y Discusión*

## 3.1. Creación de una base de datos de determinantes moleculares relacionados a la protección frente al daño por frío en duraznos

El presente trabajo es parte de una línea de investigación cuyo propósito es identificar y caracterizar determinantes moleculares involucrados en la protección frente al daño por frío en frutos de durazno. Con este fin, se han aplicado en nuestro laboratorio y en otros, diversas técnicas ómicas que proveyeron múltiples evidencias a nivel molecular de los procesos subyacentes al tratamiento térmico, una estrategia aplicada con éxito para la prevención de los trastornos asociados al daño por frío en duraznos.

Con las mencionadas evidencias, se decidió en este trabajo crear una base de datos con los determinantes identificados luego de tratamiento térmico en duraznos por nuestro laboratorio y otros grupos que estudiaron la misma problemática.

Primeramente, se incluyeron los determinantes, el/los método/s mediante los que fueron detectados y la referencia al trabajo. En segundo lugar, dado que las diversas técnicas incluidas arrojan diferentes resultados moleculares, se utilizaron los mismos vía estrategias variadas para asignar, a cada determinante, al menos un código de acceso génico y transcripcional del genoma de duraznero (tablas 2.2 y 2.3 de Materiales y Métodos). Fragmentos de ADNc obtenidos por Differential Display y amplicones de RT-qPCR fueron sujeto de blastn contra la base de datos completa de transcriptos de *P. persica*; péptidos, sus PMs y sus pIs obtenidos por 2D-DIGE-MS/MS fueron alineados por blastp contra la base de datos completa de péptidos de *P. persica;* proteínas específicas detectadas por western blot y actividades enzimáticas ensayadas se relacionaron, a través de EC IDs que las representan y descripciones génicas, con códigos de acceso génicos y transcripcionales del genoma de duraznero. Se incluyó la presencia de variantes transcripcionales para un mismo gen, no siendo siempre posible asignar el determinante identificado experimentalmente a una de ellas.

Seguidamente, se mencionarán casos posibles durante la asignación realizada. Un fragmento de ADNc secuenciado puede pertenecer a una y solo una de las variantes transcripcionales de un gen y esto detectarse mediante el blastn, o puede que la secuencia identificada sea común a ellas y que la estrategia no permita la asignación específica. Otro caso es aquel en que el péptido identificado

mediante 2D-DIGE-MS/MS pertenezca a un spot del gel 2D con ciertos pI y PM que asistan al blastp para identificar una variante por sobre otra. Asignar específicamente una enzima detectada por Western blot podría realizarse con cierta exactitud si al obtener todos los códigos de acceso génicos y transcripcionales que retienen el EC ID de su actividad catalítica, uno podría diferenciarlos mediante su PM experimental. Respecto a los ensayos de medición de actividades enzimáticas incluidos, la única estrategia realizada fue la de obtener todos los genes de duraznero y sus variantes transcripcionales, que retengan el EC ID correspondiente a la actividad enzimática particular. Una vez incluidos los determinantes moleculares y asignados a códigos de acceso génicos y transcripcionales, la base de datos se completó con descripciones y anotaciones a través de la plataforma Biomart.

Finalmente, se construyeron sets de códigos de acceso transcripcionales con expresión génica en común frente al TT: transcriptos inducidos / reprimidos / afectados, péptidos o proteínas inducidos / reprimidos / afectados, transcriptos y péptidos inducidos / reprimidos / afectados, transcriptos asociados a actividades enzimáticas inducidas / reprimidas por TT. Estos grupos de códigos de acceso construidos se pueden encontrar en la pestaña o solapa final de la base de datos de determinantes.

La tabla suplementaria I (*Tabla SI.xls*) contiene la base de datos construida. Cabe destacarse que, aparte de su uso por parte del presente trabajo, la misma será de gran utilidad para futuros experimentos en nuestra línea de investigación, hacia la caracterización funcional de genes preventivos del daño por frío en frutos de durazno.

### 3.2. PASF db: Plant Abiotic Stress- and Fitohormone-related motifs database

Con el interés de caracterizar en particular la respuesta molecular al tratamiento térmico, decidimos recopilar información acerca de factores de transcripción y sus sitios de unión que puedan relacionarse con el mismo. Un aspecto importante a considerar para el sistema biológico central del presente trabajo, es la falta de información acerca de factores de transcripción y sus sitios de unión. Es por eso que nos remitimos a bases de datos pre-existentes con sitios de unión de factores de transcripción presentes en variadas especies vegetales. Una ventaja de esta estrategia es la conservación de familias de factores de transcripción en plantas, que supone la preservación de sus sitios de unión en las mismas; sin embargo, realizar búsquedas de sitios pertenecientes a otras especies de plantas puede llevar a la pérdida de información valiosa para duraznero debido a que los motivos consenso pueden no contemplar alternativas particulares en esta especie. Cabe destacar que

hasta el momento se han descripto elementos *cis* en promotores y se han estudiado factores de transcripción de duraznero en varias publicaciones, apoyando a la posible acción de mecanismos clásicos para la regulación del inicio transcripcional en estas células vegetales (Wisniewski y col., 2006; Artlip y col., 2013; Pons y col., 2014; Genero y col., 2016).

Una vez revisada la bibliografía, proseguimos a la construcción de la base de datos deseada. Como se describe en Materiales y Métodos (sección 2.3), se utilizaron las bases de datos de motivos preexistentes PlantPAN (que nuclea PLACE, PlantCARE, AGRIS, TRANSFAC y JASPER) y footprintDB, así como referencias bibliográficas adicionales de interés. La selección realizada garantiza (luego de una extensiva revisión para cada motivo) que los sitios de unión de factores de transcripción incluidos sean aquellos probados experimentalmente (*in vivo* e/o *in vitro*). Si bien se respetó lo anterior para la gran mayoría de los motivos, solo unos pocos fueron seleccionados por ser de interés para los conocimientos que nuestro grupo posee hoy en día acerca del tratamiento térmico como estímulo.

La revisión de la/s referencia/s asociadas a cada elemento *cis* no solo permitió comprobar su validez como sitio de unión para factor de transcripción, sino que también hizo posible conocer las características de las familias de factores de transcripción de plantas, su conservación entre diversas especies, los métodos experimentales y bioinformáticos utilizados para el estudio de estas interacciones y su confiabilidad, la alta complejidad y el crosstalk existentes entre las señalizaciones posteriores a cada estímulo que la planta recibe... Esta gran cantidad de información se capitalizó en la misma base de datos de motivos: teniendo en cuenta bajo qué estímulo/s (estrés abiótico o señalización de fitohormona) se reportó que cada par de elementos *cis-trans* actuaría, se construyeron manualmente 18 grupos de motivos lineales, cada uno relacionado a un dado proceso biológico. Los mismos se listan seguidamente:

***Light stimulus***-related
***Cold stimulus***-related
***Heat stimulus***-related
***Oxidative signalling***-related
***Wounding stimulus***-related
***Unfolded Protein Response***-related
***Dehydration stimulus* (*ABA-independent*)**-related

***Biotic stimulus***-related
***Sugar response***-related

***Auxin signalling***-related
***Dehydration stimulus* (*ABA-dependent*)**-related
***Ethylene (ET) signalling***-related
***Jasmonic acid (JA) signalling***-related
***Salycilic acid (SA) signalling***-related
***Cytokinin (CK) signalling***-related
***Gibberellin (GA) signalling***-related

***Circadian rhythm***-related
***Miscellaneous group***

El resultado final fue la base de datos **PASF db**, por **Plant Abiotic Stress- and Fitohormone-related** motifs database. Entre sus características se destacan que: ha sido curada manualmente, posee motivos descriptos en diversas especies vegetales, contiene grupos de motivos lo que permitirá agregar un valor adicional al usuario cuando el mismo realice búsquedas en sets de promotores vegetales, y provee la/s referencia/s bibliográfica/s que demuestran experimentalmente la funcionalidad de los motivos y justifican su inclusión en los diferentes grupos creados.

La tabla suplementaria II (*Tabla SII.xls*) contiene la base de datos PASF db completa y adicionalmente se incluye la lista suplementaria I con las referencias correspondientes a los motivos seleccionados e incluidos (*Lista SI.pdf*). PASF db está abierta al agregado de nuevos motivos que amplíen los grupos de elementos *cis* preexistentes y nuevos grupos de motivos que cumplan roles biológicos bajo un mismo estímulo.

### 3.3. cisAnalyzer: un programa para búsqueda en *cis* y análisis de sobre secuencias biológicas

#### 3.3.1. Descripción de cisAnalyzer

Los objetivos propuestos en este trabajo permitieron diseñar un programa capaz de realizar análisis de aparición, enriquecimiento y co-ocurrencia de motivos en secuencias biológicas.

Uno de los detalles iniciales que se pensó incluir fue la posibilidad de trabajar con secuencias provenientes de cualquier especie biológica, más aún cuando existen muchos genomas secuenciados (como el caso de duraznero) y plataformas que permiten fácilmente su obtención. Otra cualidad que se deseó fue la de sumar información a los resultados clásicos que toda herramienta para análisis de promotores rinde. En particular, se buscó profundizar los análisis posibles a realizarse con la misma información que se recopila al encontrar motivos en una o más secuencias (que incluye la aparición, si la misma sucede en hebra más o menos en caso de acido nucleico y su posición en la secuencia).

La naturaleza de las secuencias en las que el programa realiza la búsqueda de motivos puede ser nucleotídica o aminoacídica. A pesar de los objetivos iniciales, resultó interesante adicionar las posibilidades de análisis de secuencias de ARN y peptídicas. Esto aumenta la versatilidad del programa diseñado, permitiendo una gran cantidad de aplicaciones biológicas.

Finalmente, fue logrado el programa con las características de interés mencionadas e incluso sumando nuevas cualidades que amplían su potencialidad. El nombre **cisAnalyzer** deriva de ellas: eligiendo/insertando motivos o patrones lineales (substrings para Perl) y proveyendo una o más

secuencias (strings para Perl), es capaz de realizar búsquedas (vía pattern matching) en *cis* y de maximizar la información obtenida de esos resultados mediante análisis posteriores de la calidad de las apariciones (o matches), el enriquecimiento y la coocurrencia de los motivos identificados en el grupo de secuencias insertado.

En cuanto al trabajo de programación realizado para la creación de cisAnalyzer, se implementaron principalmente comandos y recursos del lenguaje Perl, aunque también se combinaron con ellos, herramientas de R. El diseño estuvo muy influenciado por el potencial uso del programa por parte de un investigador en biología y, por lo tanto, por las bases de datos de secuencias que él puede obtener fácilmente de las plataformas web de genomas como Phytozome (Biomart). Sin embargo, no se descarta la posible construcción de los archivos input necesarios "a gusto" del usuario.

### 3.3.2. Organización y requisitos de cisAnalyzer

El archivo suplementario comprimido **cisAnalyzer.zip** cuenta con toda la estructura de directorios y los archivos para el funcionamiento correcto del programa cisAnalyzer. Para la instalación, solamente debe descomprimirse el archivo en la ubicación deseada por el usuario. Como requisitos, se solicita poseer sistema operativo Linux, el programa Perl con los módulos *GD::Graph* y *List::MoreUtils* instalados, y el programa R con los paquetes *RColorBrewer* y *gplots* (en el tutorial asociado a cisAnalyzer se encuentran explícitas estas y otras informaciones para permitir al usuario el correcto uso del programa). A modo de ejemplo, se presenta en la figura 3.1, el directorio cisAnalyzer a generarse luego de la descompresión.

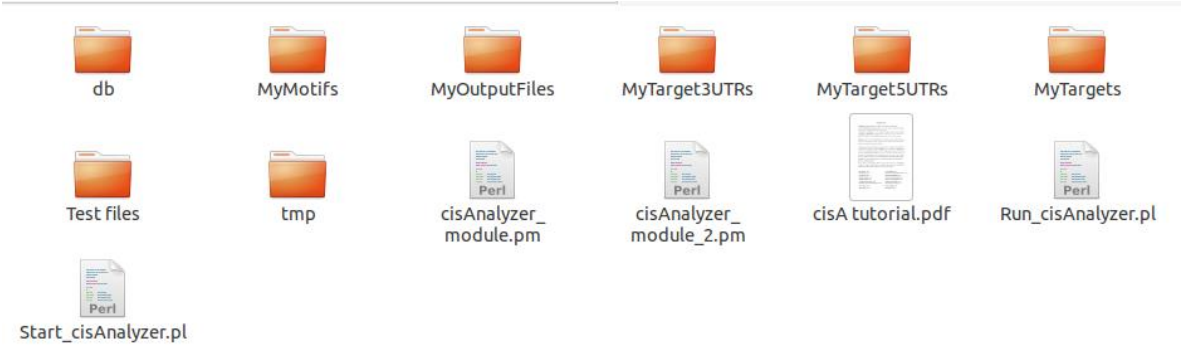


**Figura 3.1. Directorios y archivos incluidos en cisAnalyzer.zip**

Como se observa en la figura, el directorio /cisAnalyzer cuenta con:

dos scripts de Perl (con extensión .pl), denominados Start_cisAnalyzer.pl y Run_cisAnalyzer.pl, que se ejecutan en forma secuencial y que, junto a dos archivos de módulos (extensión .pm), denominados cisAnalyzer_modules.pm y cisAnalyzer_modules2.pm, conforman el core del programa;

el tutorial de uso cisA_tutorial.pdf;

el directorio /Test Files con un archivo para el testeo del correcto funcionamiento;

los directorios /MyMotifs, /MyTargets, /My Target5UTRs y /MyTarget3UTRs permiten ubicar los archivos de motivos y secuencias necesarios como input según el caso; el directorio tmp posee de antemano y es destino transitorio de archivos importantes para la generación de outputs;

el directorio /db que contiene archivos asociados al uso de la base de datos PASF db; y

el directorio /MyOutputFiles contendrá los archivos generados por cada par set de motivos – set de secuencias target, entregado al programa para análisis.

### 3.3.3. Potencialidades de cisAnalyzer

Seguidamente, se describen las diferentes opciones de análisis que el programa ofrece al usuario en consola:

Bajo la opción **CUSTOM MOTIF analysis**, el usuario podrá trabajar con motivos y secuencias target de ADN, ARN o aminoacídicos. Deberá proveer las secuencias target de búsqueda así como el o los motivos que desea buscar y analizar. Esta opción amplía enormemente la cantidad de aplicaciones de cisAnalyzer y cede aún más el control del análisis al usuario, que podría realizar búsquedas y análisis de rutina con sus propios datos. A continuación se mencionan pares motivos - secuencias target a modo de ejemplo: sitios de unión de factores de transcripción - secuencias upstream; sitios de splicing o límites exón-intrón - ARNms; sitios para modificaciones traduccionales, señales de localización subcelular o motivos peptídicos para interacción proteica - proteoma completo de una especie; etc… El alcance de los análisis es dependiente del usuario, de su creatividad e interés.

Cuando se seleccione **PASF db - assisted KNOWN MOTIF analysis**, se podrá caracterizar una o más secuencias upstream (promotor, promotor+5'UTR, 5'UTR, zonas downstream proximales al inicio transcripcional) de plantas, en busca de sitios de unión para factores de transcripción presentes en la base de datos PASF db. Entre las opciones que PASF db provee, se puede elegir buscar con todos los motivos presentes en PASF db, seleccionar grupos de motivos relacionados a estreses abióticos o señalización por fitohormonas, elegir uno o más grupos de elementos *cis* completos, y hasta subseleccionar uno o más motivos de cualquier grupo. El aprovechamiento de la base PASF no reside solamente en los elementos *cis* que contiene (que fueron seleccionados uno por uno y fue revisada la

bibliografía asociada), sino que también implica la posibilidad de relacionar rápidamente a aquellos motivos encontrados en el set de secuencias target, con el o los grupos de estímulos que lo/s contiene/n, facilitando el análisis.

Por último, la elección de **UNKNOWN MOTIF analysis** permite descubrir patrones lineales enriquecidos (presentes en más del 50%) en el grupo de secuencias nucleotídicas insertado. Se registra la aparición de patrones nucelotídicos lineales empezando por todos aquellos (de 5 residuos) posibles de generarse al azar tomando como entrada las 4 bases nitrogenadas A, C, G, T/U. Solo se rendirán aquellos motivos de 5 o más residuos cuyo enriquecimiento es el mayor registrado. Es evidente la participación del usuario para la obtención de resultados de calidad. Solo él controla qué secuencias son insertadas en el programa: su calidad, su longitud y su relación biológica son los factores más importantes para lograr descubrir motivos nuevos en su set de secuencias.

Como consideraciones extra, en los casos en que se analicen promotores, cabe destacarse que es posible obtener información adicional realizando el análisis de secuencias promotoras que incluyan la región 5'UTR. Así mismo, cuando se desee analizar ARNms completos, solo que en este caso deberán incluirse ambos archivos con secuencias 5 y 3'UTR.

Seguidamente, se resumen rápidamente las opciones de análisis en la tabla 3.1.

**Tabla 3.1. Opciones de análisis de cisAnalyzer**

| Opción | Secuencia/s target de búsqueda | | Motivo/s | |
|---|---|---|---|---|
| | **Tipo** | **Origen** | **Tipo** | **Origen** |
| **C) CUSTOM MOTIF analysis** | | | | |
| | ADN ARN peptídica | Usuario | ADN ARN peptídica | Usuario |
| **U) UNKNOWN MOTIF ENRICHMENT analysis** | | | | |
| | ADN ARN | Usuario | ADN ARN | cisAnalyzer detecta motivos enriquecidos |
| **P) PASF db - assisted KNOWN MOTIF analysis** | | | | |
| | ADN<br>Ideado para promotores vegetales | Usuario | ADN<br>Sitios de unión de factores de transcripción | PASF db |

### 3.3.4. Funcionamiento general de cisAnalyzer

Una vez instalado, cisAnalyzer puede ser manejado como cualquier script de Perl. A continuación se detalla un protocolo de uso general:

**1)** ***Navegar a la carpeta cisAnalyzer desde consola***

**2)** ***Ejecutar el script*** **Start_cisAnalyzer.pl**

**3)** ***Completar datos requeridos y elegir opciones de análisis de interés***

**4)** ***Insertar el archivo con las secuencias target de búsqueda en*** **/MyTargets**

**5)** ***Ejecutar el script*** **Run_cisAnalyzer.pl**

**6)** ***Verificar los archivos output generados en*** **/MyOutputFiles**

El programa guía al usuario para que realice el paso a paso requerido, durante la misma ejecución. Las decisiones se realizan a través de inputs de caracteres (letras o números) por parte del usuario; los mismos son identificadores presentados en consola o, para un único caso, en archivos output. Debido a los distintos caminos a los que puede llevar y para agregar consideraciones de interés, seguidamente, se detalla cada paso del protocolo de uso.

**1)** ***Navegar a la carpeta cisAnalyzer desde consola***

**2)** ***Ejecutar el script*** **Start_cisAnalyzer.pl**

```
…/cisAnalyzer$ perl Start_cisAnalyzer.pl
```

Este primer script contiene código en lenguaje Perl que recibe las opciones de análisis que el usuario desea, permite que el mismo inserte el archivo de secuencias target de búsqueda (entre otros dependiendo del análisis elegido), y realiza el procesamiento y control de calidad de las secuencias a utilizar.

**3)** ***Completar datos requeridos y elegir opciones de análisis de interés***

Se pedirá por única vez que el usuario tipee el path hacia la carpeta cisAnalyzer. Esta acción genera un archivo adicional en el directorio /cisAnalyzer que servirá en el futuro para evitar este pedido. Por ello es necesario evitar la eliminación del mismo. Por otro lado y en todo análisis, se pedirá además el tipeo de un nombre de usuario. Recomendamos que el mismo identifique el análisis.

En caso de CUSTOM MOTIF ANALYSIS, se requiere un archivo de extensión .txt, con el/los motivo/s (nucleotídicos o aminoacídicos) a ser buscados en las secuencias targets, ubicado en la carpeta MyMotifs vacía. Los elementos del mismo deben estar organizados como sigue (ejemplo de ADN):

> TTTTAATTT
> ATAATT

Adicionalmente, los motivos solamente pueden contar con caracteres IUPAC, siendo solo éste el medio vía el cual se introducen consensos (motivos con variantes) para buscar (ver tabla 3.2).

**Tabla 3.2. Caracteres nucleotídicos y aminoacídicos IUPAC**

| Caracter/es | | Denominación | Caracter/es | Abreviatura | Denominación |
|---|---|---|---|---|---|
| **A** | | Adenina | **G** | Gly | Glicina |
| **C** | | Citosina | **A** | Ala | Alanina |
| **G** | | Guanina | **V** | Val | Valina |
| **T** | | Timina | **L** | Leu | Leucina |
| **U*** | | Uracilo | **I** | Ile | Isoleucina |
| **Y** | C o T | pirimidina (pYrimidine) | **P** | Pro | Prolina |
| **R** | A o G | puRina | **F** | Phe | Fenilalanina |
| **S** | C o G | unión fuerte (Strong) | **Y** | Tyr | Tirosina |
| **W** | A o T | unión débil (Weak) | **C** | Cys | Cisteína |
| **M** | A o C | grupo Amino | **M** | Met | Metionina |
| **K** | G o T | grupo ceto (Keto) | **H** | His | Histidina |
| **V** | A, C o G | no T | **K** | Lys | Lisina |
| **B** | C, G o T | no A | **R** | Arg | Arginina |
| **D** | A, G o T | no C | **W** | Trp | Triptófano |
| **H** | A, C o G | no G | **S** | Ser | Serina |
| **N** | A, C, G o T | Cualquiera | **T** | Thr | Treonina |
| | | | **D** | Asp | Aspartato |
| | | | **E** | Glu | Glutamato |
| | | | **N** | Asn | Asparagina |
| | | | **Q** | Gln | Glutamina |
| | | | **X** | - | Cualquiera |

*Uracilo reemplaza a Timina en secuencias ribonucleicas por tanto se aplica lo mismo que para la segunda en cuanto a las variantes

cisAnalyzer reportará en /MyOutputFiles motivos con residuos indebidos (*cisA_user_date_Removed motifs.tsv*) y aquellos válidos (*cisA_user_date_Input motifs.tsv*).

### 4) *Insertar el archivo con las secuencias target de búsqueda en* /MyTargets

Previo verificar que el directorio /MyTargets esté vacío, para todo análisis como mínimo debe colocarse un archivo multifasta de extesión .txt con la/s secuencias target de búsqueda. A modo de ejemplo de ADN, la estructura del header de cada secuencia deberá ser la siguiente:

> GeneName | TranscriptName | Organism name
> TTTTAATTTATAATTCTATTGGTTTCACGATTTGGTTTGGGTGCCGAGG

Las secuencias deben poseer caracteres según la naturaleza declarada de las mismas, sino serán removidas (se reportará en /MyOutputFiles un archivo de secuencias target removidas -cisA_user_date_Removed seqs.txt- y otro con aquellas validos -cisA_user_date_Input seqs.txt-). Eso refleja la importancia del control previo (aparte de cisAnalyzer) que el usuario puede realizar sobre los datos de secuencia que inserta en el programa.
En caso de desear incluir el análisis de promotor+5UTR o ARNm completo (con 5' y 3'UTR), luego de seleccionar estas opciones, cisAnalyzer pedirá adicionalmente que se coloquen los archivos equivalentes (en cuanto a headers fasta) al archivo de targets (es decir que, idealmente, para toda secuencia target debería haber una secuencia 5'UTR, por ejemplo).
El/los archivos requeridos (de targets, de 5'UTR y de 3'UTR) en este particular formato fasta son fácilmente obtenidos mediante el control de filtros y atributos que provee Biomart, disponible en muchos browsers de genomas secuenciados. La figura 3.2 muestra esta herramienta disponible en Phytozome para genomas vegetales (http://www.phytozome.net/; Goodstein y col., 2011).

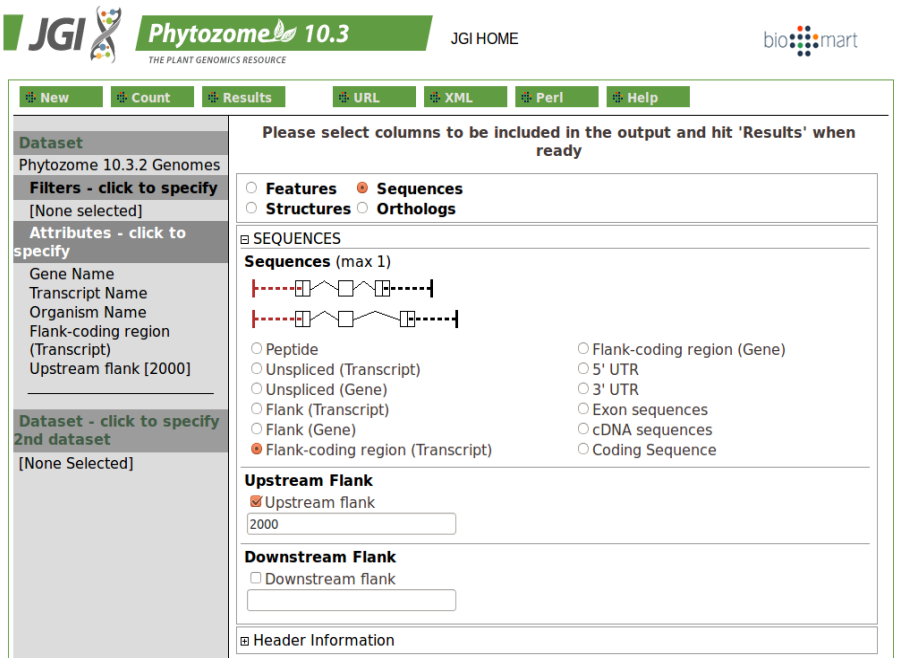


**Figura 3.2. Plataforma Biomart en Phytozome.** Permite diseñar y crear bases de datos de secuencias aplicando filtros a los datos asociados a sus genomas y seleccionando atributos deseados. Luego, se descarga directamente la base creada.

Otra consideración es que cisAnalyzer, pensado inicialmente para análisis de promotores vegetales, permite el análisis de secuencias biológicas provenientes de cualquier organismo. Por cuestiones del programa y para facilitar el camino que recorre el usuario en su utilización, se incluyeron en él las denominaciones que Phytozome utiliza para los genomas (específica de la especie) que contiene. Es por ello que se visualizarán los siguientes cuestionamientos en consola:

*> Do all your target sequences belong to one or more species included in Phytozome? (Y/N):*

Si el set de secuencias targets proviene de una especie no presente en Phytozome (podría ser una especie no vegetal como *H. sapiens*), se sugiere contestar con N (No). Luego:

*> please answer, how many species not included in Phytozome are within your set of sequences?*

Se deberá indicar el número de especies no listadas en Phytozome y finalmente se deberá ingresar su denominación como se presenta en el header fasta de la/s secuencia/s target en cuestión.
De esta manera, se procesará correctamente el archivo. Algún error en el tipeo o no declaración de las especies del set de secuencias target causará que cisAnalyzer remueva las secuencias cuyo header contiene una especie desconocida para él.

Cabe destacarse que el uso de información proveniente de cualquier genoma secuenciado está sujeto a su publicación. Recae en el usuario de cisAnalyzer la responsabilidad de verificar la categorización del/de los mismo/s y la accesibilidad de aquellas secuencias a utilizar debido a que algunos genomas no han sido publicados en su versión definitiva y existen restricciones en su uso. La tabla suplementaria III (*Tabla SIII.xls*) permite ver, entre otros detalles, disponibilidad y accesibilidad de genomas de Phytozome, verificadas para diciembre de 2015.

***Fin del programa Start_cisAnalyzer.pl.***

Se han generado archivos necesarios para el análisis en sí que seguidamente se realizará.

### 5) *Ejecutar el script* Run_cisAnalyzer.pl

```
…/cisAnalyzer$ perl Run_cisAnalyzer.pl
```

Con los archivos procesados de targets (motivos y UTRs cuando sea el caso) y las decisiones del usuario, este segundo programa utiliza los módulos creados con las subrutinas exclusivas del pattern matching en sí (que realiza las búsquedas) y de los análisis posteriores que hacen a los resultados y a su presentación como outputs. No hay nuevas decisiones que se presenten en consola, solo se visualizan aquellas tomadas durante la ejecución de Start_cisAnalyzer.pl y, luego del tiempo que requiera el script, se informa si se obtuvieron resultados positivos o no en la búsqueda realizada.

El pattern matching incluido se realiza para cada motivo y, de manera *global*, para cada secuencia target del set. Es decir, cada motivo será la *expresión regular* (regex) a buscar sobre cada secuencia (una por vez) y cuando la búsqueda de un motivo culmine para todas las secuencias target, se

proseguirá con el siguiente motivo. La estructura del código que permite aplicar este tipo de pattern matching es la siguiente:

```
while ($target[3] =~ m/$DNAmotif_with_variants/g){
    guardado y procesado de la información para cada motivo-target
};
```

En el mismo, se presentan variados elementos:

`while ( condicional ) { acción a ejecutar }` es el loop que testea el condicional impuesto y luego itera hasta que deje de ser cierto;

`$target[3]` es el escalar tipo string que contiene la secuencia target de búsqueda;

`$DNAmotif_with_variants` es el escalar tipo string que describe un patrón lineal, es la expresión regular (regex) a buscar;

`m//` es el operador de match para una regex;

`//g` es un modificador global que permite el match de un regex en un string tantas veces como sea posible;

el operador `=~` especifica un string para el operador de match, los asocia;

Una aclaración importante es que el método utilizado registra la presencia de los motivos buscados, incluso si ellos admiten variantes. Cuando el motivo no presenta variantes, un match implica encontrarlo tal y como es (ej: motivo TTACG, posible match sobre secuencia = …TTACG…). Por otro lado, si el motivo es un *consenso,* será match encontrar cualquiera de sus variantes tal y como es cada una, y se sumará el registro de todas ellas para el total (ejemplo: motivo TTACGTY o TTACGT[T/C], posible match 1 sobre secuencia = …TTACGTC…, posible match 2 sobre secuencia = … TTACGTT…).

Luego del pattern matching descripto, cisAnalyzer trabaja con esa información para obtener otro tipo de resultados deseados. Mediante el manejo de diferentes listas que retienen la información anterior, **cisAnalyzer clasifica los matches de cada motivo según los mismos se presenten: single o múltiple** (presentes una o más de una vez en las secuencias target), **en hebra más o menos** (plus o minus) y, si es posible, **en diferentes zonas de las secuencias targets.**

En cuanto a la información de posición, el método utilizado consiste en dividir cada una de las secuencias en 2 ó 4 zonas, cuando sea posible. Por ello, fue necesario tener en cuenta que el largo de las secuencias targets del set puede o no ser el mismo y sus implicancias en los resultados, y que el largo de los motivos a buscar (verificado en *Start_cisAnalyzer.pl* de ser mayor a la secuencia target menor) compromete la división a realizar. En base a esto se realiza la división deseada cuando es posible y se reportan los matches en cada zona.

Una consideración extra es acerca de la inclusión de la información de posición debida a UTRs. Los resultados para estas regiones están sujetos a aquellas secuencias que poseen UTRs anotados.

Como se trabaja con un set secuencias targets (con más de una secuencia), **los resultados obtenidos para cada motivo se presentan como frecuencias relativas y las relativizaciones son respecto a dos totales en general: total de matches y total de secuencias target con al menos un match.** Este último "tipo" de totales es el de interés cuando se pretende analizar enriquecimiento de motivos en un set de secuencias.

Otro detalle que debió considerarse es la cantidad de motivos, que permite análisis de **co-ocurrencia** o no (presencia compartida de dos o más motivos) y compromete la calidad de los archivos output. Análisis con un motivo genera gráficos de barras para sus frecuencias relativas en las diferentes categorías; analizar con 2 a 4 motivos presentarán además los resultados de coocurrencia en todas las combinaciones posibles; estudios con 4 a 40 motivos no reportarán coocurrencia; trabajar con más de 40 motivos obliga a que se presenten (gráficamente) solo aquellos con mayores frecuencias, por cuestiones de visualización del output.

### 6) *Verificar los archivos output generados en* /MyOutputFiles

En el caso en que se hayan obtenido resultados positivos, cisAnalyzer (mediante sus dos programas Start y Run) generará los siguientes archivos de output en la carpeta MyOutputFiles:

Compendio de las decisiones tomadas por el usuario

*cisA_user_date_Made decisions.txt*

Se reportan las secuencias target utilizadas y las removidas para permitir el análisis

*cisA_user_date_Input seqs.txt*

*cisA_user_date_Removed seqs.txt*

Para CUSTOM MOTIF analysis, se reportan los motivos utilizados y removidos para permitir el análisis

*cisA_user_date_Input motifs.xls*

*cisA_user_date_Removed motifs.xls*

En caso de PASF db-assisted analysis, se reportan los grupos de motivos primeramente seleccionados con sus identificadores, y la información de los motivos finalmente elegidos a través de consola para realizar el análisis

*cisA_user_date_Selected PASFdb group(s).txt*

*cisA_user_date_Selected PASFdb motif(s).xls*

En caso de Unknown motif analysis, se reporta el resultado del análisis de motivos desconocidos enriquecidos en el set de secuencias target incluido

*cisA_user_date_Motif discovery.txt*

Tabla clásica con los resultados del pattern matching de cada motivo en cada secuencia target

*cisA_user_date_Table.xls*

Reporte con la información del pattern matching analizada en profundidad para permitir la caracterización de cada motivo en el set de secuencias target. Se presentan frecuencias relativas de

cada motivo y de grupos (en caso de seleccionar más de uno de ellos de PASF db), resultados para cada categoría, coocurrencia, esquemas de localización. Lo más importante de este reporte es la **posibilidad de obtener la identidad de las secuencias target que hacen a cada frecuencia relativa**, información de gran utilidad para el biólogo.

*cisA_user_date_Report.xls*

Reporte con la información de enriquecimiento de los motivos elegidos o incluidos en el análisis. Se presentan rápidamente los motivos presentes, al menos una vez, en más del 50% del total de secuencias target incluidas para analizar.

*cisA_user_date_Enrichment Report.xls*

Gráficos asociados al reporte antes mencionado. Representan una manera rápida de visualizar los resultados centrados en cada motivo, independientemente de la identidad de las secuencias target que hacen a las frecuencias relativas. Son dos grafícos de barras para el análisis con un único motivo (A y C) y se generan en Perl mediante su módulo GD::Graph. Por otro lado, son cinco heatmaps para cuando se encuentra más de un motivo (A-E) y ellos se generan a partir de scripts de R que se ejecutan desde los módulos de Perl creados para cisAnalyzer. A y B muestran frecuencias relativas de matches de cada motivo vs. total de matches de todos los motivos (A) o total de matches de cada motivo (B). C, D y E representan frecuencias relativas de secuencias target con al menos una aparición del dado motivo vs. total de secuencias target insertadas (C), total de secuencias target con al menos un match de cualquier motivo analizado (D) o total de secuencias target con al menos un match del mismo motivo (E). Los diferentes gráficos muestran distintos resultados aunque, para visualizar el enriquecimiento de un motivo en el set de secuencias, el gráfico C sería el indicado. Un detalle presente en estos cinco gráficos por análisis es que a partir de los datos de frecuencia obtenidos, se añade un clustering jerárquico (distancia obtenida por método Manhattan y método de clustering Ward; aunque pueden cambiarse en los scripts) que agrupa motivos según su *calidad* de matching en el set de secuencias target en cuestión.

*cisA_user_date_Graph A.png*
*cisA_user_date_Graph B.png*
*cisA_user_date_Graph C.png*
*cisA_user_date_Graph D.png*
*cisA_user_date_Graph E.png*
*cisA_user_date_Groups Graph.png* (si los motivos provienen de diferentes grupos de PASF db)

Leyendas de los gráficos obtenidos

*cisA_user_date_Legends.txt*
*cisA_user_date_Legend2.txt* (sólo si los motivos pertenecen a diferentes grupos de PASF)

### 3.3.5. Aplicaciones y perspectivas a futuro de cisAnalizer

Las características logradas en cisAnalyzer resultan interesantes para diversas problemáticas que el investigador en biología enfrenta seguidamente. Las posibilidades de cisAnalyzer dependen principalmente de las necesidades y la creatividad del usuario. Entre otros, creemos que vale la pena destacar dos usos extras que cisAnalyzer permite, ya que no se reportan resultados similares a ellos en este trabajo final.

Uno de ellos es el análisis de conservación de motivos en promotores de ortólogos proteicos. Obteniendo los ortólogos de un gen de interés y luego sus promotores, se podrá verificar la conservación, incluso posicional, de los motivos encontrados en el promotor de interés. Esta información confiere más validez al primer análisis, filtrándose los resultados obtenidos para quedarse con los motivos más importantes que harían a la funcionalidad común de todos los ortólogos. Y cisAnalyzer permite obtenerla en muy poco tiempo.

La segunda aplicación de cisAnalyzer a destacar es la posibilidad de realizar análisis genome-wide sobre genomas de interés para verificar *in silico* la validez de ciertos futuros experimentos y predecir resultados. Un ejemplo sería el de poseer los motivos peptídicos mediante los que una proteína de interés se haya demostrado interaccionante con otra proteína. Si uno deseara predecir otros factores interaccionantes, lo mínimo que podría buscar inicialmente es qué otros polipéptidos (putativamente presentes en el proteoma completo del organismo de interés) poseen esos motivos peptídicos para elegir nuevos polipéptidos a probar, etc… Para estos casos, se hace evidente que el se verá comprometido el tiempo computacional cuando se provea al programa una gran cantidad de motivos y de secuencias target (que además pueden poseer importantes longitudes) para analizar. El tiempo requerido será dependiente del análisis solicitado y de la capacidad de hardware disponible para el programa.

El archivo comprimido *cisAnalyzer.zip* se encuentra disponible con la publicación del presente trabajo. El programa está abierto a modificaciones en pos de permitir nuevos usos por parte del usuario.

### 3.4. Caracterización de secuencias upstream de genes cuya expresión responde al tratamiento térmico: una aplicación del programa cisAnalyzer

Como se ha mencionado en la introducción, el tratamiento térmico es una de las estrategias preventivas del daño por frío en más de una especie frutícola. En nuestro laboratorio, se trabajó con duraznos de la variedad Dixiland y se demostró que el almacenamiento de la fruta cosechada por 3 días a 39 °C era exitoso en la protección de las células del fruto cuando se exponían a bajas temperaturas e incluso tenía efectos benefactores. Es por ello que se aplicaron diversas técnicas ómicas sobre estas muestras, identificándose genes con funcionalidad posiblemente protectora frente al frío. Para ahondar aún más en las bases genéticas y moleculares de la acción preventiva mencionada, entre otros enfoques, estamos trabajando sobre la regulación del inicio transcripcional de los genes identificados y un aspecto importante de la misma se cubre mediante el estudio de sus regiones promotoras.

La construcción de la base de datos de determinantes realizada en el presente trabajo (Tabla suplementatia SI), fue necesaria para obtener las secuencias promotoras y 5'UTRs de cada set. Luego, se prosiguió con el análisis en cisAnalyzer, utilizando los grupos de motivos relacionados con estímulo calor y frío por separado (ver sección 2.4 de Materiales y Métodos), y los archivos multifasta, correspondientes a los sets de determinantes moleculares construidos:

**DET-I** transcriptos inducidos por tratamiento térmico

**DET-R** transcriptos reprimidos por tratamiento térmico

**DEP-I** péptidos inducidos por tratamiento térmico

**DEP-R** péptidos reprimidos por tratamiento térmico

**DET-A** transcriptos afectados por tratamiento térmico: inducidos o reprimidos

**DEP-A** péptidos afectados por tratamiento térmico

**DETyP-I** transcriptos y péptidos inducidos por tratamiento térmico

A modo de ejemplo descriptivo de la aplicación de cisAnalyzer para el problema, se presentan los gráficos de resultados para el set DET-I, con los motivos del grupo heat-related (relacionados a estímulo por altas temperaturas), en las figuras 3.3 a 3.7. Los diferentes colores elegidos, ayudan al usuario a identificar los gráficos una vez que realice más de un análisis.

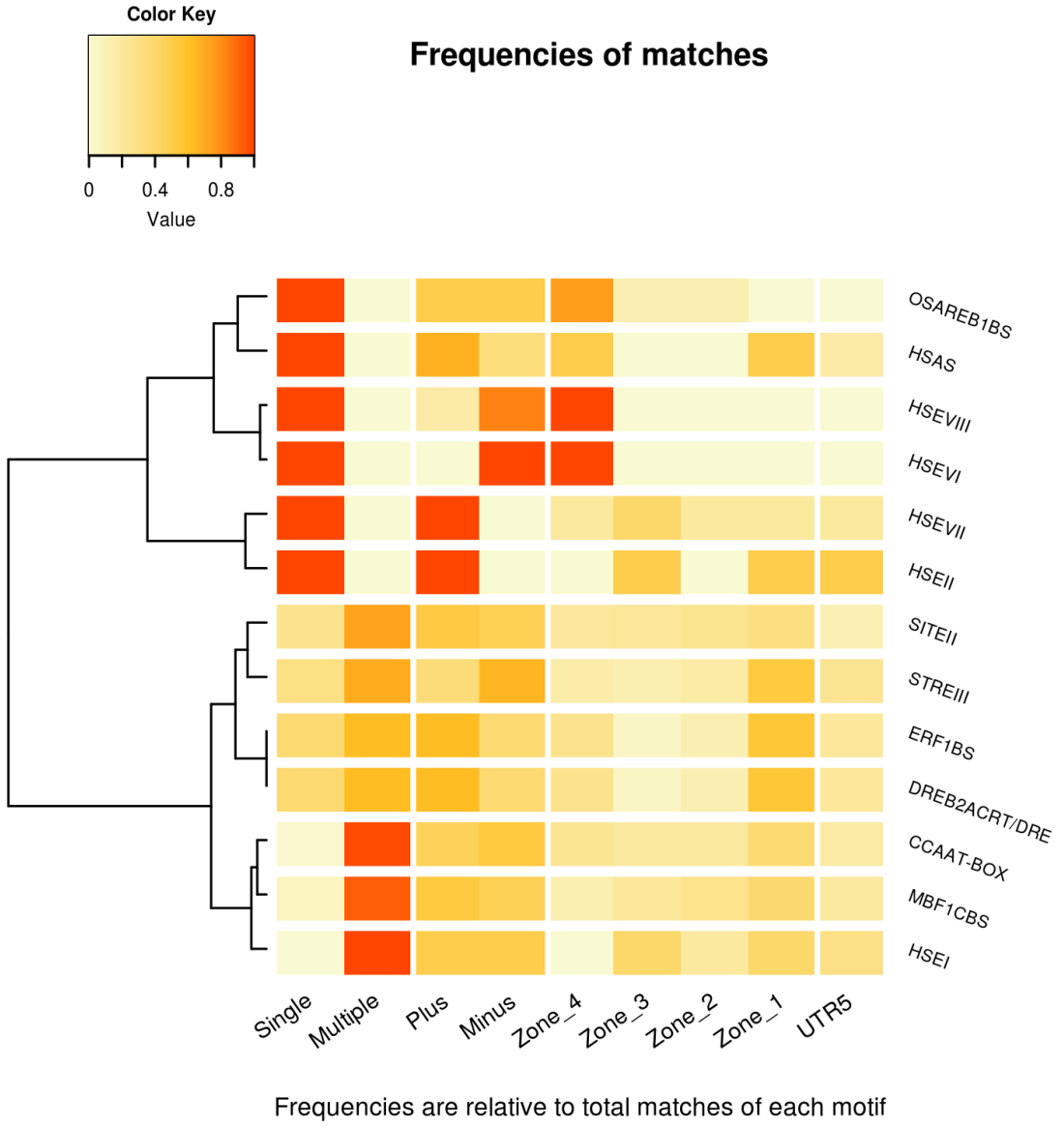


**Figura 3.3. Gráfico A, Set DET-I, Motivos Heat-related.** Se grafican frecuencias relativas de matches (apariciones) de cada motivo, versus el total de matches particular del mismo motivo.

El gráfico es un heatmap de frecuencias relativas de cada motivo que se encontró al menos una vez en el set de secuencias DET-I (transcriptos inducidos por tratamiento térmico). Se pueden visualizar las frecuencias relativas de match, por ejemplo del motivo HSEII, en las categorías analizadas, versus el total de matches del mismo HSEII en el set DET-I. El elemento *cis* HSEII posee (respecto a todos sus matches) alta cantidad de matches single (uno por secuencia) y en hebra +, y que se localiza en zonas 1 y 3, y en 5'UTR. Esto causa su clustering junto a HSEVII cuya *calidad* de matching en el set DET-I es similar frente a la de otros motivos.

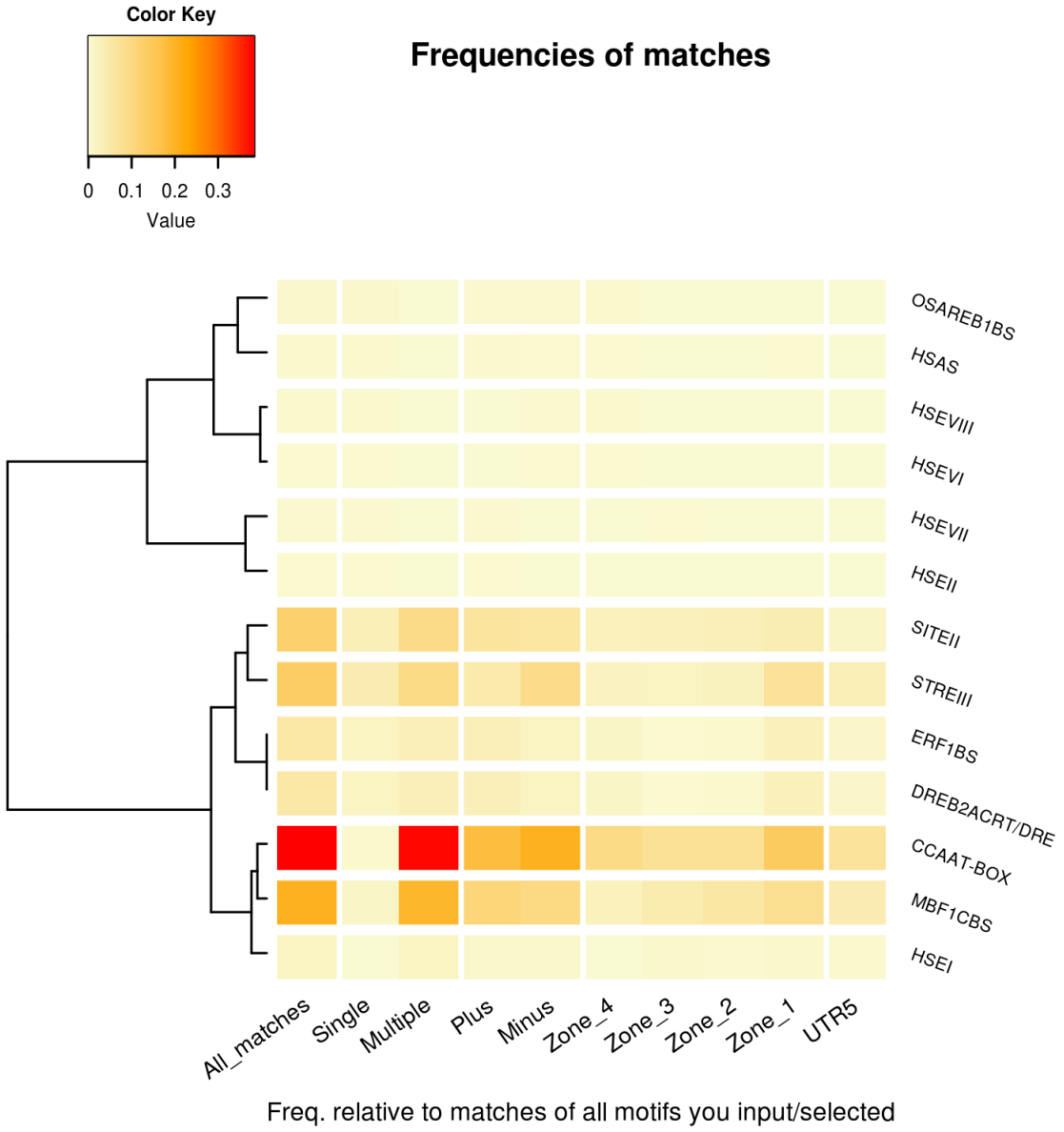


**Figura 3.4. Gráfico B, Set DET-I, Motivos Heat-related.** Se grafican frecuencias relativas de matches (apariciones) de cada motivo, versus el total de matches de todos los motivos heat-related. Aparece All_matches con las frecuencias de matches de cada motivo, relativas al total de matches de todos los motivos encontrados.

El gráfico B, como el A, también presenta frecuencias relativas de matches de cada motivo, sólo que relativizadas al total de matches por parte de todos los motivos encontrados del grupo heat-related, en el set de secuencias DET-I. CCAAT-box, MBF1c bs, STREIII y SiteII, son los motivos con mayor cantidad de matches aunque no están enriquecidos en DET-I ya que los matches podrían concentrarse en la minoría de secuencias. Las altas frecuencias de match de estos 4 motivos en el total de matches que ocurrió, desdibuja las frecuencias de otros motivos como HSEII.

Debido a cuestiones recientemente mencionadas, se vuelven outputs importantes los gráficos C, D y E que siguen, ya que las frecuencias relativas que se grafican en ellos son otras en concepto.

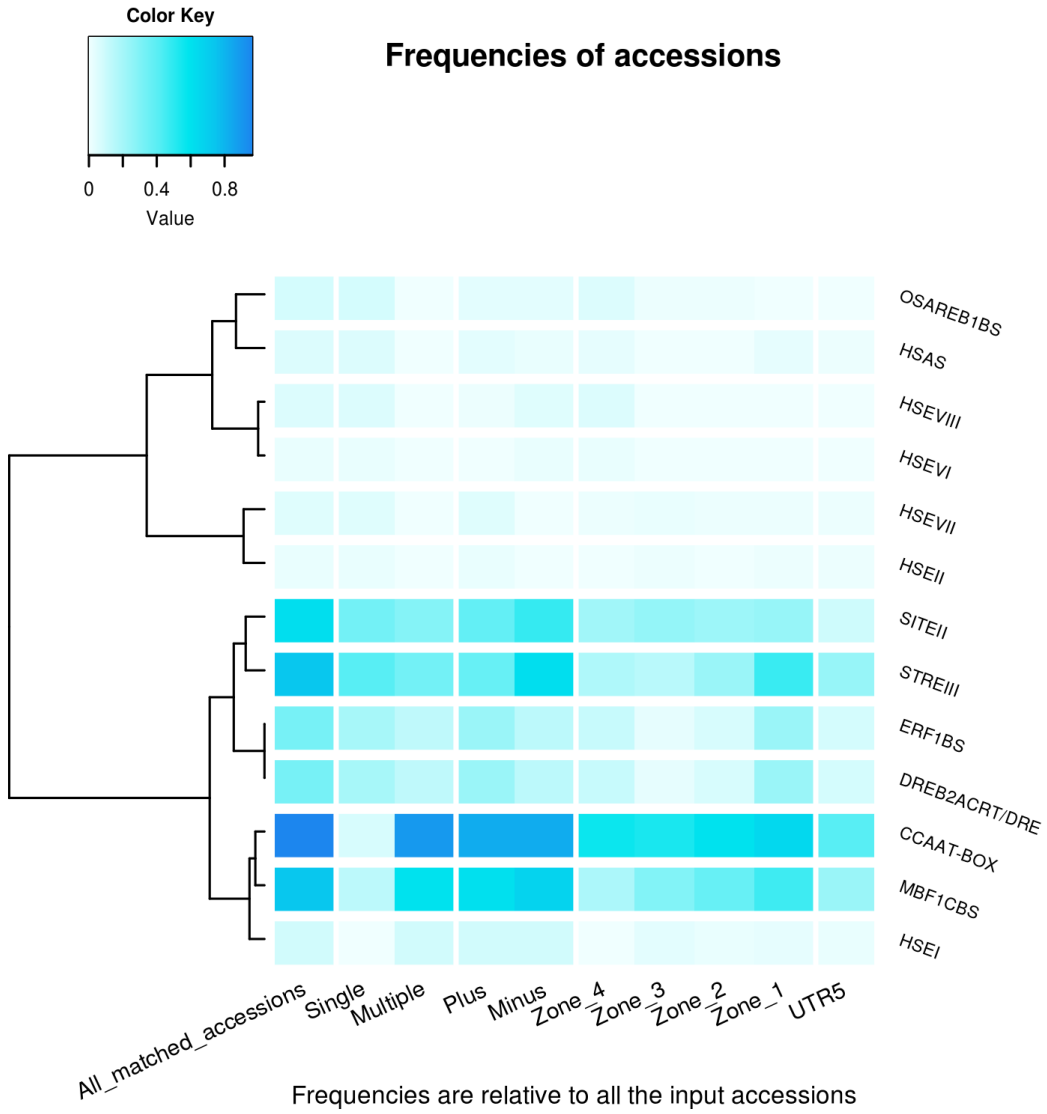


**Figura 3.5. Gráfico C, Set DET-I, Motivos Heat-related.** Se grafican frecuencias relativas de secuencias (accessions) con al menos un match por un dado motivo, versus el total secuencias del set DET-I. Aparece All_matched_accessions para mostrar las frecuencias relativas de secuencias target con al menos un match de cada motivo, respecto al total insertado.

Este gráfico muestra el enriquecimiento de motivos heat-realted en DET-I. Las frecuencias relativas son de secuencias target con al menos un match vs. todas las secuencias de DET-I (incluidas aquellas sin matches). CCAAT-box y MBF1c bs son los más enriquecidos y, junto con STRE III y Site II, son los 4 motivos presentes en más del 50% de las secuencias de DET-I. ERF1 bs y DREB2A CRT/DRE, pero no los HSE (clásicos en estrés por calor), cobran importancia aunque en menos del 50%.

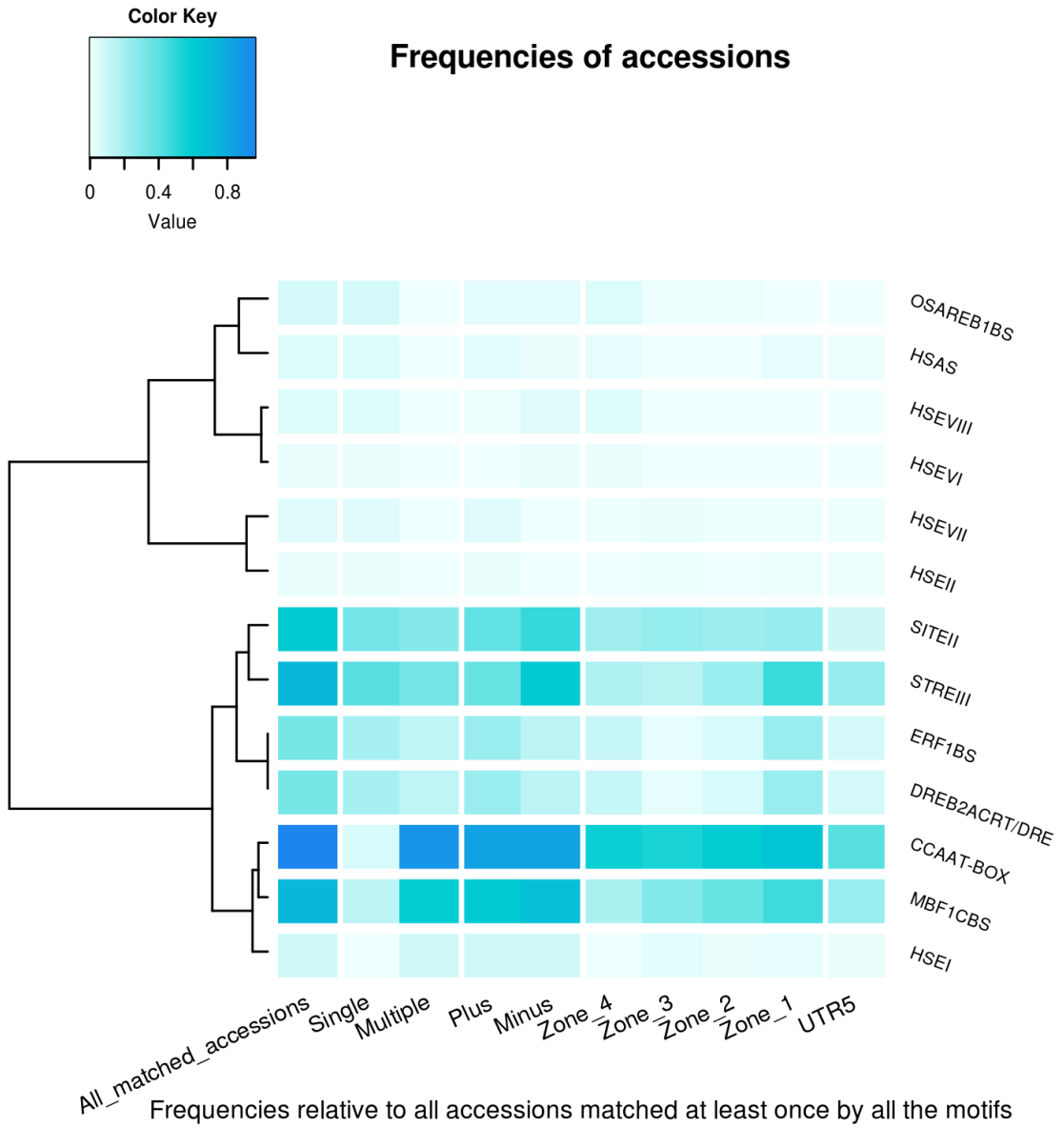


**Figura 3.6. Gráfico D, Set DET-I, Motivos Heat-related.** Se grafican frecuencias relativas de secuencias con al menos un match por un dado motivo, versus el total secuencias con al menos un match de cualquiera de los motivos heat-related.

El gráfico D es similar al C pero el total al que las frecuencias absolutas son relativizadas, es ahora la cantidad de secuencias con al menos un match de cualquiera de los motivos elegidos. En otras palabras, se relativiza respecto al total de secuencias con al menos un match por parte de los motivos heat-related.

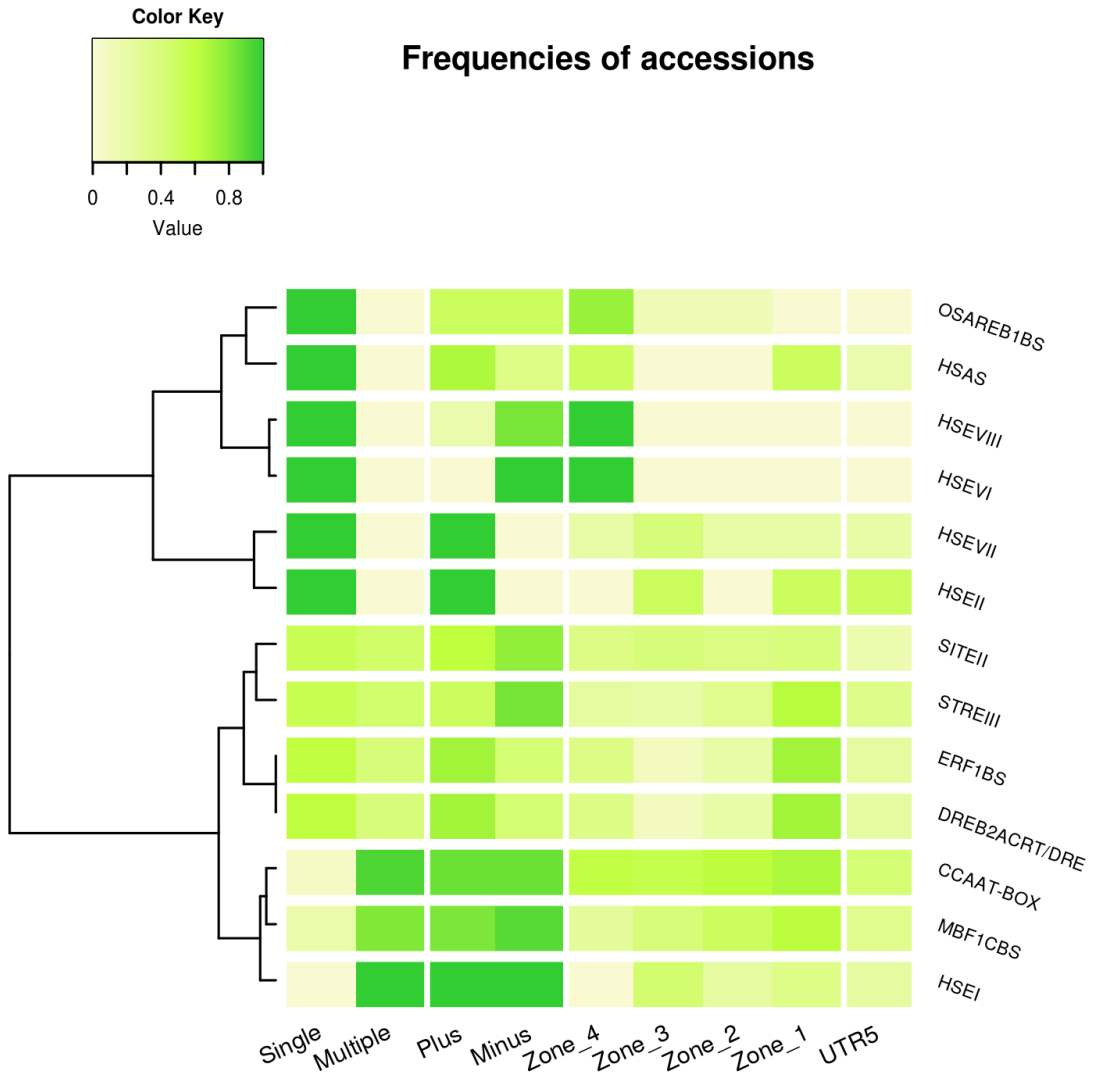


**Figura 3.7. Gráfico E, Set DET-I, Motivos Heat-related.** Se grafican frecuencias relativas de secuencias con al menos un match de un dado motivo, versus el total de secuencias con al menos un match del mismo motivo.

Este último gráfico output de cisAnalyzer para este par de motivos–secuencias target, deja ver las frecuencias relativas de secuencias con al menos un match de cada motivo, respecto al total de secuencias que poseen match del mismo motivo. Se ve claramente que todas las secuencias con single match por parte de HSEVI y VIII lo poseen localizado en la zona más distal. Esa ubicación común confiere importancia biológica a esos motivos solo en esas secuencias, aunque las mismas no representen (como se observa en el grafico C) más del 50% dentro del set DET-I (promotores de transcriptos inducidos por tratamiento térmico).

El análisis siguiente consistió en comparar los resultados obtenidos para DET-I con su control VsDET-I y con los resultados para el set de promotores total del genoma de duraznero. VsDET-I es un set de secuencias target, equivalente a DET-I en cantidad de secuencias y largo de las mismas, aunque construido al azar. Estas comparaciones permiten validar que el enriquecimiento observado con el set de interés posee mayores posibilidades de deberse a la conexión biológica que poseen las secuencias en DET-I. Los motivos CCAAT-box, MBF1c bs y STREIII poseen "altas probabilidades de ser encontrados" en más de un set de secuencias de promotores de duraznero, incluido el set de promotores del genoma completo, haciendo desconsiderable su presencia. Por otro lado, observando los gráficos E para DET-I y su control, se observa que la localización común de HSEVI y VIII es particular de ése su subset de secuencias ya que no se observa lo mismo para el control. A partir del reporte completo que se puede obtener la frecuencia relativa visible en el gráfico C). La tabla 3.3 resume estos resultados discutidos.

**Tabla 3.3. Resultados obtenidos con el set de DET-I y los motivos Heat-related.**

| Motivo | | Enriquecimiento en DET-I | Enriquecimiento en set control VsDET-I | Enriquecimiento en genoma |
|---|---|---|---|---|
| **HSEVI** | * | 0.02 | 0.06 | 0.03 |
| **HSEVIII** | ** | 0.06 | No presente | 0.01 |
| **ERF1 bs** | RCCGAC | 0.32 | 0.24 | 0.29 |
| **DREB2A CRT/DRE** | | | | |
| **HSAS** | GGGTGTC | 0.06 | 0.04 | 0.06 |
| **HSEVII** | *** | 0.05 | 0.08 | 0.08 |
| **OsAREB1 bs** | ACGTGCC | 0.07 | 0.05 | 0.03 |
| **HSEII** | **** | 0.02 | 0.01 | 0.03 |
| **HSEI** | $AGAAN_2TTCT$ | 0.08 | 0.10 | 0.05 |
| **STREIII** | AGGGG | 0.74 | 0.76 | 0.78 |
| **CCAAT-box** | CCAAT | 0.96 | 0.97 | 0.98 |
| **MBF1c bs** | CTAGA | 0.74 | 0.81 | 0.80 |
| **SITEII** | TGGGCY | 0.62 | 0.52 | 0.64 |

* $TTCN_4GAAN_7GAA$ ** $TTCN_2TTCN_{10}TTCN_3TC$ *** $TTCN_3TTCN_8TTC$ **** $GAAN_2TTCN_2GAA$

32% de DET-I posee el motivo RCCGAC al menos una vez, compartido por los factores AP2 DREB2A y ERF1 (Sakuma y col., 2006; Cheng y col., 2013). Sin embargo, este enriquecimiento es levemente mayor a aquel observable en el set control y menor al que posee el genoma.

El mismo procedimiento se llevó a cabo para analizar los motivos que conforman el grupo relacionado a estímulo por bajas temperaturas. La figura 3.8 deja visualizar directamente el gráfico C.

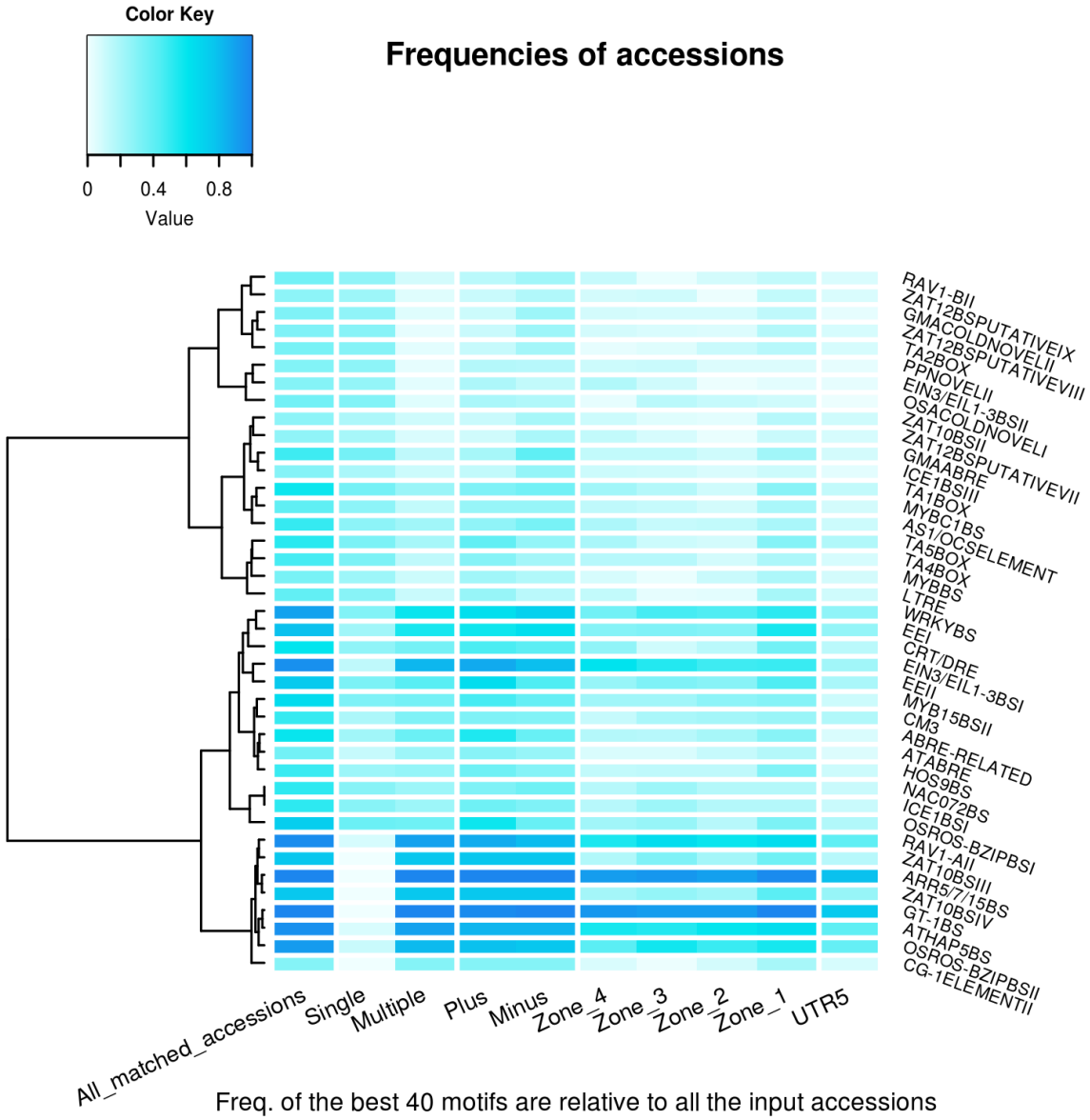


**Figura 3.8. Gráfico C, Set DET-I, Motivos Cold-related.** Se grafican frecuencias relativas de secuencias con al menos un match por un dado motivo, versus el total secuencias del set DET-I.

El heatmap C y el reporte de enriquecimiento permiten rápidamente observar que los motivos CRT/DRE (RYCGAC), OsROS-bZIP bs I (TTGATC), y ABRE-RELATED (MACGYGB) se presentan en más del 50% de las secuencias de DET-I, y en procentajes que superan al set control y a su ocurrencia en el genoma. Esto les da importancia como actores en la regulación de estos promotores bajo el tratamiento térmico. No se discuten motivos adicionales que se encuentran enriquecidos y otros, con

porcentajes menores al 50, cuyos subsets podrían indagarse particularmente (a partir del reporte completo).

Los pasos realizados y descriptos anteriormente para el set de secuencias target DET-I se llevaron a cabo y luego se evaluaron los resultados, para todos los sets de secuencias de interés (y sus controles), con ambos grupos de motivos de PASF db, heat- y cold-related. El archivo comprimido *cisA_Ppe_FastasAndResults.zip* posee los archivos multifasta descargados para cada set y los resultados obtenidos en su totalidad (*cisA_Ppe_Analysis_Names.xls* detalla los identificadores de cada análisis, ej: DET-I Cold es el análisis con motivos Cold stimulus-related sobre las secuencias DET-I).

A modo de resumen, la tabla 3.4 muestra los enriquecimientos (presencia en más del 50%) para análisis con motivos cold-related pero presencias menores a 50% para los heat-related.

**Tabla 3.4. Resultados obtenidos para motivos Heat- y Cold-related y set de secuencias target**

<table>
<tr><th>Análisis</th><th colspan="2">Enriquecimiento (vs. set control y set genoma)</th></tr>
<tr><td>DETI - Heat</td><td rowspan="6">No hubo enriquecimiento mayor a 50% de ningún motivo heat-related en ningún set de secuencias</td><td rowspan="2">DREB2A CRT/DRE o ERF1 bs, HSEI, HSEIII</td></tr>
<tr><td>DEPI - Heat</td></tr>
<tr><td>DETR - Heat</td><td>DREB2A CRT/DRE</td></tr>
<tr><td>DEPR - Heat</td><td></td></tr>
<tr><td>DETA - Heat</td><td rowspan="2">DREB2A CRT/DRE</td></tr>
<tr><td>DETyPI - Heat</td></tr>
<tr><td>DETI - Cold</td><td colspan="2">CRT/DRE<br>ABRE-RELATED, OsROS-bZIP bs I, EEII</td></tr>
<tr><td>DEPI - Cold</td><td colspan="2">CRT/DRE, ICE1 bs I<br>MYBC1 bs, OsROS-bZIP bs I<br>EEII, ZAT10 bs III</td></tr>
<tr><td>DETR - Cold</td><td colspan="2">CRT/DRE, EEII, HOS9 bs<br>ZAT10 bs IV, ICE1 bs I, MYBC1 bs, ABRE-RELATED<br>CM3, AS1/OCS element, TA5 box</td></tr>
<tr><td>DEPR - Cold</td><td colspan="2">GmaColdNovel II, CG-1 element II, AS1/OCS element<br>TA5 box, TA1 box, MYB15 bs II, EEII</td></tr>
<tr><td>DETA - Cold</td><td colspan="2">CRT/DRE, ABRE-RELATED, OsROS-bZIP bs I<br>TA5 box, ZAT10 bs IV<br>EEII, ICE1 bs I</td></tr>
<tr><td>DETyPI - Cold</td><td colspan="2">CRT/DRE, ABRE-RELATED, OsROS-bZIP bs I<br>TA5 box</td></tr>
</table>

Los resultados obtenidos con los motivos heat-related, revelaron que los factores HSF (clásicos en la respuesta al estrés por calor) no serían responsables generales de las diferencias transcripcionales observadas bajo el tratamiento térmico. Sin embargo, si bien no se vio enriquecido, ciertos sets de secuencias target presentaron subsets (de promedio 30% de los sets originales) con la presencia de sitios de unión DREB2A y/o ERF1 bs, ambos demostrados importantes para la tolerancia al calor (Sakuma y col., 2006; Cheng y col., 2013).

Por otro lado, los análisis realizados con los mismos sets y el grupo de motivos relacionados con estrés por bajas temperaturas, revelaron diversos motivos enriquecidos en más del 50% de las secuencias target de cada set. CRT/DRE es un sitio de unión a factores AP2 de tipo DREB1/CBF, conservados en varias especies vegetales (sensibles o no al frío) y clásicos en la respuesta a bajas temperaturas (Vogel y col., 2005, Pons y col., 2014). ABRE-RELATED es un motivo que se encontró enriquecido en los promotores de transcriptos expresados por frío y plantas de *A.thaliana* y *O.sativa* que poseen expresión constitutiva de factores DREB1/CBF (Lindlof y col., 2009). ICE1 es otro factor de transcripción que actúa upstream a los factores *CBF*s para la señalización por bajas temperaturas, en variadas especies vegetales (Chinnusamy y col., 2003). EEII se encontró enriquecido en promotores de A. thaliana inducibles por frío (Maruyama y col. 2012). ZAT10 es un factor con roles positivos para estrés por frío y salino (Sakamoto y col., 2004). CG-1 element es el sitio de unión de factores CAMTA1, 3 y 5, y, junto a CM3, se encontró presente en los promotores de *ZAT12* y *CBF2* (Doherty y col., 2009). MYB15 bs II es otro sitio presente en promotores de *AtCBF*s y MYB15 se demostró interaccionar con ICE1 y ser regulador negativo de la adquisición de tolerancia al frío (Agarwal y col., 2006).

OsROS-bZIP II y III (as1/ocs element) fue descubierto en arroz, como parte de un potencial regulon CBF-independiente que nuclea señalización por ROS con estímulo por frío (Cheng y col., 2007). MYBC1 se demostró independiente a los regulones CBF y ZAT12, y regulador negativo de la tolerancia al frío de *A.thaliana* (Zhai y col., 2010). GmaColdNovel II se encontró enriquecido en promotores de soja inducibles por frío (Maruyama y col., 2012). TA5 box y TA1 box son potenciales sitios de unión de OsMYBS3, un factor de arroz importante en la tolerancia al frío de esta especie (Su y col., 2010).

El enriquecimiento de los motivos descriptos, podría ser la característica que define a los promotores estudiados como afectados por el tratamiento térmico y, a las familias de factores de transcripción que unirían a estos sitios, como posibles actores principales en la regulación

transcripcional bajo el mismo estímulo. Es interesante que los elementos *cis* enriquecidos en los sets de genes afectados por tratamiento térmico hayan sido reportados de poseer roles en condiciones de bajas temperaturas. Esta podría indicar la interacción de los estímulos de altas y bajas temperaturas, combinación que efectivamente sucede al aplicar el tratamiento térmico preventivo del daño por frío y luego refrigerar la fruta. Los resultados obtenidos invitan a realizar estudios acerca de las familias de factores de transcripción de duraznero que unirían a los sitios identificados como enriquecidos.

Es interesante destacar la posibilidad de que las familias de factores de transcripción conocidas en las respuestas al calor y al frío, no necesariamente conserven sus rasgos (cantidad de miembros, estímulo/s bajo el/los que actúan, regulación) en las distintas especies vegetales. Por ejemplo, Artlip y col. (2013) identificaron motivos CRT/DRE en los propios promotores de factores *CBF*s de duraznero, familia de cinco miembros en comparación con los tres reportados de *A. thaliana*. Otro aspecto a considerar es la promiscuidad que ciertos factores de transcripción pueden presentar (un ejemplo se puede visualizar en Cheng y col., 2013). Estos aspectos permiten admitir variadas posibilidades de funcionalidad de los elementos *cis* identificados.

En cuanto al método utilizado para los análisis realizados, se podría decir que trabajar solamente con dos grupos de motivos que contienen sitios de unión para factores de transcripción caracterizados en diversas especies vegetales, tiene ventajas y desventajas. No elegir otros motivos, desconsidera a otros factores de transcripción que actúen bajo el tratamiento térmico. Sin embargo, dada la gran cantidad de información que se puede obtener, es una ventaja poder empezar con motivos relacionados (en al menos una especie vegetal) con el estímulo de interés, acortando el camino para la identificación de factores de transcripción importantes bajo el mismo. No se descartan la utilidad de nuevos análisis que complementen los resultados obtenidos, aplicando variabilidad en el largo de las secuencias a tomar, generando nuevos grupos de promotores más relacionados aún (por ejemplo, mediante la bibliografía), y eligiendo otros grupos de motivos que complementen los análisis realizados.

# *4. Conclusiones*

Como resultados del presente trabajo fue posible:

**Desarrollar el programa cisAnalyzer,** diseñado en lenguajes Perl y R. El mismo permite realizar búsquedas, analizar enriquecimiento y co-ocurrencia de motivos en grupos de secuencias biológicas, de tipo nucleotídicas o peptídicas. Es de fácil utilización y deja disponible al usuario diversas opciones que permiten una gran variedad análisis según sus necesidades y creatividad.

**Generar la base de elementos *cis* de promotores vegetales PASF (Plant Abiotic Stress and Phytohormone-related motifs)**, la cual fue construida en asociación a cisAnalyzer y curada manualmente. La misma posee más de 300 motivos probados experimentalmente como sitios de unión de factores de transcripción en diversas especies vegetales.

**Construir una base de datos a partir de los determinantes moleculares identificados en durazno por los enfoques ómicos aplicados** en nuestro laboratorio, la cual es de suma importancia para el conocimiento de los mecanismos moleculares del tratamiento térmico en duraznos y la selección de factores que intervienen en sus efectos preventivos frente al daño por frío.

**Caracterizar *in silico* de regiones promotoras de genes con expresión modificada bajo tratamiento térmico en duraznos, utilizando cisAnalyzer y los grupos de motivos de PASF db relacionados a estímulo por altas y bajas temperaturas.** Se destaca el enriquecimiento de los motivos CRT/DRE, OsROS-bZIP bs I, ABRE-RELATED e ICE1 bs en la mayoría de los sets de promotores construidos, debido a que participan en la respuesta al frío en diversas especies vegetales. Estos resultados sugieren la **movilización de mecanismos moleculares de respuesta al frío (a nivel de la regulación transcripcional) como efecto del tratamiento térmico en duraznos.**

## *5. Referencias*


Agarwal, M., Hao, Y., Kapoor, A., Dong, C-H., Fujii, H., Zheng, X., y Zhu, J-K. (2006) A R2R3 type Myb transcription factor is involved in the cold regulation of cbf genes and in acquired freezing tolerance. *J Biol Chem.* **281**(49), 37636-37645.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., y Walter, P. (2011) Molecular Biology of the Cell. 4th Edition. New York: Garland Science.

Artlip, T.S., Wisniewski, M. E., Bassett, C. L., y Norelli, J. L. (2013) CBF gene expression in peach leaf and bark tissues is gated by a circadian clock. *Tree Physiol* 33, 866-877.

Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25:3389-3402.

Bailey, T. L., Bodén, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., Noble, W. S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, 202-208.

Brummell, D.A., Dal Cin, V., Lurie, S., Crisosto, C.H. y Labavitch, J.M. (2004) Cell wall metabolism during the development of chilling injury in cold-stored peach fruit: association of mealiness with arrested disassembly of cell wall pectins. *J Exp Bot* 155, 2041-2052.

Bustamante, C.A., Budde, C.O., Borsani, J., Lombardo, V.A., Lauxmann, M.A., Andreo, C.S., Lara, M.V. y Drincovich, M.F. (2012) Heat treatment of peach fruit: modifications in the extracellular compartment and identification of novel extracellular proteins. *Plant Physiol Biochem* 60, 35-45.

Chang, W-C., Lee, T-Y., Huang, H-D., Huang, H-Y., y Pan, R-L. (2008) PlantPAN: Plant promoter analysis navigator, for identifying combinatorial cis-regulatory elements with distance constraint in plant gene groups. *BMC Genomics*. **9**, 561.

Cheng, C., Yun, K-Y., Ressom, H. W., Mohanty, B., Bajic, V. B., Jia, Y., Yun, S., J., y de los Reyes, B. G. (2007) An early response regulatory cluster induced by low temperature and hydrogen peroxide in seedlings of chilling-tolerant japonica rice. *BMC Genomics.* **8**, 175.

Cheng, M-C., Liao, P-M., Kuo, W-W. y Lin, T-P. (2013) The Arabidopsis ETHYLENE RESPONSE FACTOR1 regulates abiotic stress-responsive gene expression by binding to different cis-acting

elements in response to different stress signals. *Plant Physhiol.* **162**, 1566-1582.

Chinnusamy, V., Ohta, M., Kanrar, S., Lee, B.H., Hong, X., Agarwal, M., y Zhu, J. K. (2003) ICE1: a regulator of cold-induced transcriptome and freezing tolerance in Arabidopsis. *Genes Dev*. **17**(8), 1043-1054.

Crick, F.H.C. (1958) On Protein Synthesis. Symp. Soc. Exp. Biol. XII, 139-163.

Davuluri, R. V., Sun, H., Palaniswamy, S. K., Matthews, N., Molina, C., Kurtz, M., y Grotewold, E. (2003) AGRIS: Arabidopsis Gene Regulatory Information Server, an information resource of Arabidopsis cis-regulatory elements and transcription factors. *BMC Bioinformatics*, **4**, 25.

Doherty, C.J., Van Buskirk, H.A., Myers, S.J., y Thomashow, M.F. (2009) Roles for Arabidopsis CAMTA transcription factors in cold-regulated gene expression and freezing tolerance. Plant Cell 21, 972-984.

Genero, M., Gismondi, M., Monti, L. L., Gabilondo, J., Budde, C. O., Andreo, C. S., Lara, M. V., Drincovich, M. F., y Bustamante, C. A. (2016) Cell wall-related genes studies on peach cultivars with differential susceptibility to woolliness: looking for candidates as indicators of chilling tolerance. *Plant Cell Reports*.

Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N. y Rokhsar D.S. (2011) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* 40, 1178-118.

Hertz, G. Z., y Stormo, G. D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563-577.

Higo, K., Ugawa, Y., Iwamoto y M., Korenaga, T. (1999) Plant cis-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Res.*, **27**, 297–300.

Jin, P., Zheng, Y., Tang, S., Rui, H., y Wang C. H. (2009) A combination of hot air and methyl jasmonate vapor treatment alleviates chilling injury of peach fruit. *Postharvest Biol. Technol.* **52**, 24-29.

Jung, S., Ficklin, S., Lee, T., Cheng, C-H., Blenda, A., Zheng, P., Yu, J., Bombarely, A., Cho, I., Ru., S., Evans, K., Peace., C., Abbott, A. G., Mueller, L. A., Olmstead, M. A., y Main, D. (2014) The Genome Database for Rosaceae (GDR): year 10 update. *Nucleic Acids Res.* **42** (1), 1237-1244.

Lamesch, P., Berardini, T.Z., Li, D.,Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D.L., Garcia-Hernandez, M.,Karthikeyan, A.S., Lee, C.H., Nelson, W.D., Ploetz, L., Singh, S.,Wensel, A. y Huala, E. (2011) The Arabidopsis Information Resource (TAIR): improved gene

annotation and new tools. *Nucleic Acids Res* 40, D1202–D1210.

Lara, M.V., Borsani, J., Budde, C.O., Lauxmann, M.A., Lombardo, V., Murray, R., Andreo, C.S. y Drincovich, M.F. (2009) Biochemical and proteomic analysis of "Dixiland" peach fruit (Prunus persica) upon heat treatment. *J Exp Bot* 60, 4315-4333.

Lauxmann, M.A., Brun, B., Borsani, J., Bustamante, C. A., Budde, C. O., Lara M.V., y Drincovich, M.F. (2012) Transcriptomic profiling during the post-harvest of heat-treated Dixiland Prunus persica fruits: common and distinct response to heat and cold. *PloS ONE* 7 (12), e51052.

Lauxmann, M.A., Borsani, J., Osorio, S., Lombardo, V.A., Budde, C. O., Bustamante, C. A., Monti, L.L., Andreo, C.S., Fernie, A.R., Drincovich, M.F. y Lara M.V. (2014) Deciphering the metabolic pathways influencing heat and cold responses during post-harvest physiology of peach fruit. *Plant Cell Environ* 37(3), 601-616.

Lawrence, C. E., Altschul, S. F., Bogouski, M. S., Liu, J. S., Neuwald, A. F. , y Wooten, J. C. (1993) Detecting sublte sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**,208-214.

Lindlof, A., Bräutigam, M., Chawade, A., Olsson, O. y Olsson, B. (2009) In silico analysis of promoter regions from cold-induced genes in rice (Oryza sativa L.) and Arabidopsis thaliana reveals the importance of combinatorial control. *Bioinformatics*. 25(11), 1345-1348.

Lescot, M., Déhais, P., Thijs, G., Marchal, K., Moreau, Y., Van de Peer., Y., Rouzé, P. y Rombauts, S. (2002) PlantCARE, a database of plant *cis*-acting regulatory elements and a portal to tools for *in silico* analysis of promoter sequences. *Nucleic Acids Res.*, **30**, 325–327.

Lurie, S. (1998) Postharvest heat treatments. *Postharvest Biol Technol* 14, 257-269.

Lurie, S. y Crisosto, C.H. (2005) Chilling injury in peach and nectarine. *Postharvest Biol Technol* 37, 195-208.

Maruyama, K., Todaka, D., Mizoi, J., Yoshida, T., Kidokoro, S., Matsukura, S., Takasaki, H., Sakurai, T., Yamamoto, Y.Y., Yoshiwara, K., Kojima, M., Sakakibara, H., Shinozaki, K., y Yamaguchi-Shinozaki, K. (2012) Identification of cis-acting promoter elements in cold- and dehydration-induced transcriptional pathways in Arabidopsis, rice, and soybean. *DNA Res* 19, 37-39.

Pedreschi, R. y Lurie, S. (2015) Advances and current challenges in understanding postharvest abiotic stresses in perishables. *Postharvest Biol Technol* 107, 77-89.

Peng, C-H., Hsu, J-T., Chung, Y-S., Lin, Y-J., Chow, W-C., Hsu, D. F., y Tang, C. Y. (2006)

Identification of degenerate motifs using position restricted selection and hybrid ranking combination. *Nucleic Acids Res.* **34,** 6379-6391.

Pons, C., Marti, C., Forment, J., Crisosto, C. H., Dandekar, A. M., y Granell, A. (2014). A bulk segregant gene expression analysis of a peach population reveals components of the underlying mechanism of the fruit cold response. *Plos ONE*, **9** (3), e90706.

Sakuma, Y., Maruyama, K., Qin, F., Osakabe, Y., Shinozaki, K., y Yamaguchi-Shinozaki, K. (2006) Dual function of an Arabidopsis transcription factor DREB2A in water-stress-responsive and heat-stress-responsive gene expression. *Proc Natl Acad Sci*. **103**(49), 18822-18827.

Sakamoto, H., Maruyama, K., Sakuma,Y., Meshi, T., Iwabuchi, M., Shinozaki, K., y Yamaguchi-Shinozaki, K. (2004) Arabidopsis Cys2/His2-type zinc-finger proteins function as transcription repressors under drought, cold, and high-salinity stress conditions. *Plant Physiol* 136, 2734-2746.

Sandelin, A., Alkema, W., Engström, P., Wasserman, W., y Lenhard, B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*. **32**, 91-94.

Sebastian, A., y Contreras-Moreira, B. (2014) footprintDB: a database of transcription factors with annotated cis elements and binding interfaces.

Su, C.F., Wang, Y.C., Hsieh, T.H., Lu, C.A., Tseng, T.H., y Yu, S.M (2010) A novel MYBS3-dependent pathway confers cold tolerance in rice. *Plant Physiol*. **153**(1), 145-158.

Tao, T. Standalone BLAST Setup for Unix. 2010 May 31 [Updated 2014 Apr 18]. In: BLAST® Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2008-. Available from: http://www.ncbi.nlm.nih.gov/books/NBK52640/.

Tisdall, J. (2001) Beginning Perl for Bioinformatics. 1era edición. Editorial O'Reilly.

Thompson, W., Rouchka, E. C., y Lawrence, C. E. (2003) Gibbs recursive sampler: finding transcription factor binding sites. *Nucleic Acids Res*. **31**, 3580-3585.

Vogel, J.T., Zarka, D.G., Van Buskirk, H.A., Fowler, S.G. y Thomashow, M.F. (2005) Roles of the CBF2 and ZAT12 transcription factors in configuring the low temperature transcriptome of Arabidopsis. *Plant J* 41, 195-211.

Wall, L., Christiansen, T., y Orwant, J. (2000) Programming Perl. 3era edición. Editorial O'Reilly.

Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I., Matys, V., Meinhardt, T., Prüss, M., Reuter, I., y Schacherer, F. (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res*. **28**(1), 316-319.

Wisniewski, M. E., Bassett, C.,L., Renaut, J., Farrell, R. Jr., Tworkoski, T., y Artlip, T. S. (2006) Differential regulation of two dehydrin genes from peach (Prunus persica) by photoperiod, low temperature and water deficit. *Tree Physiology*, **26**, 575-584.

Zhai, H., Bai, X., Zhu, Y., Li, Y., Cai, H., Ji, W., Ji, Z., Liu, X., Liu, X., y Li, J. (2010) A single-repeat R3-MYB transcription factor MYBC1 negatively regulates freezing tolerance in Arabidopsis. *Biochem Biophys Res Commun*. **394**(4), 1018-1023.

Zhang, L., Yu, Z., Jiang, L., Jiang, J., Luo, H., y Fu, L. (2011) Effect of post-harvest heat treatment on proteome change of peach fruit during ripening. *J. Proteomics,* **74**, 1135-1149.