Análisis *in-silico* de la expresión de genes sHSPs en frutos de tomate *Solanum lycopersicum*.

Débora Pamela Arce

Farmacéutica, Universidad Nacional "Juan A. Maza"

Dra. en Ciencias: área Biología, Universidad Nacional de Mar Del Plata

Este Trabajo Final de Especialización es presentado como parte de los requisitos para optar al grado académico de Especialista en Bioinformática, de la Universidad Nacional de Rosario y no ha sido presentada previamente para la obtención de otro título en esta u otra Universidad. La misma contiene los resultados obtenidos en investigaciones llevadas a cabo en el grupo de BioAgroinformática del Centro Internacional Franco Argentino de Ciencias de la Información y de Sistemas (CIFASIS-CONICET), en la Cátedra de Genética de la Facultad de Ciencias Agrarias de la Universidad Nacional de Rosario y en el Grupo de Análisis, Desarrollos e Investigaciones Biomédicas (GADIB) dependiente de la Facultad Regional San Nicolás de la Universidad Tecnológica Nacional, bajo la dirección de la Dra. Flavia J. Krsticevic.

Índice

Introducción

El tomate (*Solanum lycopersicum*) es nativo de Sudamérica, se halla ampliamente distribuido en todo el planeta y es considerado una de las principales especies hortícolas del mundo[1]. Sin embargo, a pesar de su importancia económica, la secuenciación completa del genoma de la variedad comercial S. *lycopersicum* (cv. Heinz 1706) fue realizada por el Tomato Genome Consortium en el año 2012 (Tomato & Consortium, 2012). Así, fue estimado que aproximadamente 35.000 genes codificantes de proteínas componen el genoma de *S. lycopersicum* en ~ 950 Mb secuenciadas. Comparativamente, el tamaño del genoma del tomate con el de otras especies cultivables resulta pequeño. Esta característica mencionada, sumada la facilidad de controlar experimentalmente las condiciones de cultivo, además de la experiencia acumulada en la selección artificial por genetistas y fitomejoradores, hacen al tomate un modelo biológico de referencia para programas de mejoramiento vegetal dentro de las Solanáceas (Aoki et al., 2013). Adicionalmente, diversos análisis focalizados en el estudio de especies silvestres y emparentadas con la variedad comercial, han permitido estudiar más exhaustivamente la evolución del tomate a partir de la secuenciación de 360 variedades (Lin et al., 2014). Los datos obtenidos aportarán las bases para el análisis genómico comparativo y los diversos aspectos evolutivos facilitarán la selección de aquellos genes o familias génicas de interés para programas de mejoramiento vegetal.

Las *small heat shock proteins* (sHSPs) constituyen una familia multigénica que se caracteriza por el bajo peso molecular las proteínas (~ 20 kDa) codificadas por sus miembros. Estas proteínas, presentan actividad de chaperonas moleculares en respuesta al estrés térmico (*heat shock* o HS) previniendo la agregación proteica irreversible dentro de la célula (Basha, O'Neill, & Vierling, 2012; Poulain, Gelly, & Flatters, 2010). Sin embargo, el HS no es el único estímulo que induce la expresión de las sHSPs. La síntesis de sHSPs se induce durante la maduración de

[1] Estadísticas de la producción del tomate según la FAO: http://faostat3.fao.org/browse/Q/QC/E

los frutos (Lawrence, Cline, & Moore, 1997; Löw, Brändle, Nover, & Forreiter, 2000; Neta-Sharir, Isaacson, Lurie, & Weiss, 2005) y en ciertos estadios del desarrollo en Arabidopsis y Solanáceas (Faurobert et al., 2007; Prasinos, Krampis, Samakovli, & Hatzopoulos, 2005). De forma característica, esta familia es antigua (450 m.a) ya que se encuentra en la base del árbol de la vida. Así, su presencia está ampliamente distribuida en los tres dominios de organismos, Archaea, Bacteria y Eukarya (Waters, 2012). Sin embargo, aunque todas comparten la función de chaperona su número es variable entre los organismos. Notablemente en plantas, comparado con otros organismos (Fig. 1), su frecuencia es alta: 19 genes en *A. thaliana* (Scharf, Siddique, & Vierling, 2001), 39 en *O. sativa* (Ouyang, Chen, Xie, Wang, & Zhang, 2009) y 51 en *G. max* (Lopes-Caitar et al., 2013). Durante la era pre-genómica, el estatus de la familia sHSP en tomate era difuso, siendo posible caracterizar experimentalmente a sólo 14 sHSPs (Alba et al., 2005; Baniwal et al., 2004; Frank et al., 2009; Lee et al., 2012; Sanmiya, Suzuki, Egawa, & Shono, 2004). En la era post-genómica, el número de sHSPs caracterizadas se duplicó, identificándose un total de 26 genes correspondientes a sHSPs involucrados en diferentes situaciones de estrés en diversos tejidos, incluyendo HS en hojas (Fragkostefanakis et al., 2015) y microesporas (Frank et al., 2009). Sin embargo, en este sentido y tal como ocurrió para otras familias multigénicas en tomate (Andolfo et al., 2014) la familia de sHSPs permanece indefinida.

La poliploidía y por lo tanto los genes de copia múltiple sobrevivientes a la selección natural o artificial son frecuentes en los genomas vegetales. En este contexto, tres eventos antiguos de duplicación del genoma fueron detectados en el tomate (Tomato & Consortium, 2012). Por lo que se espera que la familia multigénica de sHSP pueda presentar copias múltiples. Las copias múltiples y también las secuencias repetitivas y altamente repetitivas presentan un problema a la hora de ensamblar un genoma: se podría decir que son resistentes a la mayoría de las estrategias de ensamblaje de genomas. Esto se debe a que, si por ejemplo existe una alta identidad de secuencia entre parálogos, sus secuencias son colapsadas en una única copia no siendo posible distinguirlas (Krsticevic, Santos, Januário, Schrago, & Carvalho, 2010). Por lo tanto, su número

queda sub-representado en el genoma analizado. También, si la secuencia es repetitiva, como en el caso de transposones o regiones heterocromáticas, los software de ensamblado son incapaces de decidir en qué lugar del genoma mapearía esa secuencia. Por lo que quedan *gaps* en algunas regiones del genoma que no pueden ser completadas. Un genoma completo debe incluir todas las regiones del genoma. Actualmente, esta limitante intrínseca de la tecnología de secuenciación Sanger o Illumina (*reads* cortos o muy cortos) se resolvió con las nuevas tecnologías de moléculas largas y únicas, como SMRT- PacBio (Krsticevic, Schrago, & Carvalho, 2015). Es por ese motivo, que en un principio se planteó en este trabajo final de curso, hacer una búsqueda exhaustiva de secuencias relacionadas y candidatas/ putativas a sHSPs, en las diferentes bases de datos genómicas y en los transcriptomas.

Simplemente, la presencia del dominio característico conservado en las sHSPs o *alpha-crystallin domain* IPR008978 (ACD o HSP20 domain) no es suficiente para definir la cantidad y las relaciones de los genes que conforman la familia de sHSPs (Bondino, Valle, & Ten Have, 2012). Es por ello, que la expresión diferencial de las secuencias putativas de sHSPs, podría dar indicios acerca de su función biológica como chaperonas. Gracias a la introducción de tecnologías RNA-seq es posible caracterizar los transcriptomas y cuantificar la expresión génica durante diferentes situaciones de estrés y estadios del desarrollo en plantas, así como para la identificación de genes o sus isoformas, (Conesa et al., 2016; Nagalakshmi, Waern, & Snyder, 2010). Brevemente, las etapas del análisis de datos provenientes de un experimento típico de RNA-seq, constan principalmente del control de calidad, alineamiento de *reads* (con o sin genoma de referencia), medición de los niveles de transcriptos y aproximaciones estadísticas para analizar la expresión diferencial génica entre tratamientos o condiciones. De esta manera, mediante el análisis de la expresión diferencial de secuencias putativas sHSPs se pretende obtener evidencias adicionales a la de identidad de secuencia y a la de estructura del dominio ACD o HSP20, que ayuden a definir el status funcional de los posibles miembros la familia sHSP.

Basados en estos antecedentes, nuestro objetivo general fue primeramente anotar la familia de genes de sHSPs en el tomate *S. lycopersicum* (cv.Heinz 1706). Para eso se ensambló el transcriptoma de esta especie a partir de resultados experimentales de RNA-seq públicos y paralelamente se realizó una búsqueda exhaustiva en diferentes bases de datos. Los *reads* para tres condiciones fisiológicas del fruto (estadios de madurez del tomate, EMT: verde maduro o V, naranja o N y rojo maduro o R) fueron descargados de la base de datos DDBJ. Así, mediante el análisis de expresión diferencial de sHSPs en frutos N y R comparados con fruto V, se pretendió anotar funcionalmente y cuantificar los transcriptos de sHSPs que estarían vinculados con la maduración en *S. lycopersicum* (cv. Heinz 1706).

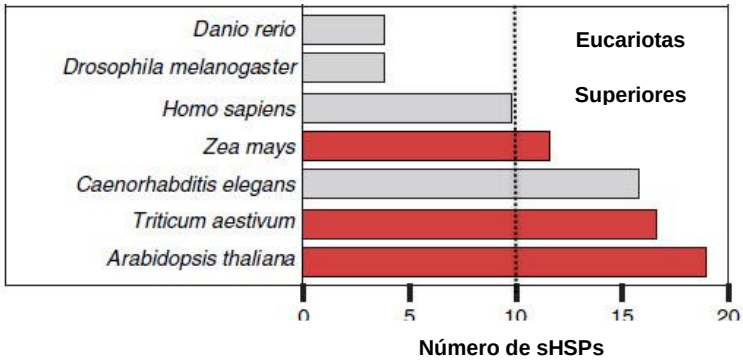


Figura 1. Número de miembros de sHSPs. El nùmero varía entre los organismos eucariotas superiores; en plantas (barras coloreadas en rojo) la mayoría exceden los 10 miembros. Adaptada de Haslbeck, et al.,(Haslbeck, Franzmann, Weinfurtner, & Buchner, 2005).

Objetivos específicos

1. Ensamblar los transcriptomas de distintos estadios de madurez del fruto de tomate *S. lycopersicum* (cv. Heinz 1706) a partir de lecturas de secuenciación de nueva generación disponibles en repositorios públicos.
2. A partir de diferentes bases de datos y de los transcriptomas ensamblados en el Obj.1, anotar las secuencias de los genes putativos sHSPs en *S. lycopersicum* (cv. Heinz 1706).
3. Constatar si los genes sHSPs estarían vinculados al proceso de maduración del fruto de tomate mediante la cuantificación de la abundancia de transcriptos sHSPs en distintos estadios de maduración del fruto de tomate y su expresión diferencial entre dichos estadios.

Hipótesis de trabajo

Las secuencias putativas sHSPs de *S. lycopersicum* (cv. Heinz 1706) serán consideradas con funcionalidad de chaperona si, alguno de los genes parálogos se expresa diferencialmente durante los estadios de maduración del fruto (V, N y R).

Materiales y métodos

Diseño experimental

El diseño experimental fue descripto en la sección 1.18.1 del material suplementario (doi:10.1038/nature11119) por el Tomato Genome Consortium (Tomato & Consortium, 2012). Las plantas de tomate *S. lycopersicum* (cv. Heinz 1706) se crecieron en invernadero bajo un fotoperiodo de 14 hs/26 °C día y 10 hs/16 °C noche. Las plántulas crecidas en sustrato se utilizaron para cosechar raíces y hojas. Las plantas maduras se utilizaron para cosechar botones florales cerrados y abiertos (flores). El resto de las flores se autopolinizaron y se cosecharon los frutos en diversos estadios: verde inmaduro de 1 cm, 2 cm, 3 cm, verde maduro (V), naranja (N, maduración temprana) y rojo maduro (R, 10 días post N). Los tejidos se pulverizaron con nitrógeno líquido y se conservaron a -80 °C. La extracción de ARN y la preparación de las bibliotecas se realizaron como describe Zhong, S. et al. (Zhong et al., 2011) con 12 muestras independientes multiplexando con barcodes las secuenciadas en una línea con la plataforma Illumina HiSeq2000. Se prepararon dos réplicas biológicas de los diferentes órganos de la misma edad y estadio para cada tejido y luego fueron secuenciadas. La secuenciación resultó en 7.4 – 14 millones de *reads*, de 48 - 53 pb de longitud, single ended por réplica y por muestra (ver Tabla S1). Las *reads* obtenidas se depositaron en el NCBI Sequence Read Archive bajo el número de acceso SRA049915.

Mapeo con un genoma de referencia de los transcriptomas de frutos V, N y R

Para estimar el tamaño de las bibliotecas para el análisis de perfiles de expresión de las sHSPs putativas durante la maduración, los *reads* de *S. lycopersicum* (cv. Heinz 1706) correspondientes a dos réplicas biológicas del fruto R (SRR404328, SRR404329), fruto N (SRR404327, SRR404326) y del fruto V (SRR404324, SRR404325) se combinaron en un único set de datos de RNA-seq utilizando los recursos bioinformáticos del DDBJ *Read Annotation*

Pipeline y se mapearon contra el genoma de referencia de *A. thaliana* (TAIR 8) utilizando el programa Bowtie2 version 2.0 (Langmead, Trapnell, Pop, & Salzberg, 2009).

Adicionalmente, el número de *reads* de las sHSPs putativas fue calculado con respecto al número de *reads* totales. Este porcentaje refleja actividad transcripcional de los miembros de la familia génica respecto del transcriptoma específico (V, N y R) y total (ver Tabla 2).

*Ensamblaje "*de novo*" de los transcriptomas de frutos V, N y R*

Se utilizaron algunos programas disponibles para el ensamblaje *de novo* en el DDBJ, como Velvet-Oases (Zerbino & Birney, 2008) y Trinity (Haas et al., 2013). El paquete Trinity ofrece la ventaja de trabajar con datos single-ended, y tal es el caso de los datos que se utilizaron en el presente trabajo.

Estadística del ensamblaje

Las pruebas estadísticas del ensamblaje se realizaron utilizando un script en lenguaje Perl diseñado por Joseph Fass y modificado por Brad Sickler del The Bioinformatics Core at UC Davis Genome Center (https://github.com/DeWitP/SFG/blob/master/scripts/count_fasta.pl). Se seleccionó el ensamblaje del transcriptoma que presentaba una mayor cantidad de contigs y un buen N50.

Búsqueda de secuencias putativas sHSPs y análisis de datos transcriptómicos de S. lycopersicum *(cv. Heinz 1706)*

Se realizó un alineamiento BlastP contra la base de datos de proteínas Tomato protein database (ITAG2.4 Release, Sol Genomics Network) utilizando Solyc09g015020 como secuencia aminoacídica "*query*" debido a que presenta el dominio ACD característico de esta familia génica. Con el objeto de capturar todos los posibles miembros de la familia sHSP se buscaron todas las secuencias de proteínas que presentaran el dominio IPR008978 HSP20- like anotado

en Interpro en la Sol Genomics Network database (ITAG release 2.40), https://solgenomics.net/. Paralelamente, las secuencias relacionadas con la palabra clave "small HSP" se recogieron de la base de datos de tomate Helmholtz-Muenchen, http://pgsb.helmholtz-muenchen.de/plant/tomato/. Se obtuvo un total de 58 secuencias putativas de la familia sHSP en *S. lycopersicum* (cv. Heinz 1706). Debido a que las láminas β son características en el dominio conservado ACD (Poulain et al., 2010), se cuantificó el número de láminas β en las sHSPs putativas. El ACD se identificó con el programa PROTEUS2, que permitió acceder a la predicción de estructuras proteicas, http://wishart.biology.ualberta.ca/proteus2 (Montgomerie et al., 2008). Además, el número de láminas β se estimó con el servicio web Promals 3D, http://prodata.swmed.edu/promals3d/promals3d.php, que predice estructuras secundarias a partir de alineamientos múltiples a nivel de secuencias y de estructuras 3 D (Pei & Grishin, 2007).

Análisis de abundancia de transcriptos (cuantificación en FPKM) y expresión diferencial de genes

Los fragmentos por kilobase de transcripto por millón de fragmentos mapeados (del inglés, *Fragments Per Kilobase of transcript per Million mapped fragments* o FPKM) se calcularon para determinar la abundancia de transcriptos (ausencia o presencia de ARNm) de cada gen analizado. Las *reads* mapeadas se obtuvieron como se indica más arriba y los valores de FPKM promedio se calcularon a partir de 2 réplicas biológicas de cada muestra (estadios V, N y R) para cada gen. Los genes se consideraron como expresos cuando el valor promedio de FPKM fue mayor o igual a 2, tal como describe (Steijger et al., 2013).

Los análisis estadísticos descriptivos y para los experimentos de RNA-seq fueron realizados con el paquete edgeR de Bioconductor en lenguaje R, con dos modificaciones (Robinson, McCarthy, & Smyth, 2010). El pipeline estándar de edgeR calcula los tamaños de las librerías como la suma de los *reads* para cada gen, asumiendo que la librería se mapea contra el total del

genoma. En el presente trabajo, los *reads* se mapearon contra 58 regiones definidas del genoma correspondientes a las posibles secuencias de sHSPs por lo que los valores se ajustaron manualmente utilizando comandos GNU. La función calcNormFactors fue omitida debido a que se utiliza cuando se realizan análisis de expresión diferencial génica para todos los transcriptos presentes en un transcriptoma, teniendo en cuenta la totalidad del genoma. El estado de madurez V se tomó como línea base para normalizar los valores correspondientes a frutos N y R relativos a V, donde el log2 fold change (log2FC) se calculó por edgeR. Los valores positivos de log2FC indican inducción, los negativos represión y los valores igual a 0, expresión constante relativa a estado V. Los valores de log2FC mayores a 1 con un p-valor menor a 0.01 fueron incluidos en los gráficos tanto para los estadios N como para R.

Resultados

Ensamblaje de los transcriptomas de frutos de tomate y anotación de sHSPs

Con el objeto de identificar si existen secuencias adicionales en el transcriptoma que no hubiesen sido anotadas por el consorcio de tomate al momento de la secuenciación del genoma (Tomato & Consortium, 2012), se ensamblaron los transcriptomas de provenientes de tres EMT de *S. lycopersicum* (cv. Heinz 1706) mediante dos técnicas de ensamblaje *de novo*: Trinity y Velvet/Oases (Tabla 1). Se realizaron los análisis estadísticos de calidad del ensamblaje para los tres EMT: V, N y R. Se seleccionó el ensamblaje realizado con Trinity debido a que se obtuvieron mayores valores en porcentaje de GC y número total de secuencias (aproximadamente 28000) con respecto a Velvet/Oases (aproximadamente 24000) para los tres EMT analizados (Tabla 1).

Tabla 1. Estadística del ensamblaje *de novo* de los transcriptomas durante la maduración de los frutos de *S. lycopersicum* (cv. Heinz 1706)

Algoritmo	**Trinity**			**Velvet/Oases**		
EMT	**R**	**N**	**V**	**R**	**N**	**V**
Largo de secuencia total (pb)	14506907	12207697	11882634	10805013	9739931	9957309
Nº total de secuencias	28766	26860	28096	23857	24610	27591
N50 (pb)	613	515	460	668	556	495
Conteo GC	6132472	5199716	5073289	4547829	4119990	4213933
GC (%)	42.27	42.59	42.69	42.09	42.3	42.32

Los ensamblajes obtenidos se analizaron con el objeto de identificar transcriptos correspondientes a las posibles sHSPs. Las secuencias identificadas se anotaron por identidad de secuencia utilizando tBlastn. Los resultados de análisis de anotación de secuencias proteicas luego del ensamblaje se adicionaron a las secuencias analizadas en las bases de datos de Solgenomics y Helmholtz-Müenchen como se indica en Materiales y Métodos.

Expresión de genes sHSPs durante la maduración en S. lycopersicum *(cv. Heinz 1706)*

Se analizaron los niveles de expresión en número de FPKM para cada secuencia sHSP putativa. Los *reads* (plataforma Illumina) provenientes de cada situación fisiológica (V, N y R) se alinearon contra las secuencias de 5 kb conteniendo las sHSPs putativas (provenientes del análisis de anotación proteica, datos no mostrados) utilizando el algoritmo Bowtie2 (Langmead et al., 2009). Luego de alinear los *reads* correspondientes a los estadios V, N y R contra todo el genoma de tomate, 0.37%-0.66% del total de los transcriptos corresponden a sHSPs putativas, constituyendo a esta familia génica como un grupo que representa un alto porcentaje del total de transcriptos en cada EMT (Tabla 2). Asimismo, la representatividad de estos transcriptos aumenta con el tiempo de maduración desde frutos V a R (0.65%) cuando se alinearon los *reads* contra todo el genoma de tomate (Tabla 2). El presente trabajo se focalizó en estudiar aquellas posibles sHSPs involucradas en la maduración mediante el análisis temporal de este proceso (desde frutos V a R).

Tabla 2. Alineamiento de *reads* provenientes de tres EMT contra el genoma de referencia durante la maduración *S. lycopersicum* (cv. Heinz 1706)

Muestras Fruto Nº	***Reads* procesadas**	***Reads* procesadas alineadas contra el genoma de referencia**
V-1	8423857	31500(0.37%)
V-2	10214504	38383 (0.38%)
N-1	11534326	44576 (0.39%)
N-2	11111053	43172 (0.39%)
R-1	14085509	93108 (0.66%)
R-2	13515035	88137 (0.65%)

La expresión durante la maduración se lleva a cabo mediante un proceso lineal que se desarrolla en el tiempo (abundancia). Se observó un incremento de los niveles de expresión de 17 secuencias sHSPs desde el EMT V hacia los EMT N y R (ver Tabla 3, Solyc06g076520,

Solyc06g076570, Solyc06g076560, Solyc08g062450, Solyc05g014280, Solyc09g015020, Solyc09g01500, Solyc08g078700, Solyc03g082420, Solyc08g062340, Solyc01g102960, , Solyc03g11390, Solyc11g020330, Solyc06g076540, Solyc04g082720, Solyc12g042830 y Solyc03g123540). También se observaron 3 sHSPs con expresión heterogénea según el EMT: Solyc01g014480 (valores de FPKM de 66.16-229.64-187.44 en V, N y R, respectivamente), Solyc08g078720 (FPKM de 48.28-26.17-37.85) y Solyc02g093600 (FPKM de 2.85-18.48-6.61). Por otro lado se observaron diferencias en el patrón de expresión de 9 secuencias, en donde los valores de FPKM disminuyeron desde el EMT V al R (Solyc07g0624020, Solyc09g011710, Solyc01g098810, Solyc04g082740, Solyc11g071560, Solyc10g076880, Solyc02g080410, Solyc01g009200 y Solyc09g007140). Dos secuencias se expresaron a niveles casi no detectables (Solyc10g086680 y Solyc01g098790). Las diferencias en los niveles de expresión registrados indicarían una posible diferenciación en la regulación transcripcional de estos 31 genes identificados.

Tabla 3. Promedio de la abundancia de transcriptos sHSPs putativos expresados en FPKM.

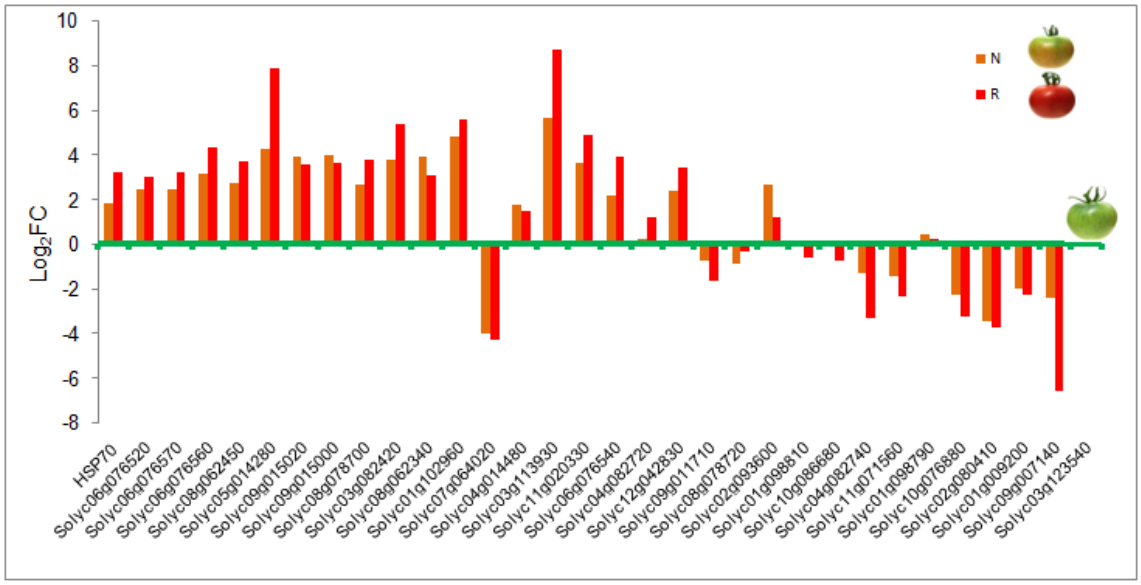
	FPKM		
ID	V	N	R
Solyc06g076520	278.430434	1519.82841	2256.68799
Solyc06g076570	202.801257	1097.8104	1877.25308
Solyc06g076560	78.4012029	689.261176	1555.35156
Solyc08g062450	64.853396	421.062959	819.664401
Solyc05g014280	2.9490299	56.2471681	697.356354
Solyc09g015020	34.0435468	511.23135	410.651139
Solyc09g015000	31.9065733	506.367245	405.392312
Solyc08g078700	29.7094548	186.403575	391.81333
Solyc03g082420	9.56575271	131.105738	390.419182
Solyc08g062340	34.3827865	526.515612	290.482485
Solyc01g102960	5.64556501	156.290185	275.235438
Solyc07g064020	4705.21623	290.949951	245.871806
Solyc04g014480	66.1616902	229.646746	187.449061
Solyc03g113930	0.43090209	22.910089	183.479457
Solyc11g020330	5.82710537	72.5374874	171.197974
Solyc06g076540	10.064224	44.9252109	152.404253
Solyc04g082720	44.0794412	52.4421379	103.57674
Solyc12g042830	8.47584812	44.4285468	87.5254453
Solyc03g123540	17.4079602	19.5617973	52.9168575
Solyc09g011710	186.614016	115.135764	58.9089129
Solyc08g078720	48.2842547	26.1777147	37.8541824
Solyc02g093600	2.85155798	18.4824561	6.61239551
Solyc01g098810	9.28909591	8.97380435	5.98045634
Solyc10g086680	0.97826852	1.11848496	0.59178913
Solyc04g082740	5.50757926	2.29629047	0.55631675
Solyc11g071560	2.61355489	0.92436711	0.50057198
Solyc01g098790	0.26709592	0.37646615	0.33152983
Solyc10g076880	2.71619492	0.54361589	0.26641777
Solyc02g080410	2.59283392	0.22720119	0.17899381
Solyc01g009200	0.85206138	0.19633501	0.16521124
Solyc09g007140	1.446975	0.25895341	0

Expresión diferencial de genes sHSPs durante la maduración de tomate

Con el objeto de evaluar si estas secuencias estarían vinculadas específicamente a los EMT, N y R, se observó la expresión diferencial de 31 transcriptos sHSPs putativas, en frutos N y R

comparando con los niveles de transcriptos en fruto V (Fig. 2). Diecinueve genes se inducen tanto en fruto N como en R: entre ellos 14 muestran mayores niveles para el estadio R que para el N (Solyc06g076520, Solyc06g076570, Solyc06g076560, Solyc08g062450, Solyc05g014280, Solyc08g078700, Solyc03g082420, Solyc01g102960, Solyc03g113930, Solyc11g020330, Solyc06g076540, Solyc04g082720, Solyc03g123540 y Solyc12g042830), y las 5 secuencias restantes poseen mayores niveles relativos en fruto N con respecto al R (Solyc09g015020, Solyc09g015000, Solyc08g062340, Solyc04g014480 y Solyc02g093600). Por otro lado, 8 secuencias se reprimen tanto en fruto N como en R (Solyc07g064020, Solyc09g011710, Solyc04g082740, Solyc11g071560, Solyc10g076880, Solyc02g080410, Solyc01g009200 y Solyc09g007140). Finalmente, 4 secuencias no se expresaron diferencialmente en frutos N o R con respecto al V (Solyc08g078720, Solyc01g098810, Solyc10g086680 y Solyc01g098790).

Figura 2. . Análisis de expresión diferencial en genes sHSPs durante la maduración en *S. lycopersicum* (cv. Heinz 1706).



Funcionalidad de sHSPs

En la Tabla 4 se observa el resultado de expresión diferencial y el análisis de la estructura proteica (ausencia o presencia de ACD, número de láminas β) para cada posible sHSP (ID Solgenomics). La última columna muestra la posible función como chaperona asignada para cada miembro. En este trabajo se consideró funcionalidad como chaperona (sHSP) durante la maduración de los frutos de tomate, a aquellas secuencias (22 en total) que se expresaron diferencialmente (inducción o represión) en los EMT N y/o R comparando con V y con presencia del ACD. Una secuencia con presencia de ACD no se expresó diferencialmente (NDE): Solyc10g086680. Otras dos secuencias no se expresaron diferencialmente (Solyc08g078710 y Solyc08g078720) con ausencia de ACD y se consideraron proteínas HSP20-like o bien aún no podría asignárseles una función mediante los análisis realizados en este trabajo. Hubo 5 casos (Solyc01g009200, Solyc09g007140, Solyc10g076880, Solyc04g082720 y Solyc04g082740) para los cuales se observó expresión diferencial pero con ausencia de ACD que aún no se les puede asignar funcionalidad. Para ello, trabajos futuros en el análisis de sus secuencias promotoras, análisis filogenéticos, como así también modelado de sus estructuras proteicas tridimensionales serán necesarios para dilucidar su funcionalidad. Por último, Solyc01g009220 y Solyc10g086680 son nuevas proteínas, una HSP20-like y la otra sHSP, respectivamente. Es interesante que casi la mitad de las proteínas analizadas (15 secuencias) no se les pudiera asignar localización subcelular con los predictores más utilizados (Target P, http://www.cbs.dtu.dk/services/TargetP/; WoLF PSORT, http://www.genscript.com/wolf-psort.html).

Tabla 4. Anotación de sHSPs en *S. lycopersicum* (cv. Heinz 1706). Expresión diferencial en frutos N y R comparado con fruto V (Ind: inducción; Rep: represión; NDE: no diferencialmente expreso). Estructura proteica y asignación de funcionalidad como chaperonas. (C: citosólica, P: cloroplástica, M: mitocondrial, RE: retículo endoplásmico, PX: peroxisomal, ND: no determinado, * Nuevas proteínas)

ID Solgenomics	Loclización Subcelular	Expresión diferencial	ACD	No. Lam β	Función Chaperona
Solyc06g076520	C	Ind	si	9	sHSP
Solyc06g076570	C	Ind	si	9	sHSP
Solyc06g076560	C	Ind	si	9	sHSP
Solyc06g076540	C	Ind	si	9	sHSP
Solyc09g015020	C	Ind	si	9	sHSP
Solyc09g015000	ND	Ind	si	9	sHSP
Solyc08g062450	C	Ind	si	9	sHSP
Solyc08g062340	C	Ind	si	9	sHSP
Solyc03g123540	C	Ind	si	9	sHSP
Solyc11g020330	RE	Ind	si	9	sHSP
Solyc01g102960	RE	Ind	si	9	sHSP
Solyc03g113930	RE	Ind	si	9	sHSP
Solyc04g014480	PX	Ind	si	9	sHSP
Solyc02g080410	C	Rep	si	8	sHSP
Solyc02g093600	C	Ind	si	9	sHSP
Solyc08g078710	ND	NDE	no	9	HSP20-like
Solyc08g078720	ND	NDE	no	7	HSP20-like
Solyc08g078700	M	Ind	si	9	sHSP
Solyc12g042830	M	Ind	si	9	sHSP
Solyc07g064020	ND	Rep	si	8	sHSP
Solyc03g082420	P	Ind	si	9	sHSP
Solyc05g014280	P	Ind	si	9	sHSP
Solyc01g009200	ND	Rep	no	9	HSP20-like?
Solyc01g009220*	ND	NDE	no	8	HSP20-like
Solyc09g007140	ND	Rep	no	9	HSP20-like?
Solyc11g071560	ND	Rep	si	9	sHSP
Solyc10g076880	ND	Rep	no	7	HSP20-like?
Solyc04g082720	ND	Ind	no	9	HSP20-like?
Solyc04g082740	ND	Rep	no	9	HSP20-like?
Solyc01g098790	ND	NDE	no	7	HSP20-like
Solyc01g098810	ND	NDE	no	9	HSP20-like
Solyc09g011710	ND	Rep	si	9	sHSP
Solyc10g086680*	ND	NDE	si	9	sHSP

Discusión

Treinta y una secuencias se analizaron y se evaluó su expresión diferencial en frutos N y R comparados con fruto V (ver Fig.2 Resultados). Entre ellas, 19 se indujeron y 8 se reprimieron. Seis secuencias diferencialmente expresadas no se les pudieron asignar función como sHSPs. Las secuencias que pertenecen (Solyc06g076520, Solyc06g076570 y Solyc06g076560) mostraron los mayores niveles de abundancia, tal como se observa en los análisis de expresión (ver Tabla 3), expresión diferencial (Fig. 2), mostrados en resultados, y las evidencias experimentales previas provenientes de otros grupos de investigación. Tal como fue reportado por el grupo de Goyal et al (2012) en donde tres sHSPs (Sl17.6, Sl20.0, y Sl20.1) localizadas en el cromosoma 6, se expresan durante la maduración del fruto de tomate en *S. lycopersicum* variedad Ohio 8245 y presentan abundancia relativa en otros tejidos de la planta.

Entre las secuencias de sHSPs analizadas en el presente trabajo, se encuentran 3 proteínas en retículo endoplásmico (RE, Solyc11g020330, Solyc01g102960 y Solyc03g113930), 2 cloroplásticas (P, Solyc03g082420 y Solyc05g014280), 10 citoplasmáticas (C, Solyc06g076520, Solyc06g076570, Solyc06g076560, Solyc06g076540, Solyc09g015020, Solyc08g062450, Solyc08g062340, Solyc03g123540, Solyc02g080410, Solyc02g093600), y 2 secuencias mitocondriales (M, Solyc08g078700, Solyc12g042830). La presencia de este set de secuencias sHSPs como así también su patrón de expresión se repite en otros sistemas experimentales tal como hojas (Fragkostefanakis et al., 2015) y microesporas (Frank et al., 2009) inducidas por HS, aunque las sHSPs peroxisomales no se observaron en el último caso. La clase RE de sHSPs (Solyc11g020330 y Solyc03g113930) es crítica para la síntesis de lípidos de membrana, plegamiento y homeostasis celular (Zeng et al., 2015). Evidencias experimentales sugieren que Solyc11g020330 (BAA97658 protein) provee protección térmica de proteínas bacterianas solubles *in vitro* en *E. coli* (Mamedov & Shono, 2008).

Al menos una sHSP mitocondrial debe estar presente (Tabla 4), aunque se encontraron otras 4

sHSPs en el genoma de tomate que poseen localización subcelular mitocondrial: Solyc08g078700 se induce en fruto R y en microesporas tratadas con HS, mientras que Solyc12g042830 se induce en hojas y microesporas tratadas con HS (Fragkostefanakis et al., 2015). El fruto de tomate puede experimentar HS durante el desarrollo y la maduración de los frutos (Sato et al., 2006; Sumesh & Ghildiyal, 2008). Aún en aquellas zonas donde el tomate se cultiva para su comercialización pueden experimentar altas temperaturas (Ferguson, Snelgar, Lay-Yee, Watkins, & Bowen, 1998). Aparentemente las sHSPs M que se inducen rápidamente durante el HS en respuesta a este tipo de estrés ambiental (Sanmiya, Suzuki, Egawa, & Shono, 2004). Matas et al (Matas et al., 2011) reportaron que Solyc08g078700 expresa diferencialmente (TU031561unigene) en cuatro tejidos de pericarpio. Estos resultados indicarían que al menos una sHSP M es necesaria durante la respuesta al HS y al estrés que sucede en el fruto. El estrés oxidativo durante el desarrollo y la maduración como así también el estrés abiótico térmico desencadenan respuestas en la expresión de las sHSPs, sugiriendo un rol importante de esta familia génica en la homeostasis celular.

Solyc03g082420 se expresa diferencialmente en el fruto R, hojas y microesporas estresadas térmicamente (Fragkostefanakis et al., 2015). Esta secuencia codifica para LeHsp21 que es una proteína cloroplástica (P) que ha sido descripta por numerosos autores: se trasloca al cromoplasto en desarrollo y se induce durante la maduración de tomate (Lawrence, Cline, & Moore, 1997; Matas et al., 2011; Srivastava, Gupta, Datsenka, Mattoo, & Handa, 2010). Las líneas transgénicas sobreexpresantes de esta proteína P acumulan carotenoides tempranamente cuando se compara con líneas normales en ausencia de estrés. Todas estas evidencias indican que LeHsp21 participa en la acumulación de carotenoides durante la maduración y en la conversión de cloroplastos en cromoplastos (Neta-Sharir, Isaacson, Lurie, & Weiss, 2005) protegiendo a otras proteínas del estrés oxidativo en el fruto (Lambert et al., 2011). El gen P Solyc05g014280 se expresa en forma similar a LeHsp21 pero está ausente en pericarpio (TU048847 in Matas et al., 2011). Además, las sHSPs pueden actuar en las membranas

tilacoides cloroplásticas como proteínas estabilizadoras estabilizando membranas con funciones antioxidantes frente a HS y estrés oxidativo en organismos fotosintetizadores (Yu et al., 2012). Todos estos hallazgos sugieren que el proceso de maduración involucra un cambio masivo estructural en los plástidos que simulan un estrés de tipo ambiental (Giovannoni, 2001).

Por otro lado, 8 sHSPs se reprimen durante el proceso de maduración (ver Resultados). Fragkostefanakis et al. (2014) reportaron represión en la transcripción de dos sHSPs (Solyc09g011710 y Solyc07g064020) en hojas de *S. lycopersicum* bajo estrés térmico. La represión en la expresión génica puede esperarse dado que el proceso de maduración constituye un fenómeno de senescencia en donde los procesos de estabilización y degradación proteica desempeñan roles determinantes (Srivastava, 2010). Las secuencias Solyc10g076880 y Solyc03g005190 se reprimen y aún no puede establecerse si son proteínas de tipo HSP20-like dado que no presentan el ACD. En arroz, algunos miembros de tipo HSP20-like mostraron similares patrones de expresión diferencial: muchos de ellos mostraron expresión constitutiva en órganos vegetativos y otros fueron reprimidos durante el HS (Sarkar, Kim, & Grover, 2009).

Por último, los resultados de expresión diferencial por represión observados en el presente trabajo para las proteínas de tipo sHSPs y HSP20-like se analizaron mediante el estudio de sus regiones promotoras (datos no mostrados). Ninguno de los promotores de estas secuencias reprimidas mostró el típico motivo de unión a factores de transcripción *Heat Shock Element* (HSE). El HSE se encuentra presente en todos los promotores génicos de sHSPs en diferentes organismos eucariotas en respuesta al estrés (Pelham, 1985). Esta podría ser una de las causas de su patrón diferencial y debe ser estudiado en mayor profundidad.

Conclusiones

Utilizando datos públicos disponibles de experimentos de RNA-seq, se analizó la expresión de esta familia génica en tres estadios de la maduración del fruto de tomate (EMT: verde V, naranja N y rojo R). El análisis de sHSPs y otras proteínas HSP20-like durante EMT se abordó por dos estrategias: cuantificación de niveles de transcriptos y comparación de los mismos en N y R con respecto a V (expresión diferencial). Los patrones de expresión fueron diferenciales según EMT: de 58 secuencias analizadas, 19 se indujeron diferencialmente en N y R, siendo todas sHSPs funcionales y 8 se reprimieron (4 sHSPs funcionales). El resto no se expresaron diferencialmente en fruto N y fruto R con respecto al V. Adicionalmente, las secuencias Solyc01g009220 y Solyc10g086680 son novedosas, una HSP20-like y la otra sHSP, respectivamente y aportaron nueva evidencia experimental durante EMT. Se identificó un set mínimo de sHSPs con diferente localización subcelular que estarían involucradas en EMT que implicaría un rol diferencial de estas sHSPs en el proceso fisiológico de la maduración. Cabe mencionar que la mayoría de aquellas secuencias a las que no pudo asignárseles funcionalidad como sHSP durante la maduración, tampoco pudo asignárseles localización subcelular, por lo que futuros estudios a nivel de proteínas y secuencias regulatorias podrían elucidar su posible funcionalidad como chaperonas o no durante este proceso fisiológico.

Por todo lo anterior, se montaron los transcriptomas de frutos V, N y R y se buscaron las secuencias putativas sHSPs que contenían el dominio HSP20-like, no encontrándose más miembros, por lo que podría tratarse de todas las sHSPs existentes de *S. lycopersicum* variedad Heinz 1706. Estas proteínas se anotaron funcionalmente de acuerdo con los criterios de expresión diferencial en frutos N y R con respecto a V, y su estructura proteica. Algunas secuencias no pudo asignárseles funcionalidad como sHSP o HSP20-like bajo los criterios utilizados en el presente trabajo. Finalmente la expresión diferencial tanto inducción como represión génica se correlaciona con el EMT según el avance de este proceso fisiológico lineal en el tiempo y da cuenta del proceso de estrés que conlleva la maduración en donde estas

proteínas se hallan involucradas.

Resumen

El presente informe de trabajo fue el inicio de la línea de investigación de la familia de sHSPs durante la maduración en diferentes cultivares de tomate, entre los grupos de investigación de la Universidad Tecnológica Nacional (GADIB-UTN), el CIFASIS-CONICET y la cátedra de Genética de la Facultad de Ciencias Agrarias de la UNR. El objetivo principal fue la anotación funcional de los miembros putativos sHSP en *S. lycopersicum* (cv. Heinz 1706). Se abordó el análisis de expresión RNA-seq de la familia génica sHSPs durante la maduración de los frutos de tomate, comparando los transcriptomas en frutos N y R con respecto a fruto V. Se identificaron los perfiles de expresión de los miembros sHSPs que estarían involucrados en la maduración. La complejidad del proceso de maduración se evidencia en la diversidad de las sHSPs. Finalmente, los mecanismos de regulación de la transcripción de esta familia génica aún deben ser elucidados.

Referencias bibliográficas

Alba, R., Payton, P., Fei, Z., McQuinn, R., Debbie, P., Martin, G. B., … James J. Giovannoni. (2005). Transcriptome and Selected Metabolite Analyses Reveal Multiple Points of Ethylene Control during Tomato Fruit Development. *Plant Cell*, *17*(November), 2954–2965. http://doi.org/10.1105/tpc.105.036053.1

Andolfo, G., Jupe, F., Witek, K., Etherington, G. J., Ercolano, M. R., & Jones, J. D. G. (2014). Defining the full tomato NB-LRR resistance gene repertoire using genomic and cDNA RenSeq. *BMC Plant Biology*, *14*(1), 120. http://doi.org/10.1186/1471-2229-14-120

Aoki, K., Ogata, Y., Igarashi, K., Yano, K., Nagasaki, H., Kaminuma, E., & Toyoda, A. (2013). Functional genomics of tomato in a post-genome-sequencing phase. *Breeding Science*, *63*(1), 14–20. http://doi.org/10.1270/jsbbs.63.14

Baniwal, S. K., Bharti, K., Chan, K. Y., Fauth, M., Ganguli, A., Kotak, S., … von Koskull-Döring, P. (2004). Heat stress response in plants: a complex game with chaperones and more than twenty heat stress transcription factors. *Journal of Biosciences*, *29*, 471–487. http://doi.org/10.1007/BF02712120

Basha, E., O'Neill, H., & Vierling, E. (2012). Small heat shock proteins and α-crystallins: dynamic proteins with flexible functions. *Trends in Biochemical Sciences*. http://doi.org/10.1016/j.tibs.2011.11.005

Bondino, H. G., Valle, E. M., & Ten Have, A. (2012). Evolution and functional diversification of the small heat shock protein/α-crystallin family in higher plants. *Planta*, *235*(6), 1299–313. http://doi.org/10.1007/s00425-011-1575-9

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., … Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, *17*, 13. http://doi.org/10.1186/s13059-016-0881-8

Faurobert, M., Mihr, C., Bertin, N., Pawlowski, T., Negroni, L., Sommerer, N., & Causse, M. (2007). Major proteome variations associated with cherry tomato pericarp development

and ripening. *Plant Physiology*, *143*(3), 1327–46. http://doi.org/10.1104/pp.106.092817

Fragkostefanakis, S., Simm, S., Paul, P., Bublak, D., Scharf, K. D., & Schleiff, E. (2015). Chaperone network composition in Solanum lycopersicum explored by transcriptome profiling and microarray meta-analysis. *Plant, Cell and Environment*, *38*(4), 693–709. http://doi.org/10.1111/pce.12426

Frank, G., Pressman, E., Ophir, R., Althan, L., Shaked, R., Freedman, M., … Firon, N. (2009). Transcriptional profiling of maturing tomato (Solanum lycopersicum L.) microspores reveals the involvement of heat shock proteins, ROS scavengers, hormones, and sugars in the heat stress response. *Journal of Experimental Botany*, *60*(13), 3891–3908. http://doi.org/10.1093/jxb/erp234

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., … Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, *8*, 1494–512. http://doi.org/10.1038/nprot.2013.084

Haslbeck, M., Franzmann, T., Weinfurtner, D., & Buchner, J. (2005). Some like it hot: the structure and function of small heat-shock proteins. *Nature Structural & Molecular Biology*, *12*, 842–846. http://doi.org/10.1038/nsmb993

Krsticevic, F. J., Santos, H. L., Januário, S., Schrago, C. G., & Carvalho, a B. (2010). Functional copies of the Mst77F gene on the Y chromosome of Drosophila melanogaster. *Genetics*, *184*(1), 295–307. http://doi.org/10.1534/genetics.109.107516

Krsticevic, F. J., Schrago, C. G., & Carvalho, a B. (2015). Long-Read Single Molecule Sequencing to Resolve Tandem Gene Copies: The Mst77Y Region on the Drosophila melanogaster Y Chromosome. *G3 (Bethesda, Md.)*, *5*(6), 1145–50. http://doi.org/10.1534/g3.115.017277

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*(3), R25.

http://doi.org/10.1186/gb-2009-10-3-r25

Lawrence, S. D., Cline, K., & Moore, G. A. (1997). Chromoplast development in ripening tomato fruit: identification of cDNAs for chromoplast-targeted proteins and characterization of a cDNA encoding a plastid-localized low-molecular-weight heat shock protein. *Plant Molecular Biology*, *33*, 483–492.

Lee, J. M., Joung, J.-G., McQuinn, R., Chung, M.-Y., Fei, Z., Tieman, D., … Giovannoni, J. (2012). Combined transcriptome, genetic diversity and metabolite profiling in tomato fruit reveals that the ethylene response factor SlERF6 plays an important role in ripening and carotenoid accumulation. *The Plant Journal : For Cell and Molecular Biology*, *70*(2), 191–204. http://doi.org/10.1111/j.1365-313X.2011.04863.x

Lin, T., Zhu, G., Zhang, J., Xu, X., Yu, Q., Zheng, Z., … Huang, S. (2014). Genomic analyses provide insights into the history of tomato breeding. *Nature Genetics*, *46*(11), 1220–6. http://doi.org/10.1038/ng.3117

Lopes-Caitar, V. S., de Carvalho, M. C. C. G., Darben, L. M., Kuwahara, M. K., Nepomuceno, A. L., Dias, W. P., … Marcelino-Guimarães, F. C. (2013). Genome-wide analysis of the Hsp20 gene family in soybean: comprehensive sequence, genomic organization and expression profile analysis under abiotic and biotic stresses. *BMC Genomics*, *14*, 577. http://doi.org/10.1186/1471-2164-14-577

Löw, D., Brändle, K., Nover, L., & Forreiter, C. (2000). Cytosolic heat-stress proteins Hsp17.7 class I and Hsp17.3 class II of tomato act as molecular chaperones in vivo. *Planta*, *211*, 575–582. http://doi.org/10.1007/s004250000315

Montgomerie, S., Cruz, J. a, Shrivastava, S., Arndt, D., Berjanskii, M., & Wishart, D. S. (2008). PROTEUS2: a web server for comprehensive protein structure prediction and structure-based annotation. *Nucleic Acids Research*, *36*(Web Server issue), W202-9. http://doi.org/10.1093/nar/gkn255

Nagalakshmi, U., Waern, K., & Snyder, M. (2010). RNA-Seq: a method for comprehensive

transcriptome analysis. *Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel ... [et Al.]*, *Chapter 4*, Unit 4.11.1-13. http://doi.org/10.1002/0471142727.mb0411s89

Neta-Sharir, I., Isaacson, T., Lurie, S., & Weiss, D. (2005). Dual role for tomato heat shock protein 21: protecting photosystem II from oxidative stress and promoting color changes during fruit maturation. *The Plant Cell*, *17*, 1829–1838. http://doi.org/10.1105/tpc.105.031914

Ouyang, Y., Chen, J., Xie, W., Wang, L., & Zhang, Q. (2009). Comprehensive sequence and expression profile analysis of Hsp20 gene family in rice. *Plant Molecular Biology*, *70*(3), 341–357. http://doi.org/10.1007/s11103-009-9477-y

Pei, J., & Grishin, N. V. (2007). PROMALS: Towards accurate multiple sequence alignments of distantly related proteins. *Bioinformatics*, *23*(7), 802–808. http://doi.org/10.1093/bioinformatics/btm017

Poulain, P., Gelly, J.-C., & Flatters, D. (2010). Detection and Architecture of Small Heat Shock Protein Monomers. *PLoS ONE*, *5*(4), e9990. http://doi.org/10.1371/journal.pone.0009990

Prasinos, C., Krampis, K., Samakovli, D., & Hatzopoulos, P. (2005). Tight regulation of expression of two Arabidopsis cytosolic Hsp90 genes during embryo development. *Journal of Experimental Botany*, *56*, 633–644. http://doi.org/10.1093/jxb/eri035

Sanmiya, K., Suzuki, K., Egawa, Y., & Shono, M. (2004). Mitochondrial small heat-shock protein enhances thermotolerance in tobacco plants. *FEBS Letters*, *557*(1), 265–268.

Scharf, K., Siddique, M., & Vierling, E. (2001). The expanding family of Arabidopsis thaliana small heat stress proteins and a new family of proteins containing ⎵ -crystallin domains (Acd proteins). *Cell Stress & Chaperones*, *6*, 225–237.

Tomato, T., & Consortium, G. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, *485*, 635–41. http://doi.org/10.1038/nature11119

Waters, E. R. (2012). The evolution, function, structure, and expression of the plant sHSPs.

Journal of Experimental Botany, *64*(2), 391–403. http://doi.org/10.1093/jxb/ers355

Zerbino, D. R., & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, *18*(5), 821–9. http://doi.org/10.1101/gr.074492.107

Zhong, S., Joung, J. G., Zheng, Y., Chen, Y. R., Liu, B., Shao, Y., … Giovannoni, J. J. (2011). High-throughput illumina strand-specific RNA sequencing library preparation. *Cold Spring Harbor Protocols*, *6*(8), 940–949. http://doi.org/10.1101/pdb.prot5652