PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR

FACULTAD DE HÁBITAT, INFRAESTRUCTURA Y CREATIVDAD

TÍTULO DE MAGISTER EN BIOLOGÍA COMPUTACIONAL

Comprendiendo la diversidad genética de *Glycine max* a nivel de célula única.

Trabajo de Titulación

Autora: Ing. Rojas Chango, Mélida Guadalupe MSc.

Director: PhD. Cervantes Pérez, Alan Sergio MSc.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ÍNDICE DE ANEXOS

RESUMEN

La soya, *Glycine max,* es un cultivo de importancia alimenticia para el ser humano y para los animales de granja. Además, la soya se cultiva para propósitos de generación de biocombustibles lo que vuelve a este cultivo de uso versátil y de gran peso en la industria agroalimentaria. Es necesario entonces desarrollar planes de fitomejoramiento que permitan aprovechar de mejor manera el cultivo aumentando su productividad. Además, es importante concentrarse en mejoras de resistencia ante patógenos y formas de volver resiliente al cultivo en frente del cambio climático y sus múltiples consecuencias. Con estos antecedentes el presente estudio buscó comprender la expresión genética de genes de soya con variación estructural, genes asociados a resistencia a *Phythophthora sojae* y genes asociados al estrés hídrico. Para este propósito se intersecó datos de un pangenoma de soja de 26 genomas, una lista de genes asociados a resistencia a *P. sojae* y genes asociados a estrés hídricos con los datos pertenecientes a expresión génica de secuenciación de célula única expuestos en *Tabula Glycine*. Se encontró que los genes con variaciones estructurales se expresan altamente en células de tejido de semilla y raíz y meristemo floral, y los genes de resistencia y genes asociados a productividad se expresan en mayor medida en tejidos como haz vascular, hoja verdadera, semilla y cotiledón.

Palabras clave: soya, estudio comparativo, secuenciación de célula individual, pangenoma, fitomejoramiento, resistencia.

ABSTRACT

Glycine max, a crop of main importance to agroindustry, is a key element in the nutritional status of both human beings and farm animals. *G. max* is also important for biofuel production, making it a versatile plant. Therefore, many studies to understand its genetics have been conducted. A pangenome of soy of 26 soy varieties that shows the structural variations among the genomes has been published. Also, a tissue-specific transcriptome study of soybean obtained from single-cell RNA-sequencing is published in *Tabula Glycine*. In this study, we intersected this information using the reference genome Willams82 version 4 (Wm82v4), to find the state of expression of structural variations in Wm82v4. In addition, a list of genes of interest for better production and resistance to abiotic stress and *P. sojae* was used. It was found that the genes with structural variations had higher expression in seeds, roots, and floral meristem, whereas the genes associated with productivity and resistance were mostly expressed in tissues such as vascular bundle, true leaves, seeds, and cotyledon. This information is of great importance for the implementation of breeding plans that allow obtaining a soy cultivar of high productivity and resistance to pests such as *Phythophthora sojae* and abiotic stress.

Keywords: Soybean, Comparative analysis, single-cell sequencing, pangenome, plant breeding.

1. INTRODUCCIÓN

El cultivo de soya es importante extensamente en áreas de la industria de alimentos y agricultura. En la industria de alimentos, el aceite de soya ha sido mejorado para ser más saludable que otros aceites convencionales y para aportar compuestos anticancerígenos. En el caso de la agricultura, la soya ha sido ampliamente usada en la ganadería como alimento. Sin embargo, es necesario desarrollar variedades de soya que sean mayormente productivas, con resistencia a enfermedades y condiciones adversas del medio ambiente. Para poder alcanzar estos objetivos, grandes volúmenes de datos correspondientes a estudios de transcriptomas, genomas y pangenomas se han generado. Por un lado, se ha estudiado el pangenoma de 26 especies silvestres de soya y se ha publicado un pangenoma gráfico (Y. Liu et al., 2020) y un pangenoma tridimensional(Ni et al., 2023). Además, recientemente, se ha publicado *Tabula Glycine*, una base de datos de expresión génica correspondiente a todos los órganos de la planta y a diferentes estados de desarrollo que cuenta con una resolución a nivel de célula única (Cervantes-Pérez, Thibivilliers, et al., 2024). Esta gran cantidad de datos puede ser aprovechada para entender la variación genética de los genes de soya en cuanto a los tipos de célula y variedad de cultivo. En este estudio para poder entender la variación genética de soya en los diferentes estadíos de desarrollo y tejidos, se ha tomado los datos de expresión génica a nivel de tipo celular como referencia. Además, se han tomado los genes que han sido identificados que tienen variaciones estructurales en el pangenoma de 26 genomas de soya alineados al genoma Zhonghuang 13 (Zh13). Se han buscado los genes homólogos en el genoma Williams 82 anotación 4 (Wm82v4) y se ha observado como es el comportamiento de expresión de estos genes en *Tabula Glycine*.

También se buscó por literatura una lista de genes asociados a resistencia a enfermedades y producción de la semilla que sirvieron para estudiar la expresión génica siguiendo un enfoque de estudio dirigido a la mejora de producción de semilla, resistencia a estrés del medio ambiente y resistencia a enfermedades de *G. max*. Este análisis puede dar información valiosa para orientar los programas de fitomejoramiento en pro de alcanzar los objetivos de incremento de productividad y resistencia.

1.1. Planteamiento del problema

El estudio comparativo del transcriptoma de la soya a nivel de tipos celulares con los pangenomas de soya de 26 genomas silvestres ofrece información de la correlación de expresión génica a nivel de tipo celular y las variaciones estructurales entre genomas. Es importante estudiar si existen variaciones estructurales en Wm82 para saber si por ejemplo una inversión o una deleción podría cambiar la producción de un cultivo. Más adelante se puede asociar las variaciones estructurales con un fenotipo específico de una variedad de soya e implementar esta información en un programa de fitomejoramiento.

Además, conocer la respuesta de expresión de genes involucrados en procesos de crecimiento vegetativo, producción de semilla y resistencia a *Phythophthora sojae* a nivel de tipo celular y por especie puede arrojar información valiosa para aplicarla en programas de fitomejoramiento del cultivo. Al momento, a mi leal saber y entender aún no se ha estudiado esta correlación a nivel de tipo celular volviendo necesario el presente estudio que puede aportar a los planes de fitomejoramiento del cultivo.

La producción mundial de soya fue de 420' 580 000 toneladas métricas en 2024, lo que indica la importancia del este cultivo a nivel industrial en el mundo entero (*Soybeans | USDA Foreign Agricultural Service*, n.d.) En América del Sur, la producción de soya fue liderada por Brasil en 2024 con 169'000 000 de toneladas métricas, mientras que Ecuador tuvo una producción de 29 000 toneladas métricas en el mismo año. Si bien la producción no llega al millón de toneladas métricas, la soya es sin duda un cultivo alimenticio de importancia en el país, siendo el cuarto de mayor producción después del maíz, arroz, y aceite de palma (*Ecuador Production*, n.d.).
La importancia de este cultivo radica en su uso alimenticio para el ser humano, animales de granja, su utilidad en la producción de biocombustibles. Debido a la importancia de este cultivo existen esfuerzos de fitomejoramiento que permitan evitar pérdidas en su producción debido a enfermedades y estrés ocasionado por cambios en el medio ambiente. Además, se busca incrementar la producción mejorando las características fenotípicas y genéticas del cultivo. Para cumplir con estas metas se han desarrollado diversos estudios genéticos del cultivo y aquí se recalcan los estudios relacionados a la identificación de genes relevantes para el fitomejoramiento de soya. En cuanto a *P. sojae,* se ha desarrollado un estudio cuantitativo de resistencia de locus en el cromosoma 18 de *G. max (Robertson et al., 2018).* En el caso de pérdidas asociadas a stress abiótico, se ha estudiado los genes involucrados en la marchitez de los tallos de soya, fenómeno que está relacionado al estrés por escasez de agua (Chamarthi et al., 2021). Finalmente, en el caso del mejoramiento de la producción del cultivo, se han desarrollado estudios cuantitativos de rasgos de loci (QTL, por sus siglas en inglés) asociados a la altura de la planta y número de nodos en el tallo, lo cual está directamente relacionado con la producción del cultivo (Li et al., 2021; Wang et al., 2022).
Contando con información relevante sobre los genes involucrados en atributos de mejoramiento de la soya, se hace necesario comprender como es su expresión génica a nivel de tipo celular. Comprender la variación genética de la soya a nivel celular ofrecerá nueva información que se puede aplicar a planes de fitomejoramiento del cultivo.

1.2. Preguntas de investigación

¿Cuál es la correlación existente entre la expresión génica de los diferentes tipos de células de *G. max* presentados en *Tabula Glycine* y el pangenoma de 26 genomas de *G. max*?

¿Cuál es el nivel de expresión de los genes con variaciones estructurales en el pangenoma de de *G.max* a nivel de tipo de célula?

¿Cuáles son las variaciones estructurales que mayor expresión muestran a nivel de tipo de célula y cuáles de estos genes con variaciones puede influenciar el fitomejoramiento de *G. max*?

¿Cómo está configurada la expresión de los genes de resistencia a *P. sojae*, genes relacionados a la producción de *G. max*, y genes relacionados a la tolerancia de estrés hídrico en *Tabula Glycine*? ¿Cuáles son los tejidos con mayor expresión génica?

¿En comparación a *Tabula Glycine*, cómo es la configuración de los tejidos con mayor expresión de los genes con variaciones estructurales estudiados en el pangenoma de 26 genomas de *G. max*?

2. MARCO TEÓRICO

El cultivo de soya es importante en muchas áreas de la industria, por ejemplo, en la industria alimenticia el aceite de soya genéticamente modificada ha sido desarrollado para ofrecer un mejor cumplimiento de estándares que aceites convencionales. Modificaciones a la producción de ácido linoleico, oleico, y palmítico en soya han vuelto al aceite de soya transgénico una opción apetecible en el mercado. El aceite de soya con contenido de ácido linoleico (18:3) de 1% puede reducir la necesidad de hidrogenación para alcanzar estabilidad y vida en percha del aceite. Esto es beneficioso para la salud pues ácidos grasos que no han sido producidos a través de la hidrogenación o trans-hidrogenación no traen consecuencias negativas para el sistema circulatorio del ser humano. Por el contrario, el incremento de ácido palmítico (18:1) de 25% a 80% también mejora la estabilidad del aceite y el tiempo de vida en percha(Fehr, 2007) . Finalmente, el aceite con alto contenido de ácido oleico es rico en proteínas y además se ha observado que inhibe el crecimiento de células cancerígenas en el colon, hígado y pulmón(Rayaprolu et al., 2013).

Desde el origen de la soya en China alrededor de 5000 años atrás, han existido esfuerzos para mejorar el cultivo(Hymowitz & Shurtleff, 2005). Existen registros de hibridación artificial a partir de 1900(Bradshaw, 2017), y se continúa uniendo esfuerzos para mejorar el cultivo de soya. Uno de estos esfuerzos es el uso de la biología molecular para estudiar el perfil genético del cultivo que permita tomar decisiones más informadas al momento de hibridar. Así mismo, los avances en tecnología de secuenciación y el desarrollo de la técnica de secuenciación de célula única han permitido el análisis de transcripción de *G. max*, el cuál ha sido estudiado en diferentes niveles. Transcriptómas de nódulos de la raíz en estado de desarrollo y en estado maduro han sido estudiados(Cervantes-Pérez, Zogli, et al., 2024; Z. Liu et al., 2023). Además, recientemente se ha publicado *Tabula Glycine*, una base de datos que contiene información de transcripción correspondiente a todos los órganos de la planta en varios niveles de desarrollo con una resolución a nivel de célula única(Cervantes-Pérez, Thibivilliers, et al., 2024). Finalmente existen varias bases de datos como SoyBase, SoyKB, SoyFGB, SoybeanGDB y SoyOmics que proveen datos de genomas, pangenomas, transcriptomas y proteínas de *G. max*. Esta información permite realizar un análisis del perfil genético de soya a detalle para contribuir al fitomejoramiento de *G. max*, sin embargo, al tratarse de volúmenes de datos extensos es necesario usar herramientas bioinformáticas para realizar los análisis de forma efectiva.

2.1. Fitomejoramiento

El fitomejoramiento de plantas es un proceso de selección de plantas que realiza el ser humano y que empezó cerca de 13000 años atrás(Bradshaw, 2017). Mediante este proceso se puede seleccionar plantas con rasgos fenotípicos deseados, por ejemplo, tamaño del fruto en el caso de una fruta comestible, tamaño de la flor, en el caso de un ornamental, o tamaño de semilla en el caso de semillas comestibles. El fitomejoramiento realizado por el ser humano conglomera alrededor de 2500 especies de plantas que han sido domesticadas alrededor del mundo con más de 160 familias de plantas contribuyentes de una o más especies de cultivos(Meyer et al., 2012).

2.2. Pangenoma

El dogma de la biología molecular que establece que el flujo de información genética en los organismos vivos parte del ADN, se transcribe a ARN y se traduce en proteínas conformadas por aminoácidos ha permitido estudiar la genética de varias especies de plantas. Para ello, fue necesario conocer la molécula de ADN, hecho que fue logrado en 1962 por Watson Crick y Wilkins con el aporte de la cristalografía de ADN hecha previamente por Franklin en 1951.

Seguido de este hito, la técnica de secuenciación de cadenas de nucleótidos desarrollado por Sanger en 1977 permitió conocer las secuencias de ADN de varios organismos. *Arabidopsis thaliana* fue la primera planta cuyo genoma fue secuenciado en el año 2000 y cuya extensión fue de 115.4 Mb (mega bases). A continuación, el primer cultivo productivo secuenciado fue el del arroz publicado en la revista Nature en 2005 con un tamaño de 389 Mb (Bradshaw, 2017).

La mejora de las técnicas de secuenciación y la reducción de su costo a lo largo de los últimos 20 años ha permitido incrementar el número de especies secuenciadas. Esto ha originado gran cantidad de información de genómas y pangenómas, mucha de la cual está disponible en bases de datos de acceso libre. Un pangenoma es el esutudio genómico de varios genomas correspondientes a diferentes especies o variedades de organismos y comprende todos los genes de cada organismo comparado. En el caso de *G. max*, se cuenta con datos de dos estudios de pangenomas y varios genomas (Y. Liu et al., 2020; Ni et al., 2023; Yang et al., 2024).

2.3. Transcriptoma

El transcriptoma se puede definir como la identidad de cada gen expresado y su nivel de expresión en una población de células definida (Velculescu et al., 1997). En términos generales un transcriptoma es el conjunto completo de transcritos presentes en una célula en un tiempo dado. El transcriptoma comprende secuencias de ARN codificadoras o no y puede ser estudiado a varios niveles(Caudai et al., 2021). Se puede tomar tejido juvenil, maduro, senescente o también se puede tomar tejido de diferentes órganos de un organismo para realizar un transcriptoma. En el caso de este estudio, se utilizarán datos de un transcriptoma realizado a partir de datos obtenidos de single-cell sequencing de *G. max*. Varios órganos a varios tiempos de desarrollo de soya fueron tomados para realizar el estudio de transcriptoma que produjo una base de datos denominada *Tabula Glycine*,

la cual se usó en comparación a un pangenoma de 26 genomas de familiares de la misma especie para diferenciar los genes que están siendo expresados de los que no pero que si están presentes en el genoma de acuerdo con los datos del pangenoma.

Usar la amplia información actual sobre transcriptomas podría beneficiar a los programas de fitomejoramiento de soya. Específicamente en el caso de *G. max*, es importante mejorar la producción de semilla, resistencia a roturas por clima adverso, resistencia a condiciones climáticas adversas como sequías, resistencia a enfermedades, y mejorar la composición de aceites y proteínas dentro de las semillas, entre otras. Varios estudios se han realizado para alcanzar estos objetivos, y recientemente se desarrolló un pangenoma que comprende 26 especies silvestres de soya. Este estudio es una fuente de información genética de *G. max*, y puede ser usado como referencia para analizar los datos del transcriptoma de *Tabula Glycine* y resaltar la influencia de la variación genética de las especies de acuerdo a los órganos de *G. max.* Existen varios genes de importancia para el fitomejoramieno de la soya entre ellos se puede encontrar a Glyma.13G190400 y Glyma.19G262700, genes que proveen de resistencia vertical en contra de *Phythopthora sojae* (de Ronne et al., 2020). Además, los factores de transcripción *APETALA2 (AP2), VIVIPAROUS1/ABI3-LIKE (VAL), ENDOSPERMA INDEPENDIENTE DE FERTILIZACIÓN* (FIE), *GLABRA2 (GL2), PICKLE (PKL)*, y UNIÓN DE ADN CON UN DEDO (DOF4), además de los genes *ÁCIDO ABSÍSICO (ABA)-INSENSITIVO 3 (ABI3), COTILEDÓN FRONDOSO 2 (LEC2), FUSCA3 (FUS3), Y WRINKLED1 (WRI1)* están relacionados al desarrollo y relleno de las semillas (Cao et al., 2022; Mosquna et al., 2004; O'Rourke et al., 2014; Pelletier et al., 2017; B. Shen et al., 2006). Estos genes juegan un papel crucial en los objetivos del fitomejoramiento de soya y serán analizados con especial atención en la comparación del pangenoma con *Tabula Glycine*. Con este estudio se busca diferenciar el estado de expresión de estos genes a nivel de órgano y estudiar las posibles moléculas que interaccionan con la expresión de estos genes.

Para poder realizar la comparación de los genes expresados en diferentes órganos de *G. max*, se realizó un blast al genoma de referencia de la variedad Williams 82 (Wm82v4) de *G. max* de los genes con variación estructural del pangenoma de 26 genomas alineados previamente al genoma de referencia ZH13. Los genes homólogos obtenidos fueron buscados en *Tabula Glycine* para ver qué tejidos mostraron mayor expresión génica. Adicionalmente, los genes de importancia de resistencia a *P.sojae,* los genes relacionados a producción y tolerancia a stress hídrico también fueron analizados dentro de *Tabula Glycine* para conocer en qué tejido se expresan mayoritariamente.

3. METODOLOGÍA

Se usó la información generada por Liu et al., (2020) la cual muestra un detalle de los genes con variación genética después de haber comparado 26 genomas de soya domesticada y silvestre con el genoma de soya Zhonghuang 13 (Zh13), el cual es un cultivo de origen chino derivado de las acceciones Yudou 18 y Zhongzuo 90052-76 vía selección de pedigrí por alta producción y tolerancia al stress (Y. Shen et al., 2018). Esta

information de variación estructural génica incluye *presence absence variation (PAV), copy number variation (CNV), inversions, (INV), variability in gene regulation or interaction between elements located on different chromosomes (Trans-inter variation), variation in regulatory or physical interactions between non-adjacent regions within the same chromosome (Trans-intra variation).*

Con el nombre de los genes con variaciones se obtuvo las secuencias de correspondientes a la anotación 2 del Zh13 provenientes del sitio https://mines.legumeinfo.org/legumemine/begin.do. Para esto se usó el script descrito en el anexo1. Debido a los diferentes tipos de transcritos codificantes para cada gen, se obtuvo varias secuencias por cada gen que se usaron más tarde para realizar un blast con el genoma del cultivo de soya Willams 82, ensamblamiento 4 (Wm82v4). Para realizar el blast se descargó la herramienta blastn: 2.16.0+Package: blast 2.16.0 desde https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/ncbi-blast-2.16.0+-x64-linux.tar.gz. Se usó WSL de Windows 11 Pro para instalar blastn para Linux. Además, libgomp1, que fue instalado para completar el funcionamiento de balstn.

Se descargó el genoma ensamblado de Wm82v4 a través de la interfaz de programación de aplicaciones api publicado en la página de NCBI. Se convirtió el genoma de referencia en *data base* para blast y se realizó el blast entre las secuencias de referencia de Zh13 y las secuencias del genoma de soya Wm82v4.

Después del blast se obtuvo los nombres de las secuencias (*sequencies id*) correspondientes a las secuencias de Zh13. Se extrajo los *sequencies id* en un archivo y se creó un script para extraer los nombres de los genes correspondientes a estas secuencias, aliases o locus tag del sitio web del NCBI.

Con los nombres de los genes se procedió a eliminar duplicados. Además, se encontraron 3 genes sin asignación de alias o locus tag, por lo que se quedaron fuera del estudio. Con esta lista de genes se creó un script de R para empatar los nombres de la lista con los nombres de los genes de *Tabula Glycine* y extraer la información de los genes empatados. Se obtuvo una lista final de 15 genes de los cuáles se hizo un mapa de calor.

Siguiendo la misma metodología se extrajo de *Tabula Glycine* la expresión de los genes con *SNPs* correspondientes a líneas de soya expuestas a estrés hídrico. Además, se usó la misma metodología para extraer los genes que por literatura se encontró que tienen importancia productiva y resistencia a *P. sojae*.

Todos los scripts fueron desarrollados por mi persona y mejorados con inteligencia artificial de ChatGpt4.

4. RESULTADOS

4.1 *Tabula Glycine* muestra una gran diversidad de genes expresados. Por el contrario, los genes expresados altamente correspondientes a las variaciones estructurales son pocos.

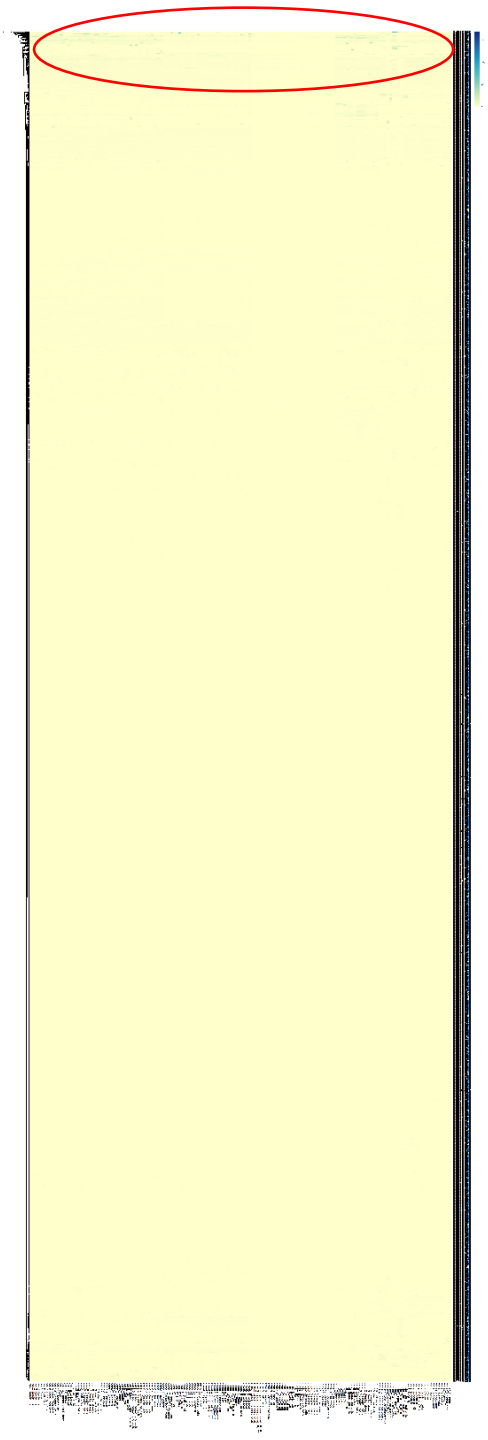

Figura 1 Expresión de todos los genes con variaciones estructurales alineados a Zh13

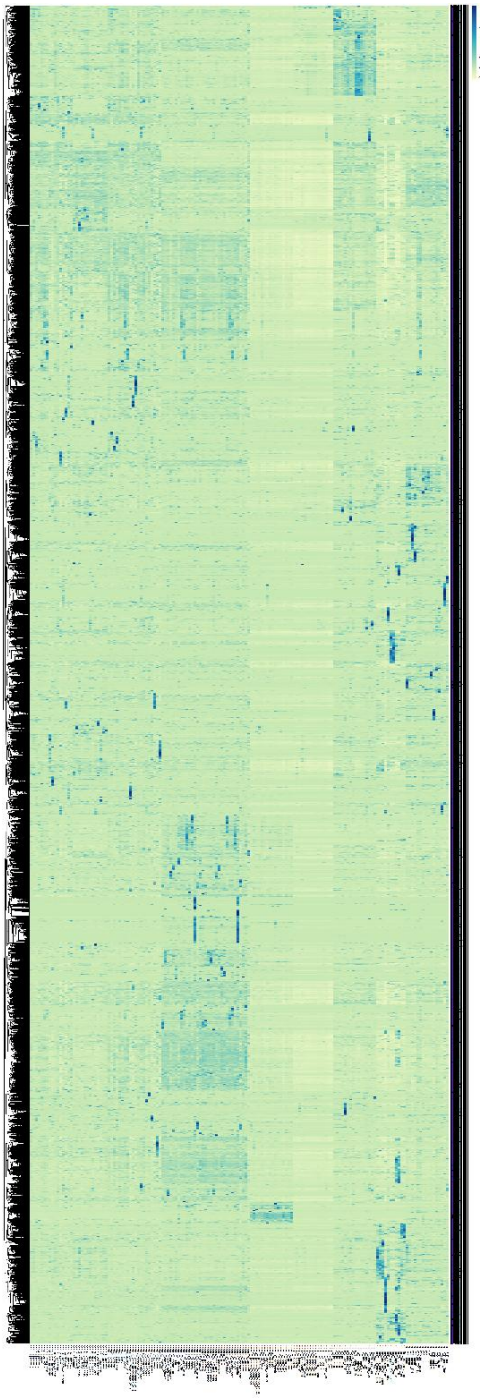

Figura 2 Tabula Glycine

4.2 La expresión de los genes con variaciones estructurales están presentes en tejidos de desarrollo y además su función está relacionada a la división celular y respuesta a estrés.

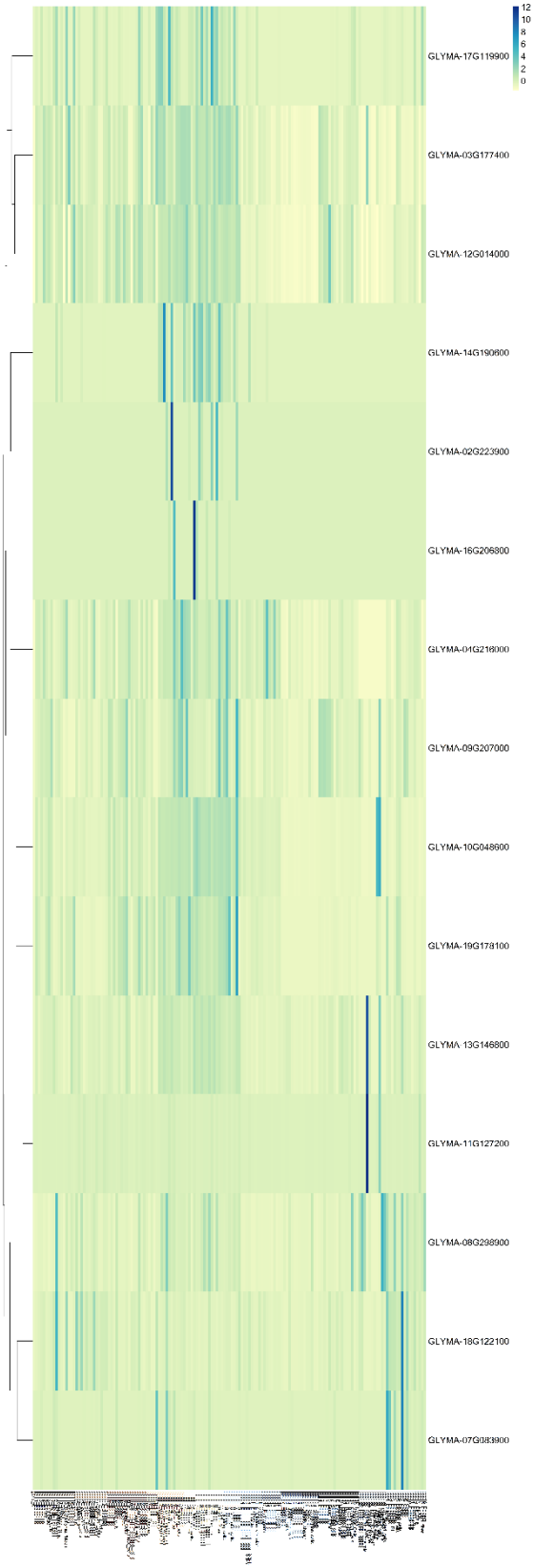

Figura 3Expresión de las variaciones estructurales del pangenoma de 26 genomas de soya alineados al genoma de Zh13.

Se partió de 1252 genes que presentaron variaciones estructurales pertenecientes al pangenoma de 26 genomas alineado a Zh13. Se usó estas secuencias para hacer un blast a Wm82v4. En este blast se encontraron varias secuencias de Wm82v4 que fueron exactamente iguales entre sí, pero que empataron diferentes secuencias con variaciones estructurales del pangenoma. Entonces, se eliminó las secuencias repetidas y se obtuvo un resultado de 15 secuencias correspondientes a 15 genes cuya expresión está representada por un mapa de calor de *Tabula Glycine* (Figura 3). Esta expresión se obtuvo de la base de datos de *Tabula Glycine* cuyas unidades son *Unique Molecular Identifiers* (UMIs). UMIs permiten saber el radio real de moléculas de ARN que inicialmente fueron extraídas de las muestras ya que son etiquetas pegadas a cada molécula lo cual asemeja a un *barcoding*. Sin embargo, a diferencia del *barcoding* que identifica la célula de donde se extrajo la molécula, UMIs identifica la molécula en sí (*What Are Unique Molecular Identifiers and Why Do We Need Them?*, n.d.).

4.3 Las variaciones estructurales identificadas muestran una firma específica de expresión dependiente del tipo celular.

En la Figura 3 se puede observar que los genes Glyma-11G127200, y Glyma-13G146800 tienen una alta expresión correspondiente a un Z-score de 12. Estos genes tienen alta expresión en la capa esclereida del nódulo que es muy superior a la expresión del mismo gen en otros tejidos (Tabla 2). Glyma-11G127200 tiene actividad kinasa de la fructuosa y participa en el proceso metabólico de la fructuosa. Esta función está categorizada como una respuesta a estrés como lo detalla Cao et al., 2022. Por otro lado, Glyma-13G146800 participa en el proceso de unión del grupo hierro-azufre. Los dos genes poseen variaciones estructurales del tipo presencia-ausencia en los genomas del pnagenoma (PAV), variación de regulación de expresión (TRANS-inter), inversión (INV), y variación en el número de copias del gen a lo largo de los genomas del pangenoma (CNV). Los resultados sugieren que esta variación estructural es muy importante para el desarrollo de un tipo celular del nódulo de soya. Lo cual es a su vez importante en el transporte de nutrientes.

Tabla 1Variaciones estructurales (Zh13 en Wm82v4)

Gene	Description	Type of Structural Variation	Literatur e
GLYMA-18G122100	No description available	PAV, TRANS-intra, TRANS-inter, INV, CNV	Liu., et al, 2020
GLYMA-04G216000	K01956 - carbamoyl-phosphate synthase small subunit (carA, CPA1)	PAV, TRANS-intra, TRANS-inter, INV, CNV	
GLYMA-10G048600	K02930 - large subunit ribosomal protein L4e (RP-L4e, RPL4)	PAV, TRANS-intra, TRANS-inter, INV, CNV	
GLYMA-19G178100	K02930 - large subunit ribosomal protein L4e (RP-L4e, RPL4)	PAV, TRANS-intra, TRANS-inter, INV, CNV	

GLYMA-09G207000	ribosome maturation factor	PAV, TRANS-inter, INV, CNV
GLYMA-14G190600	PF05910 - Plant protein of unknown function (DUF868) (DUF868)	PAV, TRANS-intra, TRANS-inter, INV, CNV
GLYMA-02G223900	PF05910 - Plant protein of unknown function (DUF868) (DUF868)	PAV, TRANS-inter, CNV
GLYMA-08G298900	PTHR24413//PTHR24413:SF138 - FAMILY NOT NAMED // F-BOX/KELCH-REPEAT PROTEIN SKIP11	PAV, TRANS-intra, TRANS-inter, INV, CNV
GLYMA-13G146800	PTHR10093 - IRON-SULFUR CLUSTER ASSEMBLY ENZYME NIFU HOMOLOG	PAV, TRANS-inter, INV, CNV
GLYMA-03G177400	K02930 - large subunit ribosomal protein L4e (RP-L4e, RPL4)	PAV, TRANS-intra, TRANS-inter, INV, CNV
GLYMA-07G083900	PTHR21234//PTHR21234:SF19 - PURINE NUCLEOSIDE PHOSPHORYLASE // PHOSPHORYLASE SUPERFAMILY PROTEIN	PAV, TRANS-inter, INV, CNV
GLYMA-17G119900	PTHR10361//PTHR10361:SF35 - SODIUM-BILE ACID COTRANSPORTER // SODIUM/PYRUVATE COTRANSPORTER BASS2, CHLOROPLASTIC	PAV, TRANS-intra, TRANS-inter, INV, CNV
GLYMA-12G014000	PTHR15572 - GLIOMA TUMOR SUPPRESSOR CANDIDATE REGION GENE 1	PAV, TRANS-inter, INV, CNV
GLYMA-16G206800	PTHR10634//PTHR10634:SF36 - AN1-TYPE ZINC FINGER PROTEIN // ZINC FINGER A20 AND AN1 DOMAIN-CONTAINING STRESS-ASSOCIATED PROTEIN 10-RELATED	PAV, TRANS-intra, INV, CNV
GLYMA-11G127200	PTHR10584:SF168 - F14I3.3 PROTEIN-RELATED	PAV, TRANS-inter, INV, CNV

Glyma-18G206800 tiene también una expresión muy marcada, con un Z-score de 12, en la epidermis de la semilla en estado de corazón (Tabla 2). Este gen no está asociado aún a un proceso metabólico, pero tiene variaciones estructurales como una inversión, variación en el número de copias del gen, variación de regulación génica TRANS-inter, y está presente o ausente en los genomas que constituyen el pangenoma de soya (Tabla 1).

Glyma-02G223900 es otro gen cuya expresión es alta en el tejido suspensor del cotiledón de la semilla (Tabla 2). Glyma-02G223900 es también una proteína de función desconocida. Sin embargo, este gen presenta variaciones estructurales como presencia o ausencia en los genes del pangenoma, presenta una variación de regulación génica trans-inter y una variabilidad en el número de copias del gen (Tabla 1). Muchas veces este tipo de variaciones estructurales se convierten en clave para asegurar la supervivencia de la semilla.

Así también, el gen Glyma-07G083900 tiene un Z-score alrededor de 10 y su alta expresión se da en tejidos como intra-tegumento de la semilla, corteza del cotiledón de la semilla, epidermis del cotiledón de la semilla y células divisorias del cotiledón de la semilla (Tabla 2). También tiene una expresión alta en el meristemo floral. Este gen tiene actividad de respuesta a heridas, es decir también es un gen de respuesta a estrés. Parece ser entonces, que algunas de las variaciones estructurales están situadas en genes de respuesta a la división celular y de estrés. Pero también se puede observar que las variaciones estructurales están siendo expresadas en tejidos de desarrollo como los nódulos de la raíz, las semillas y el meristemo floral (Tabla 2).

Tabla 2 Órganos de expresión de variaciones estructurales del pangenoma de G. max

Órgano	**Descripción**	**Gen**
Raíz	capa esclereida del nódulo	Glyma-11G127200
Raíz	capa esclereida del nódulo	Glyma-13G146800
Semilla	epidermis de la semilla en estado de corazón	Glyma-18G206800
Semilla	tejido suspensor del cotiledón de la semilla	Glyma-02G223900
Semilla	intra-tegumento de la semilla,	Glyma-07G083900
Semilla	corteza del cotiledón de la semilla	Glyma-07G083900
Semilla	epidermis del cotiledón de la semilla	Glyma-07G083900
Semilla	células divisorias del cotiledón de la semilla	Glyma-07G083900
Flor	Meristemo floral	Glyma-07G083900

La descripción de la actividad de los genes fue obtenida de la base de datos Monocots PLAZA (https://bioinformatics.psb.ugent.be/plaza.dev/instances/monocots_05/genes/view/Glyma.02G223900).

Por otro lado, se usó una tabla de genes con *SNPs* que fueron parte de un estudio de estrés hídrico para relacionar marchitamiento del dosel con loci que presenten *SNPs* distintivos (Chamarthi et al., 2021). De esta tabla se obtuvo una lista de 485 genes del cual se realizó un mapa de calor de 30 genes cuya expresión en *Tabula Glycine* es la más alta y la más baja en comparación al resto de genes con *SNPs*.

4.3 Expresión de genes importantes en respuesta a estrés hídrico con SNPs.

Tabla 3. Expresión de genes con SNPs bajo estrés hídrico

Gene	Description	Literature
GLYMA-07G121500	K15404 - aldehyde decarbonylase (K15404, CER1)	
GLYMA-20G054600	PF00443//PF12436//PF14533 - Ubiquitin carboxyl-terminal hydrolase (UCH) // ICP0-binding domain of Ubiquitin-specific protease 7 (USP7_ICP0_bdg) // Ubiquitin-specific protease C-terminal (USP7_C2)	
GLYMA-17G176300	PTHR19878:SF4 - TRANSDUCIN/WD40 DOMAIN-CONTAINING PROTEIN	
GLYMA-10G120900	K11838 - ubiquitin carboxyl-terminal hydrolase 7 (USP7, UBP15)	
GLYMA-08G068200	PTHR11945//PTHR11945:SF179 - MADS BOX PROTEIN // MADS-BOX PROTEIN SVP	
GLYMA-09G262000	PTHR16254:SF13 - K( ) EFFLUX ANTIPORTER 1, CHLOROPLASTIC-RELATED	
GLYMA-19G037200	PF14225 - Cell morphogenesis C-terminal (MOR2-PAG1_C)	Chamarti et al., 2021
GLYMA-06G213600	3.2.1.68 - Isoamylase / Debranching enzyme	
GLYMA-17G065300	PTHR31062//PTHR31062:SF32 - FAMILY NOT NAMED // XYLOGLUCAN ENDOTRANSGLUCOSYLASE/HYDROLASE PROTEIN 25-RELATED	
GLYMA-05G196900	PF01823 - MAC/Perforin domain (MACPF)	
GLYMA-20G071200	K11838 - ubiquitin carboxyl-terminal hydrolase 7 (USP7, UBP15)	
GLYMA-05G090500	PTHR19878:SF4 - TRANSDUCIN/WD40 DOMAIN-CONTAINING PROTEIN	
GLYMA-01G082700	KOG0670 - U4/U6-associated splicing factor PRP4	
GLYMA-20G250000	cytokinesis by cell plate formation	
GLYMA-10G118800	PTHR22597//PTHR22597:SF0 - POLYCOMB GROUP PROTEIN // POLYCOMB PROTEIN SUZ12	

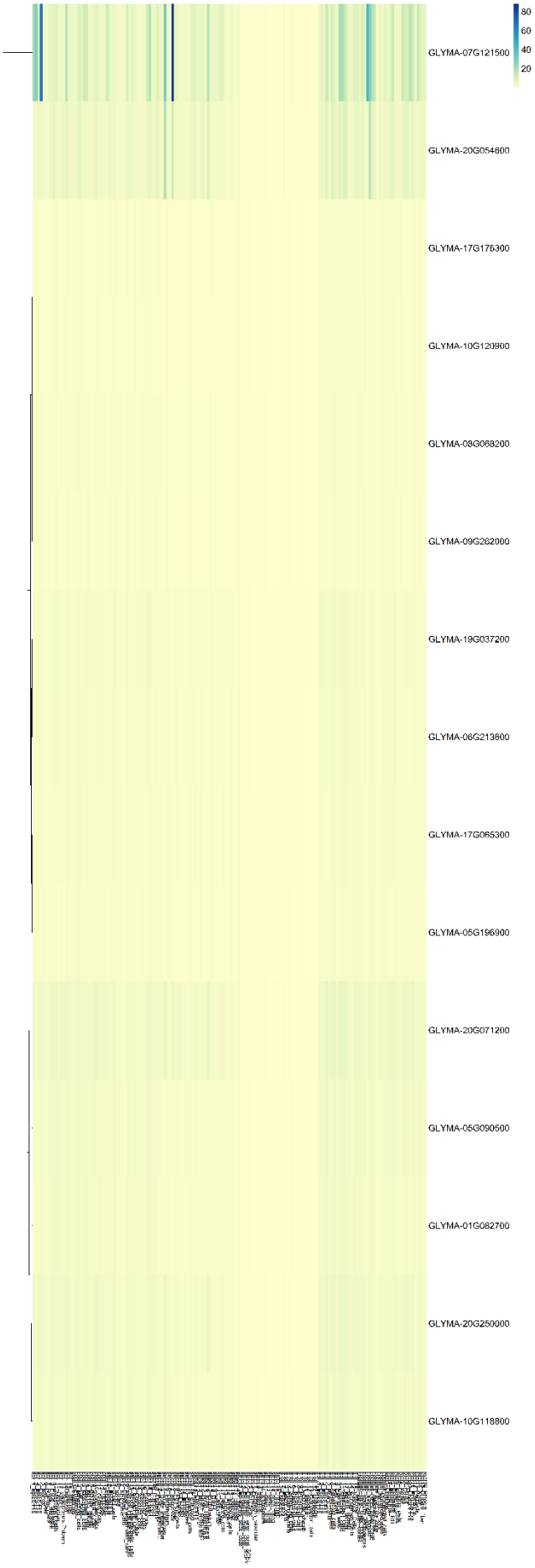

Figura 4Expresión de los genes asociados a stress en Tabula Glycine

En la Figura 4 se puede observar que la expresión génica de los valores extremos, más altos y más bajos, no es marcadamente alta. Esto puede deberse a que se están estudiando genes que tienen zonas con *SNPs* y no genes que hayan sido seleccionados debido a su expresión variante en condiciones de estrés hídrico. Sin embargo, se observa que los genes Glyma-07G121500 y Glyma-20G054600 son altamente expresados en comparación al resto, con un Z-Score de entre 40 y 80. Glyma-07G121500 es un gen responsable de respuesta a estrés hídrico, biosíntesis de cera, respuesta inmunitaria a hongos y desarrollo de anteras, polen, entre otras funciones. Esto está en línea con la expresión de este gen ampliamente en los tejidos nodulares, la hoja verdadera, semilla y cotiledón; y además su expresión es muy marcada en la flor. Se esperaría que bajo estrés hídrico la expresión de este gen sea inclusive mayor en la raíz, y nódulos (Tabla 4). Glyma-20G054600 tiene la función de desubiquitinación de proteínas. Este gen está altamente expresado en el haz vascular del tejido nodular y en el cotiledón de la semilla, en el tegumento interno. Este gen puede estar reaccionando al estrés hídrico al iniciar una respuesta en cadena a partir de la desubiquitinación de proteínas. Sin embargo, es una hipótesis que habría que investigar más adelante.

Tabla 4 Órganos de expresión de los genes con SNPs distintivos de muestras sometidas a estrés hídrico

Órgano	**Descripción**	**Gen**
Raíz	tejidos nodulares	Glyma-07G121500
Hoja	hoja verdadera	Glyma-07G121500
Semilla	semilla	Glyma-07G121500
Semilla	cotiledón	Glyma-07G121500
Flor	flor	Glyma-07G121500
Haz Vascular	haz vascular del tejido nodular	Glyma-20G054600
Semilla	cotiledón de la semilla	Glyma-20G054600
Semilla	tegumento interno	Glyma-20G054600

4.4 Expresión a nivel celular de genes involucrados en resistencia a patógenos como candidatos para el fitomejoramiento.

Finalmente, se buscó la expresión de los genes involucrados en resistencia a *P. sojae* y los genes que están involucrados en el desarrollo de la semilla y del embrión. Estos genes son de importancia productiva en la soja ya que la semilla es el producto comercializado y de uso principal en el cultivo. En la Figura 5 se puede observar que estos genes tienen una expresión a lo largo de todos los tejidos de *Tabula Glycine*. Es decir, podrían considerarse genes "*house-keeping*". Sin embargo, dentro de esta expresión generalizada, existen tejidos donde cada uno de estos genes tiene una mayor expresión que en el resto de los tejidos. Glyma-04G233300, Glyma-06G131500, Glyma-01G021100, Glyma-09G201000 y Glyma-08G195300 son genes que están involucrados en la regulación de la transcripción y codifican para la proteína con dominio DoF-Zn (Tabla 3.), la cual está envuelta en desarrollo embrionario (O'Rourke et al., 2014). Su mayor expresión se

observa en tejidos de hoja verdadera, cotiledón de semilla y meristemo apical (Tabla 4). Por otro lado, los genes de resistencia Glyma-19G626700 y Glyma-13G190400 tienen una expresión mayoritaria en los tejidos de la raíz y el tejido meristemático apical respectivamente. Glyma-19G626700 es un gen que está involucrado en funciones de transcripción y además tiene funciones de respuesta a estrés hídrico. Glyma-13G190400 es un gen de defensa ante el ataque de hongos. Glyma-13G177500 es un gen involucrado en procesos de transporte de iones y es una proteína con dominio conservado antiguo. Este gen se expresa mayoritariamente en tejidos del meristemo apical y en la hoja verdadera.

Finalmente, Glyma-06G063400 se expresa ampliamente en tejidos de raíz, nódulo, hoja verdadera, meristemo apical, vaina y flor (Tabla 4). Este gen se ve involucrado en varias rutas de desarrollo como regulación negativa de ácido abscísico, respuesta a giberelinas, auxinas, y respuesta a estrés hídrico entre otros.

Por otro lado, cabe resaltar que los tejidos de hoja trifoliada y semi desarrollo medio de células de epidermis, mesófilas, sensitoria, células guarda, relleno de semillas, cotiledón epidermis de semilla, embrión y células de ciclo celular tuvieron una expresión baja con Z-Score igual a 0 y -2 para los genes Glyma-13G190400 y Glyma-06G063400 (Tabla 5). Es interesante que Glyma-13G190400, que es un gen de resistencia para *P. sojae*, tiene una expresión muy baja en estos tejidos bajo condiciones no estresantes. Por otro lado, el gen Glyma-06G063400 juega un papel en la activación del desarrollo embrionario. En las hojas trifoliadas y en un estadio de semi desarrollo medio de los órganos de *G. max* este gen no parece estar altamente activo (Tabla 5).

Tabla 5 Órganos de expresión de genes con relación a producción de semilla y resistencia a P. sojae

Órgano	**Descripción**	**Gen**
Hoja, semilla, meristemo apical	hoja verdadera, cotiledón, semilla, meristemo apical	Glyma-04G233300
Hoja, semilla, meristemo apical	hoja verdadera, cotiledón, semilla, meristemo apical	Glyma-06G131500
Hoja, semilla, meristemo apical	hoja verdadera, cotiledón, semilla, meristemo apical	Glyma-01G021100
Hoja, semilla, meristemo apical	hoja verdadera, cotiledón, semilla, meristemo apical	Glyma-09G201000
Hoja, semilla, meristemo apical	hoja verdadera, cotiledón, semilla, meristemo apical	Glyma-08G195300
Raíz	raíz	Glyma-19G626700
Raíz, hoja	raíz, nódulo, hoja verdadera, meristemo apical, vaina y flor	Glyma-06G063400
Meristemo apical	meristemo apical	Glyma-13G190400
Hoja tejidos semi desarrollados	baja expresión tejidos semi desarrollo medio y hoja trifoliada	Glyma-13G190400
Hoja tejidos semi desarrollados	baja expresión tejidos semi desarrollo medio y hoja trifoliada	Glyma-06G063400

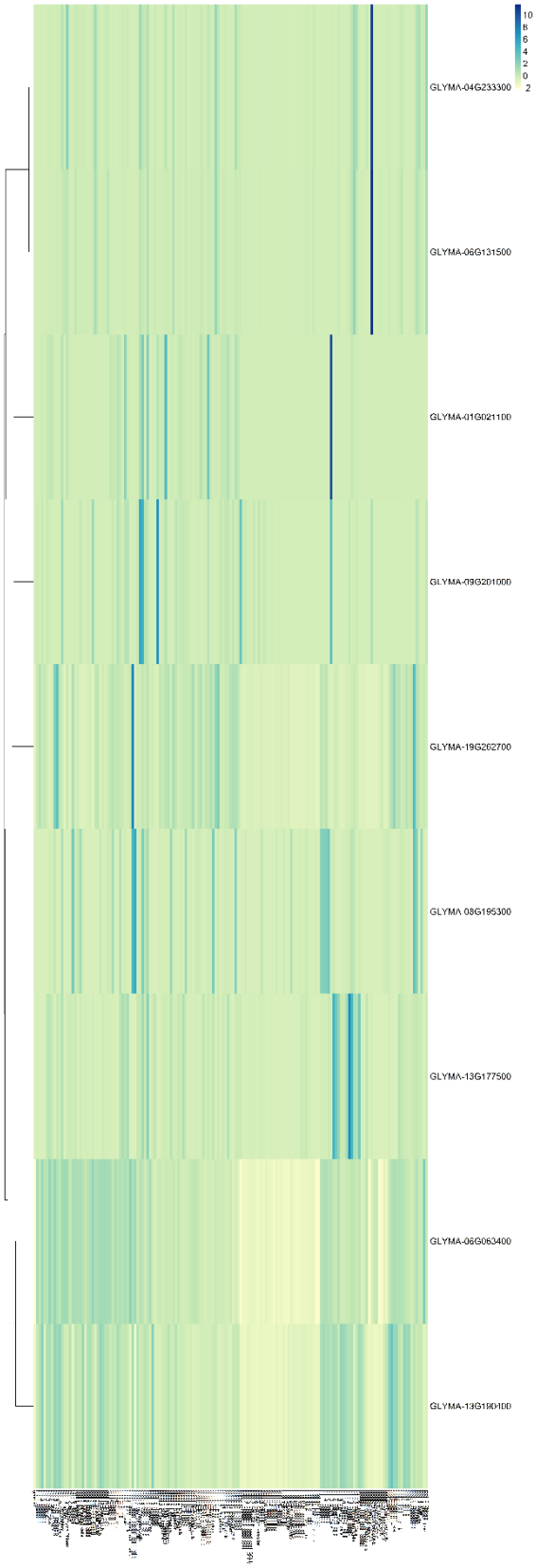

Figura 5 Expresión de los genes asociados a producción y resistencia en Tabula Glycine

Tabla 6 Production and Resistance Genes

Gene	Description	Putative Process	Literature
GLYMA-01G021100	PF02701 - Dof domain, zinc finger (zf-Dof)	Early embryo development	
GLYMA-04G233300	PF02701 - Dof domain, zinc finger (zf-Dof)	Early embryo development	
GLYMA-06G063400	K11643 - chromodomain-helicase-DNA-binding protein 4 (CHD4, MI2B)	Activator of embryo development	
GLYMA-06G131500	PF02701 - Dof domain, zinc finger (zf-Dof)	Early embryo development	
GLYMA-19G262700	PTHR31190//PTHR31190:SF4 - FAMILY NOT NAMED // ETHYLENE-RESPONSIVE TRANSCRIPTION FACTOR ERF071-RELATED	*Phytophthora sojae* resistance	O'Rourke J., et al.
GLYMA-09G201000	PF02701 - Dof domain, zinc finger (zf-Dof)	Early embryo development	
GLYMA-08G195300	PF02701 - Dof domain, zinc finger (zf-Dof)	Early embryo development	
GLYMA-13G190400	PTHR23155//PTHR23155:SF414 - LEUCINE-RICH REPEAT-CONTAINING PROTEIN // SUBFAMILY NOT NAMED	*Phytophthora sojae* resistance	
GLYMA-13G177500	PF02701 - Dof domain, zinc finger (zf-Dof)	Early embryo development	

5. ANÁLISIS DE RESULTADOS

5.1. Discusión

¿Cuál es la correlación existente entre la expresión génica de los diferentes tipos de células de *G. max* presentados en *Tabula Glycine* y el pangenoma de 26 genomas de *G. max*?

Para responder a esta pregunta se observó el mapa de calor de la expresión total de las variaciones estructurales del pangenoma de soya. Se pudo observar que pocos genes (15), tienen una expresión marcada más alta, lo que podría indicar que la mayoría tiene una expresión cuando la planta tiene algún estímulo exterior. Para poder conocer más sobre el comportamiento de los otros genes se necesitaría conocer la expresión de los genes del pangenoma completo. En este estudio no se realizó este análisis pues no se contó con una alineación de secuencias de los genomas total para compararlo a Wm82v4.

¿Cuál es el nivel de expresión de los genes con variaciones estructurales del pangenoma de 26 genomas de *G.max* a nivel de tipo de célula?

El nivel de expresión de los genes con variaciones estructurales tuvo tan poca expresión como cero y 150 de valor de Z-Score. Esto quiere decir que no todos los genes con variaciones estructurales tienen una expresión remarcada en Wm82v4. Sin embargo, se encontró que 15 genes con variaciones estructurales tienen expresión en tejidos de desarrollo y además su función está relacionada a la división celular y respuesta a estrés.

¿Cuáles son las variaciones estructurales que mayor expresión muestran a nivel de tipo de célula y cuáles de estos genes con variaciones puede influenciar el fitomejoramiento de *G. max*?

Las variaciones estructurales que mayor expresión mostraron a nivel de tipo de célula fueron todas, PAV, TRANS-intra, TRANS-inter, INV, CNV y los genes a los que pertenecen estas variaciones estructurales son Glyma-02G223900, Glyma-16G206800 que están expresadas en tejidos de epidermis de semilla, y tejido suspensor del cotiledón. Por otro lado, Glyma-13G146800, Glyma-07G083900, y Glyma-11G127200 corresponden a los tejidos de capa esclereida del nódulo en la raíz y meristemo floral. Esta información sugiere que las variaciones de este tipo están relacionadas con la producción de semilla y reproducción lo cual puede tener un peso evolutivo.

¿Cómo está configurada la expresión de los genes de resistencia a *P. sojae*, genes relacionados a la producción de *G. max*, y genes relacionados a la tolerancia de estrés hídrico en *Tabula Glycine*? ¿Cuáles son los tejidos con mayor expresión génica?

En el caso de los genes de resistencia a estrés hídrico, se encontró que Glyma-07G121500 y Glyma-20G054600 están altamente expresados en los tejidos de la raíz, y el haz vascular respectivamente, lo que va en línea con la sensibilidad de la planta a la marchitez cuando existe estrés hídrico.

En cuanto a los genes de resistencia a *P. sojae* y los genes relacionados al desarrollo de la semilla, se encontró que Glyma-19G626700 y Glyma-13G190400 tienen una expresión mayoritaria en los tejidos de la raíz y el tejido meristemático apical respectivamente. Esto sugiere que están activos en tejidos de desarrollo.

Por otro lado, el gen Glyma-13G190400, el cual es un gen de resistencia hacia hongos, tiene una muy baja expresión en la hoja trifoliada y tejidos en semi desarrollo, células guarda y células embrionarias. Esto sugiere que este gen podría estar activo en tejido en desarrollo que no tiene que ver con desarrollo embrionario.

En cuanto a los genes ligados a la producción, se encontró que Glyma-06G063400 se expresa ampliamente en tejidos de raíz, nódulo, hoja verdadera, meristemo apical, vaina y flor. Este gen se ve involucrado en varias rutas de desarrollo como regulación negativa de ácido abscísico, respuesta a giberelinas, auxinas, y respuesta a estrés hídrico entre otros. Esto sugiere que este gen podría estar activo en los tejidos de desarrollo y además en los tejidos desarrollados a término ya que es un gen de respuesta a estrés y está involucrado en varios procesos metabólicos a través de las hormonas giberelinas, y auxinas.

¿En comparación a *Tabula Glycine*, cómo es la configuración de los tejidos con mayor expresión de los genes con variaciones estructurales estudiados en el pangenoma de 26 genomas de *G. max*?

La configuración de la expresión de los genes con variaciones estructurales en Wm82v4 se ve baja en comparación a Tabula Glycine. Sin embargo, esto puede ser un artefacto de la resolución del mapa de calor de los genes con variaciones estructurales. Al ser genes repetidos y específicos su resolución es más alta que la de Tabula Glycine y muestra baja expresión. Al contrario, Tabula Glycine muestra la expresión de todos los genes de Wm82v4 y su resolución en este caso es más baja ya que tiene mayor cantidad de información. Lo que se puede apreciar es que 15 genes con variaciones estructurales tienen expresión marcada en tejidos de raíz, semilla y flor.

Finalmente, los mapas de calor observados ayudan a entender cómo los genes estudiados se encuentran expresados sin ningún estímulo externo que a propósito afecte su equilibrio. En primera instancia, al analizar el comportamiento de expresión de las variaciones estructurales en Wm82v4 se pudo observar que varias de las variaciones estructurales se encuentran codificando genes involucrados en división celular y respuesta a estrés. Además, se pudo observar que tejidos como nódulos de la raíz, semilla y meristemo floral son afectadas por las variaciones estructurales lo que hace pensar que tendrían un peso evolutivo en *G. max*. También se puede concluir que las variaciones estructurales tienen una alta expresión en la raíz y semillas lo que es de especial importancia si se quisiera usar una variedad del pangenoma para mejorar la productividad (semilla) o el vigor de la planta (raíz).

En segunda instancia, al analizar genes con presencia de *SNPs* correspondiente a plantas sometidas a estrés hídrico se pudo determinar que la diferencia entre expresión génica fue mínima. Muy probablemente esto se debe a que estos genes no son genes cuya expresión cambió al estar sometidos a estrés hídrico, pero son genes que mostraron un genotipo diferente. Sin embargo, se encontró que Glyma-07G121500 y Glyma-20G054600 fueron genes altamente expresados en tejidos nodulares, raíz, semilla y flor. Cabe recalcar que la función de estos genes es de respuesta a estrés hídrico y desarrollo de polen y anteras por lo que son de importancia a la hora de medir el estrés hídrico como un indicador de plantas sensibles o no con fines de fitomejoramiento.

Finalmente, al analizar los genes posiblemente involucrados en producción, se pudo observar que su expresión es generalizada. Sin embargo, los genes involucrados en desarrollo embrionario como Glyma-09G201000 y Glyma-08G195300 (O'Rourke et al., 2014) se vieron expresados en tejidos de hoja verdadera, meristemo apical y semilla. Esto es consecuente con su función putativa de desarrollo y relleno de semilla. Además, los genes involucrados en resistencia a *P. sojae* también fueron altamente expresados en tejidos de meristema apical, raíz y hoja verdadera. Eso quiere decir que estos genes están

ampliamente expresados quizás a la espera de un ataque patogénico para entrar en acción. Sin embargo, los tejidos en semi desarrollo medio y la hoja trifoliada tienen menor expresión, lo cual puede significar que en estos estadios la planta no usa energía en protección si no existe un ataque.

Más información e imágenes de alta resolución se pueden encontrar siguiendo el siguiente enlace de material suplementario:

Material suplementario

6. CONCLUSIONES Y RECOMENDACIONES

6.1. Conclusiones

Tabula Glycine es una herramienta robusta que permite determinar el estado de expresión general de los órganos de *G. max*. En este estudió sirvió como guía para entender como están siendo expresados los genes con variaciones estructurales pertenecientes al pangenoma de 26 genomas de soya. Además, se pudo determinar cómo es la expresión de los genes involucrados en desarrollo embrionario y por ende en producción. Se pudo también conocer como los genes con *SNPs* pertenecientes a líneas de *G.max* Wm82v4 que fueron sometidas a estrés hídrico se comportan a nivel de expresión génica.

La información recopilada y presentada en este trabajo es un inicio para desarrollar hipótesis y plantear nuevos experimentos que permitan conocer que genes de acuerdo al tejido podrían ser los más importantes a la hora de fitomejorar la soya para obtener mejor producción, resistencia patogénica y resistencia al estrés hídrico. En una era en la que el planeta sufre cambios climáticos abruptos, es trascendental continuar con este estudio.

6.2. Recomendaciones

Se recomienda repetir este estudio con más genes que sean considerados importantes para la producción de *G. max*. Es importante llenar el vacío del conocimiento sobre los genes que estén relacionados a producción para poder esbozar planes de fitomejoramiento que permitan crear variedades de mayor producción.

También es necesario combinar los estudios genómicos computacionales con estudios de fenotipado por medio de máquinas que sigan algoritmos para poder claramente discernir entre mejor o peor producción, susceptibilidad a enfermedades y estrés hídrico.

La información adquirida aquí puede servir para planificar planes de fitomejoramiento en los que se hagan barridos de detección de los genes con variaciones estructurarles que afectan el desarrollo de las semillas. Además, los genes que mostraron actividad de respuesta al estrés hídrico, así como los genes de resistencia a *P. sojae* pueden ser usados para hacer otro barrido de detección de líneas que presenten estos genes expresados o no.

Combinar los datos con observaciones del fenotipo puede ayudar a direccionar el proceso de fitomejoramiento para una mayor productividad, y mayor resistencia.

7. REFERENCIAS

Bradshaw, J. E. (2017). Plant breeding: past, present and future. *Euphytica*, *213*(3), 1–12. https://doi.org/10.1007/S10681-016-1815-Y/METRICS

Cao, P., Zhao, Y., Wu, F., Xin, D., Liu, C., Wu, X., Lv, J., Chen, Q., & Qi, Z. (2022). Multi-Omics Techniques for Soybean Molecular Breeding. *International Journal of Molecular Sciences*, *23*(9). https://doi.org/10.3390/IJMS23094994

Caudai, C., Galizia, A., Geraci, F., Le Pera, L., Morea, V., Salerno, E., Via, A., & Colombo, T. (2021). AI applications in functional genomics. *Computational and Structural Biotechnology Journal*, *19*, 5762–5790. https://doi.org/10.1016/J.CSBJ.2021.10.009

Cervantes-Pérez, S. A., Thibivilliers, S., Amini, S., Pelletier, J. M., Meyer, I., Xu, H., Tennant, S., Ma, P., Sprueill, C. M., Farmer, A. D., Coate, J. E., Nelissen, H., Yao, Q., Martin, O. C., Amézquita, E. J., Goldberg, R. B., Harada, J. J., & Libault, M. (2024). Tabula Glycine: The whole-soybean single-cell resolution transcriptome atlas. *BioRxiv*, 2024.07.08.602332. https://doi.org/10.1101/2024.07.08.602332

Cervantes-Pérez, S. A., Zogli, P., Amini, S., Thibivilliers, S., Tennant, S., Hossain, M. S., Xu, H., Meyer, I., Nooka, A., Ma, P., Yao, Q., Naldrett, M. J., Farmer, A., Martin, O., Bhattacharya, S., Kläver, J., & Libault, M. (2024). Single-cell transcriptome atlases of soybean root and mature nodule reveal new regulatory programs that control the nodulation process. *Plant Communications*, *5*(8), 100984. https://doi.org/10.1016/J.XPLC.2024.100984/ATTACHMENT/A7AAF4FD-A49F-47A3-8A3E-F4A647DC7958/MMC17.PDF

Chamarthi, S. K., Kaler, A. S., Abdel-Haleem, H., Fritschi, F. B., Gillman, J. D., Ray, J. D., Smith, J. R., Dhanapal, A. P., King, C. A., & Purcell, L. C. (2021). Identification and Confirmation of Loci Associated With Canopy Wilting in Soybean Using Genome-Wide Association Mapping. *Frontiers in Plant Science*, *12*, 698116. https://doi.org/10.3389/FPLS.2021.698116/BIBTEX

de Ronne, M., Labbé, C., Lebreton, A., Sonah, H., Deshmukh, R., Jean, M., Belzile, F., O'Donoughue, L., & Bélanger, R. (2020). Integrated QTL mapping, gene expression and nucleotide variation analyses to investigate complex quantitative traits: a case study with the soybean–Phytophthora sojae interaction. *Plant Biotechnology Journal*, *18*(7), 1492. https://doi.org/10.1111/PBI.13301

Ecuador Production. (n.d.). Retrieved April 27, 2025, from https://ipad.fas.usda.gov/countrysummary/Default.aspx?id=EC

Fehr, W. R. (2007). Breeding for Modified Fatty Acid Composition in Soybean. *Crop Science*, *47*(SUPPL. DEC.), S-72. https://doi.org/10.2135/CROPSCI2007.04.0004IPBS

Hymowitz, T., & Shurtleff, W. R. (2005). Debunking Soybean Myths and Legends in the Historical and Popular Literature. *Crop Science*, *45*(2), 473–476. https://doi.org/10.2135/CROPSCI2005.0473

Li, W. X., Wang, P., Zhao, H., Sun, X., Yang, T., Li, H., Hou, Y., Liu, C., Siyal, M., Raja veesar, R., Hu, B., & Ning, H. (2021). QTL for Main Stem Node Number and Its Response to Plant Densities in 144 Soybean FW-RILs. *Frontiers in Plant Science*, *12*. https://doi.org/10.3389/FPLS.2021.666796/FULL

Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., Zhou, G. A., Zhang, H., Liu, Z., Shi, M., Huang, X., Li, Y., Zhang, M., Wang, Z., Zhu, B., Han, B., Liang, C., & Tian, Z. (2020). Pan-Genome of Wild and Cultivated Soybeans. *Cell*, *182*(1), 162-176.e13. https://doi.org/10.1016/J.CELL.2020.05.023

Liu, Z., Kong, X., Long, Y., Liu, S., Zhang, H., Jia, J., Cui, W., Zhang, Z., Song, X., Qiu, L., Zhai, J., & Yan, Z. (2023). Integrated single-nucleus and spatial transcriptomics captures transitional states in soybean nodule maturation. *Nature Plants*, *9*(4), 515–524. https://doi.org/10.1038/S41477-023-01387-Z

Meyer, R. S., Duval, A. E., & Jensen, H. R. (2012). Patterns and processes in crop domestication: an historical review and quantitative analysis of 203 global food crops. *The New Phytologist*, *196*(1), 29–48. https://doi.org/10.1111/J.1469-8137.2012.04253.X

Mosquna, A., Katz, A., Shochat, S., Grafi, G., & Ohad, N. (2004). Interaction of FIE, a Polycomb protein, with pRb: A possible mechanism regulating endosperm development. *Molecular Genetics and Genomics*, *271*(6), 651–657. https://doi.org/10.1007/S00438-004-1024-6/METRICS

Ni, L., Liu, Y., Ma, X., Liu, T., Yang, X., Wang, Z., Liang, Q., Liu, S., Zhang, M., Wang, Z., Shen, Y., & Tian, Z. (2023). Pan-3D genome analysis reveals structural and functional differentiation of soybean genomes. *Genome Biology*, *24*(1), 1–26. https://doi.org/10.1186/S13059-023-02854-8/FIGURES/6

O'Rourke, J. A., Bolon, Y. T., Bucciarelli, B., & Vance, C. P. (2014). Legume genomics: understanding biology through DNA and RNA sequencing. *Annals of Botany*, *113*(7), 1107–1120. https://doi.org/10.1093/AOB/MCU072

Pelletier, J. M., Kwong, R. W., Park, S., Le, B. H., Baden, R., Cagliari, A., Hashimoto, M., Munoz, M. D., Fischer, R. L., Goldberg, R. B., & Harada, J. J. (2017). LEC1 sequentially regulates the transcription of genes involved in diverse developmental processes during seed development. *Proceedings of the*

National Academy of Sciences of the United States of America, *114*(32), E6710–E6719. https://doi.org/10.1073/PNAS.1707957114/SUPPL_FILE/PNAS.1707957114.SD06.XLSX

Rayaprolu, S. J., Hettiarachchy, N. S., Chen, P., Kannan, A., & Mauromostakos, A. (2013). Peptides derived from high oleic acid soybean meals inhibit colon, liver and lung cancer cell growth. *Food Research International*, *50*(1), 282–288. https://doi.org/10.1016/J.FOODRES.2012.10.021

Robertson, A. E., Cianzio, S. R., Cerra, S. M., & Pope, R. O. (2018). Within-field Pathogenic Diversity of Phytophthora sojae in Commercial Soybean Fields in Iowa. *Https://Doi.Org/10.1094/PHP-2009-0908-01-RS*, *10*(1). https://doi.org/10.1094/PHP-2009-0908-01-RS

Shen, B., Sinkevicius, K. W., Selinger, D. A., & Tarczynski, M. C. (2006). The homeobox gene GLABRA2 affects seed oil content in Arabidopsis. *Plant Molecular Biology*, *60*(3), 377–387. https://doi.org/10.1007/S11103-005-4110-1/METRICS

Shen, Y., Liu, J., Geng, H., Zhang, J., Liu, Y., Zhang, H., Xing, S., Du, J., Ma, S., & Tian, Z. (2018). De novo assembly of a Chinese soybean genome. *Science China Life Sciences 2018 61:8*, *61*(8), 871–884. https://doi.org/10.1007/S11427-018-9360-0

Soybeans | USDA Foreign Agricultural Service. (n.d.). Retrieved April 27, 2025, from https://www.fas.usda.gov/data/production/commodity/2222000

Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Bassett, D. E., Hieter, P., Vogelstein, B., & Kinzler, K. W. (1997). Characterization of the yeast transcriptome. *Cell*, *88*(2), 243–251. https://doi.org/10.1016/S0092-8674(00)81845-0

Wang, J., Hu, B., Jing, Y., Hu, X., Guo, Y., Chen, J., Liu, Y., Hao, J., Li, W. X., & Ning, H. (2022). Detecting QTL and Candidate Genes for Plant Height in Soybean via Linkage Analysis and GWAS. *Frontiers in Plant Science*, *12*, 803820. https://doi.org/10.3389/FPLS.2021.803820/BIBTEX

What are Unique Molecular Identifiers and Why do We Need Them? (n.d.). Retrieved April 26, 2025, from https://www.lexogen.com/rna-lexicon-what-are-unique-molecular-identifiers-umis-and-why-do-we-need-them/

Yang, Z., Luo, C., Pei, X., Wang, S., Huang, Y., Li, J., Liu, B., Kong, F., Yang, Q. Y., & Fang, C. (2024). SoyMD: a platform combining multi-omics data with various tools for soybean research and breeding. *Nucleic Acids Research*, *52*(D1), D1639–D1650. https://doi.org/10.1093/NAR/GKAD786

8. ANEXOS

1. Anexo 1.

```
# -*- coding: utf-8 -*-
"""
Created on Sun Feb  2 09:46:38 2025
@author: Mélida Rojas
"""
import csv
import os
import requests
from intermine.webservice import Service

# Set file paths
genes_file = r'C:\Users\LT816\OneDrive - Valleflor\Documentos\Generales\Tesis\Genes
variation and expression\genes_extracted.csv'  # Path to your genes_extracted file
output_file = r'C:\Users\LT816\OneDrive -
Valleflor\Documentos\Generales\Tesis\Genes variation and
expression\gene_query_results3.csv'  # Output file for the results

# Check if the file exists
if not os.path.exists(genes_file):
    print(f"File not found: {genes_file}")
    exit(1)

# Create a session
session = requests.Session()

# Disable SSL verification temporarily (you can remove this if SSL is not an issue)
session.verify = False  # Disabling SSL verification for troubleshooting

# Initialize the service
service = Service("https://mines.legumeinfo.org/legumemine/service")

# Use the session for making requests
service._session = session  # Manually assign the session to the service

# Open and read the CSV file, skipping the first line (header)
with open(genes_file, 'r', newline='', encoding='utf-8') as f:
    reader = csv.reader(f)

    # Skip the first line (header)
    next(reader)

    # Read the gene names and store them in a set (to avoid duplicates)
    gene_identifiers = {row[0] for row in reader}  # Now it's a list, no set needed

# Function to execute the query for each gene individually
```

```
def run_query_gene(gene, out_file):
    query = service.new_query("Gene")
    query.add_view(
        "transcripts.primaryIdentifier",
        "transcripts.sequence.residues"
    )
    query.add_sort_order("Gene.description", "ASC")

    # Add constraint for the current gene
    print(f"Adding constraint for gene: {gene}")  # Debugging: Print the gene being
processed
    query.add_constraint("secondaryIdentifier", "=", gene)  # Querying one gene at a time

    # Open the output file for writing
    with open(out_file, 'a', newline='', encoding='utf-8') as out_fh:
        out_csv = csv.writer(out_fh)

        rows = list(query.rows())  # Execute the query and retrieve rows as a list
        print(f"Number of rows returned for {gene}: {len(rows)}")  # Debugging print:
Check how many rows returned for this gene

        # Write the data to the file for this gene
        if rows:
            for row in rows:
                out_csv.writerow([row["transcripts.primaryIdentifier"],
                                  row["transcripts.sequence.residues"]])
        else:
            print(f"No rows returned for gene: {gene}")

# Process each gene individually
for gene in gene_identifiers:
    run_query_gene(gene, output_file)
    print(f"Processed gene: {gene}")
```

2. Anexo 2.

awk command line para extraer secuencias que tengan .gnm2, que se refiere a genoma assembly 2 de soya.
~$ awk '/glyma\.Zh13\.gnm2\./, /glyma\./ {print $0}' /mnt/c/Users/LT816/OneDrive\ -\ Valleflor/Documentos/Generales/Tesis/Genes\ variation\ and\ expression/gene_query_results3.csv > output_file.csv

3. Anexo 3.

Bash script para transformar el output_file en un archivo fasta para correr un blast (fastagenes1.sh). Se crea farmatted_sequenes.fasta pero más tarde se cambia nombre a formatted_sequences_gmn2.fasta

```
#!/bin/bash

# Check if the user has provided an input file as an argument
if [ -z "$1" ]; then
    echo "Usage: $0 <input_file>"
    exit 1
fi

# Get the input file path from the command line argument
input_file="$1"
output_file="formatted_sequences.fasta"

# Using AWK to format the sequences into FASTA format
awk 'BEGIN {
    # Initialize variables
    sequence_line = "";
}
{
    # Check if the line contains "glyma" (sequence header)
    if ($0 ~ /glyma/) {
        # If we have accumulated a sequence, print it out first
        if (length(sequence_line) > 0) {
            # Print the previous sequence in 80-character chunks
            while (length(sequence_line) > 80) {
                print substr(sequence_line, 1, 80);
                sequence_line = substr(sequence_line, 81);
            }
            print sequence_line;  # Print the remaining part of the sequence
        }

        # Print the sequence header with ">"
        gsub(",", "\n", $0);  # Replace commas with line breaks in the header
        print ">" $0;  # Print the header with ">" at the beginning

        # Reset sequence
        sequence_line = "";
    } else {
```

```
        # Concatenate sequence parts to the current sequence (sequence line)
        sequence_line = sequence_line $0;
    }
}
END {
    # Ensure to print the last sequence if there is one
    if (length(sequence_line) > 0) {
        # Print the last sequence in 80-character chunks
        while (length(sequence_line) > 80) {
            print substr(sequence_line, 1, 80);
            sequence_line = substr(sequence_line, 81);
        }
        print sequence_line;  # Print the remaining part of the sequence
    }
}
' "$input_file" > "$output_file"

# Print a message indicating completion
echo "Formatted sequences have been written to $output_file"
```

4. Anexo 4.

Instalación de Blastn en WSL

Installar blast 2 de NCBI. Se tomó blast2 de:
https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/
~$sudo apt update
~$wget https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/ncbi-blast-2.16.0+-x64-linux.tar.gz
~$tar -xvzf ncbi-blast-2.16.0+-x64-linux.tar.gz
~$echo 'export PATH=$PATH:~/ncbi-blast-2.16.0+/bin' >> ~/.bashrc
source ~/.bashrc
~$sudo apt update
~$sudo apt install libgomp1
~$blastn -version

~$blastn: 2.16.0+
Package: blast 2.16.0, build Jun 25 2024 08:58:03

5. Anexo 5.

~$wget --trust-server-names --content-disposition https://api.ncbi.nlm.nih.gov/datasets/v2/genome/accession/GCF_000004515.6/download?include_annotation_type=GENOME_FASTA&include_annotation_type=GENOME_GFF&include_annotation_type=RNA_FASTA&include_annotation_type=CDS_FASTA&include_annotation_type=PROT_FASTA&include_annotation_type=SEQUENCE_REPORT&hydrated=FULLY_HYDRATED

Convertir genoma de referencia en db para blast:
~$makeblastdb -in GCF_000004515.6_Glycine_max_v4.0_genomic.fna -dbtype nucl -out Wm82v4_db

6. Anexo 6.

Blast:
blastn command line:
~$ blastn -query formatted_sequences_prueba2.fasta -
db Wm82v4_db -out blast_results2.txt -outfmt 6 -evalue 1e-10 -num_threads 6

7. Anexo 7.

Configuración script para obtener nombres de genes correspondientes a la subject sequence id después de haber corrido Blast. Se usan dos archivos pues se dividió en dos el archivo blast para trabajar más eficientemente.

~$ awk '{print $2"\t"$9"\t"$10}' blast_results2.txt > sequence_ids2.txt

```
~$ awk '{print $2"\t"$9"\t"$10}' blast_results1.txt > sequence_ids1.txt
```

8. Anexo 8.

Remoción de genes duplicados

```
~$ awk '!seen[$0]++' input.txt > unique_sequences.txt
```

9. Anexo 9.

Bash script para recuperar los genes homólogos en Wm82v4 usando los nombres de los sequences id obtenidos del Blast y el archivo genomic.gff d Wm82v4

```
~$ ./fetchgene.sh sequence_ids2.txt
/home/mel/Wm82v4/ncbi_dataset/data/GCF_000004515.6/genomic.gff
```

```
#!/bin/bash

# Check if the user provided both the sequence IDs file and the GFF file as arguments
if [ -z "$1" ] || [ -z "$2" ]; then
    echo "Usage: $0 <unique_ids_file> <gff_file>"
    exit 1
fi

# The input files containing sequence IDs and the GFF data (provided as arguments)
unique_ids_file="$1"
gff_file="$2"

# Output file where all gene names and sequence IDs will be saved
output_file="gene_sequence_ids1.txt"

# Clear the output file if it exists
> "$output_file"

# Loop through each line in the unique_ids_file
while IFS=$'\t' read -r seq_id start_pos end_pos; do
    echo "Processing sequence ID: $seq_id, Start: $start_pos, End: $end_pos"  #
Debugging: show which seq_id and range are being processed

    # To track already added gene names for the current sequence ID
    declare -A gene_names

    # Search for the sequence ID in the GFF file and process the matching lines
    grep "$seq_id" "$gff_file" | while IFS= read -r line; do
        # Extract the feature type (column 3) and position range (start: column 4, end:
column 5)
        feature_type=$(echo "$line" | cut -f 3)
        gene_start=$(echo "$line" | cut -f 4)
        gene_end=$(echo "$line" | cut -f 5)

        # Only proceed if the feature is a 'gene' and it falls within the specified range
```

```
      if [ "$feature_type" == "gene" ] && [ "$gene_start" -le "$end_pos" ] && [
"$gene_end" -ge "$start_pos" ]; then
         # Extract the 9th column from the GFF line (this contains the attributes)
         attributes=$(echo "$line" | cut -f 9)

         # Extract the gene name from the 'gene' attribute in the 9th column
         gene_name=$(echo "$attributes" | grep -oP 'gene=\K[^;]+')

         # If a gene name is found and it hasn't been added before, add it to the output file
         if [ -n "$gene_name" ] && [ -z "${gene_names[$gene_name]}" ]; then
            echo "$seq_id $gene_name" >> "$output_file"
            gene_names["$gene_name"]=1  # Mark the gene name as added
            echo "Added to output file: $seq_id $gene_name"  # Debugging
         fi
      fi
   done
done < "$unique_ids_file"  # Read sequence IDs and positions from the input file

echo "Gene names and sequence IDs have been saved to $output_file"
```

10. Anexo 10.

Se crea el script fetchlocustag2.py para extraer el locus tag desde NCBI con el archivo gene_sequence_ids1 y gene_sequence_ids2 para correr el script, en el bash se debe usar la siguiene linea de código:

python3 fetchlocustag2.py gene_sequence_ids1.txt locustag1.txt

```
# -*- coding: utf-8 -*-
"""
Created on Sun Feb  2 09:46:38 2025
@author: Mélida Rojas
"""

import sys
import os
import requests
from xml.etree import ElementTree

# Set up NCBI URL for eSummary (to fetch gene information)
NCBI_ESummary_URL = "https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esummary.fcgi"

def fetch_locus_tag(gene_id):
    """Fetches Locus tag from NCBI for a given gene ID (numeric part of
LOCXXXXX)"""
    try:
        # Extract numeric part from gene_id (remove "LOC")
        gene_id_numeric = gene_id.replace("LOC", "")

        # Fetch gene summary using the numeric gene_id directly from NCBI
        summary_params = {
            'db': 'gene',
            'id': gene_id_numeric,
            'retmode': 'xml',
        }
        summary_response = requests.get(NCBI_ESummary_URL,
params=summary_params)

        # Check for valid response
        if summary_response.status_code != 200:
            print(f"Error fetching summary data for {gene_id}: HTTP
{summary_response.status_code}")
            return None

        # Parse the summary XML to extract the Locus tag
        summary_tree = ElementTree.fromstring(summary_response.content)

        # Search for the <OtherAliases> tag that contains the Locus Tag
        locus_tag = summary_tree.find(".//OtherAliases")
```

```
        if locus_tag is not None:
            return locus_tag.text

        # If no Locus Tag found, return None
        return None

    except Exception as e:
        print(f"Error fetching data for {gene_id}: {e}")
        return None

def main(input_file, output_file):
    with open(input_file, "r") as infile, open(output_file, "w") as outfile:
        # Read each line of the input file
        for line in infile:
            # Split each line into two columns: sequence_id and gene_id
            columns = line.strip().split()
            if len(columns) == 2:
                sequence_id, gene_id = columns
                # Fetch Locus tag for the given gene_id
                locus_tag = fetch_locus_tag(gene_id)

                # Write the original sequence_id, gene_id, and the Locus tag to the output file
                if locus_tag:
                    outfile.write(f"{sequence_id}\t{gene_id}\t{locus_tag}\n")
                else:
                    outfile.write(f"{sequence_id}\t{gene_id}\tNo Locus Tag Found\n")
            else:
                print(f'Skipping invalid line: {line.strip()}")

    print(f'Output written to {output_file}")

if __name__ == "__main__":
    # Check command-line arguments
    if len(sys.argv) != 3:
        print("Usage: python fetch_locus_tags.py <input_file> <output_file>")
        sys.exit(1)

    input_file = sys.argv[1]
    output_file = sys.argv[2]

    # Ensure input file exists
    if not os.path.isfile(input_file):
        print(f"Error: Input file {input_file} does not exist.")
        sys.exit(1)

    main(input_file, output_file)
```

11. Anexo 11.

Código para cambiar “_” por “-“ in locustag/aliases name of genes:
~$ awk '{gsub(/_/, "-"); print}' locustag1.txt > locustag1c.txt

12. Anexo 12.

Se elimina duplicados
~$ awk '!seen[$1, $2]++' locuag1c.txt > locustag1cnd.txt

13. Anexo 13.

Con la siguiente mitad del archivo de sequence_ids2 tomadas del genoma se elimina duplicados con el código awk:
~$ awk '!seen[$1, $2]++' gene_sequence_ids2 > geneseqid2nod

14. Anexo 14.

Con este archivo se ejecuta fetchlocusta2.py para extraer locus tag/Aliases desde NCBI
~$ python3 fetchlocustag2.py geneseqid2nod.txt locustag2nod.txt

15. Anexo 15.

Se cambia el “ _” por “-“ en el archivo usando el siguiente código:

~$ awk '{gsub(/_/, "-"); print}' locustag2nod.txt > locustag2nodc.txt

16. Anexo 16.

Cuando la tercera columna del archivo tiene "No Locus Tag Found" se copia lo respectivo en la segunda columna. Es decir se copia el nombre del gen

~$ awk -F'\t' '$3 == "No Locus Tag Found" { $3 = $2 } { print }' cleaned_ltnodc2.txt > locust
agTnodc3.txt

17. Anexo 17.

Se juntó blast_results1 and blast_results2 in one file
Extract fifth field after 4th dot in first column of blast results complete (Blast_results_gmn1) and added to end of file blast_resultsTLogT

~$ awk -F'.' '{print $0, $(5)}' blast_resultsT_gmn2.txt > blast_resultsTLogT.txt

Eliminar v4 de locustagTnodc3.txt

```
$ awk '{$3 = gensub(/v4$/, "", 1, $3); print}' locustagTnodc3.txt > locustagTnodc4.txt
```

18. Anexo 18.

Se usó el archivo blast_resultsTLogT para cambiar los nombres de sequences ids por los nombres de los genes o Locu Tag de Wm82v4
~$./exchseq.sh blast_resultsTLogT.txt locustagTnodc4.txt blast_resultsT_ZHWM.txt

```
#!/bin/bash

# Usage: ./exchseq.sh file1.txt file2.txt output_file.txt

file1="$1"
file2="$2"
output_file="$3"

# The AWK command to match and replace columns
awk '
NR==FNR {
   # Read file2 (the second file) and store column 1 as key and column 3 as value in the array
   arr[$1] = $3
   next
}
{
   # If the value of column 2 from file1 exists as a key in the array, replace column 2 with the corresponding value from file2
   if ($2 in arr) {
      $2 = arr[$2]
   }
   # Print the modified line
   print
}
' "$file2" "$file1" > "$output_file"
```

19. Anexo 19.

Usar el script coorSHWM.sh para combiner columna 2 de blast_resultsT_ZHWM.txt y archive de variaciones estructurales del pangenoma alineado al genoma de Zh 13.

$./coorSHWM.sh blast_resultsT_ZHWM.txt SVZH13.txt ZVZHWM.txt
Para esto primero se copió la tabla original (1-s2.0-S0092867420306188-mmc7) en un archivo .txt y le dio el nombre de SVZH13.txt

```
#!/bin/bash

# Usage: ./extract_matching.sh file1.txt file2.txt output_file.txt

file1="$1"
file2="$2"
output_file="$3"

# Add headings from file2 + new headings for the added columns
# This assumes file2 has column headers in the first line
head -n 1 "$file2" | awk '{print $0 "\tWM82v4\tInicial Coordinate\tFinal Coordinate"}'
> "$output_file"

# AWK command to process and append data
awk '
NR==FNR {
   # Create an associative array with column 3 as key and columns 9 and 10 as values
from file1
   arr[$13] = $2 "\t" $9 "\t" $10  # Use tabs instead of commas
   next
}
NR>1 {
   # If column 3 of file2 matches column 13 of file1, append columns 2, 9, and 10 of
file1 to the current line of file2
   if ($3 in arr) {
      print $0 "\t" arr[$3]  # Append using tab separator
   } else {
      # If no match, print the original line from file2
      print $0
   }
}
' "$file1" "$file2" >> "$output_file"
```

20. Anexo 20.

Se borraron los títulos duplicados manualmente.
Se modificó el formato de ZVZHWM.txt para que el alias o Locus Tag o nombre del gen este separado solamente por tab y no por salto de línea y tab:

$./ sed 's/\r\tGLYMA/\tGLYMA/g' input_file.txt > output_file.txt
GLYMA column in the correct position.

$./ sed 's/\r\tLOC/\tLOC/g' ZVZHWM2.txt > ZVZHWM3.txt
LOC names en la posición correcta

~$ sed 's/\r\tNFR/\tNFR/g' ZVZHWM3.txt > ZVZHWM4.txt
Otros nombres además de Glyma y LOC
~$ sed 's/\r\tNW/\tNW/g' ZVZHWM4.txt > ZVZHWM5.txt
Otros nombres además de Glyma y LOC

21. Anexo 21.

Para extraer de *Tabula Glycine* solo los valores de interés, primero se tomó las 17 secuencias id que no tuvieron Locus Tag/alias y se buscó el nombre del gen asignado en caso de existir. Además, para el gen NFR5B se buscó el nombre del Locus Tag/alias y se lo reemplazó. Se usó el archivo SVZHWM5_noduplicates2.csv. Una vez con estos valores se usó el script de R:
Extracted_SV_ZHWM82 para extraer los genes respectivos de *Tabula Glycine*:

Se obtuvieron 15 genes
Se hizo un heat map de los 15 genes con R
Script para extraer los genes que empatan la lista de *Tabula Glycine*

Extracted_SV_ZHWM82_3
title: "R Notebook"
output: html_notebook
author: Mélida Rojas

```{r}
setwd("C:/Users/LT816/OneDrive - Valleflor/Documentos/Generales/Tesis/Genes variation and expression")
```

Leer archivo

```{r}

# Leer todo el archivo sin encabezados
wm82_raw <- read.csv("exp_atlas.csv", header = FALSE, stringsAsFactors = FALSE)

# Verifica visualmente para identificar bien qué contiene cada fila/columna
```

```
head(wm82_raw, 5)

# Asignar la segunda fila como encabezados de columna
colnames(wm82_raw) <- wm82_raw[2, ]

# Quitar las primeras dos filas (ya usamos la segunda como encabezado)
wm82_clean <- wm82_raw[-c(1,2), ]

# Asignar la segunda columna como rownames
rownames(wm82_clean) <- wm82_clean[, 2]

# Opcional: eliminar la columna 2 ya que ahora es parte de los rownames
wm82_clean <- wm82_clean[, -2]

# Resultado final
head(wm82_clean)

svZHWM82 <- read.csv("SVZHWM5_noduplicates2.csv", header = TRUE, sep = ";")


```

Matching databases

```{r}

# Match column1 from database1 to column6 from database2
matching_rows3 <- wm82_clean[wm82_clean$gene_id %in%
svZHWM82$WM82v4_NG, ]

# Save matching_rows to a CSV file
write.csv(matching_rows3, "matching_rows3.csv", row.names = TRUE)


# Print the matching rows from database1
print(matching_rows3)
```

Script para elaborar mapa de calor

22. Anexo 22.

Script para elaborar mapa de calor

````
---
title: "Gene expression SHWM"
output: pdf_notebook
author: Mélida Rojas
---

Libraries

```{r}
library(pheatmap)
library(RColorBrewer)
```

Establecer carpeta de trabajo

```{r}
setwd("C:/Users/LT816/OneDrive - Valleflor/Documentos/Generales/Tesis/Genes variation and expression")
```

Leer el archivo

```{r}

sv<- read.csv("matching_rows4.csv", header = TRUE)
# Extraer data sín títulos, (Empieza desde columna 3 y fila 3)
sv_numeric <- sv[1:nrow(sv), 2:ncol(sv)]


#Convert all columns to numeric
````
````

```
sv_numeric[] <- lapply(sv_numeric, as.numeric)
sv_numeric[is.na(sv_numeric)] <- 0  # o cualquier otro valor
# Check if all columns are numeric
sapply(sv_numeric, is.numeric)
```

Remover ceros

```{r}
sv_numeric<- as.matrix(sv_numeric[rowSums(sv_numeric)>0,])
```

Z-score function

```{r}
cal_z_score <- function(x){
  (x - mean(x)) / sd(x)
}
```

##Z-score calculation using the function described above

```{r}
sv_numeric <- t(apply(sv_numeric, 1, cal_z_score))
# Mantener los nombres de las filas

# Supongamos que sv es tu data.frame original
# y que la primera columna contiene nombres o IDs
```

```
# Filtrar filas donde NO todas las columnas (excepto la primera) son 0 o NA
sv_numeric <- sv[!apply(sv[ , -1], 1, function(row) all(is.na(row) | row == 0)), ]

# Asignar nombres de fila desde la primera columna
rownames(sv_numeric) <- sv_numeric[[1]]

# Eliminar la primera columna si ya no la necesitas
sv_numeric <- sv_numeric[ , -1]



#rownames(sv_numeric) <- sv[[1]]  # Nombres de filas



# Mantener los nombres de las columnas
colnames(sv_numeric) <- colnames(sv_numeric)  # Suponiendo que top_values y bottom_values tienen las mismas columnas
# Elimina la columna de nombres
sv_numeric <- sv_numeric[ , -1]
```

```
##Heatmap color

```{r}

mycol = colorRampPalette(brewer.pal(n=7, name = "YlGnBu"))(255)
```
```

```
##Saving the heatmap plot

png("sv_genes_S.png", 12000,35000, res=900)

pheatmap(sv_numeric,color = mycol,cluster_cols = FALSE, cluster_rows = TRUE, fontsize = 4, cluster_col=FALSE,fontsize_col = 6)

dev.off()
```

23. Anexo 23.

Se hace una lista con los genes que contienen SNPs pertenecientes a líneas sometidas a estrés hídrico. Ests genes están en anotación Wm82v1, por lo que se obtienen las secuencias y luego se realiza un Blast para determinar coordenadas y nombres de secuencias.

Ya que los genes de resistencia a stress están en anotación 1, lo que se hace es primero descargar las secuencias de la lista de genes respectivas a la anotación 1

Se aplica el script Fetchseqa1.py

Se hace un blast a Wm82v4.

Se forma el archivo blast_results_matched82a1.txt

De este archivo es extraen secuencias y cordenadas

~$ awk '{print $2, $9, $10}' blast_results_matched82a1.txt> RS_genes2.txt

Se aplica el script para encontrar los genes proximales más cercanos

Entrez_fetchgene.py

Para extraer columna 2 del blast de la lista de genes en wm82a1 a wm82a4(v4):

~$ awk '{print $2}' blast_results_matched82a1.txt > wm82a1seqid.txt

También se extrajo la column 9 y 10 formando el archivo RS_genes2.txt con el que se uso un script

Entrez_fetchgene.py para extraer los genes más próximos a las coordenadas dadas y el locus escrito

Luego se usó el script siguiente para extraer solo los nombres de los genes:

~$ awk '{ for (i=1; i<=NF; i++) if ($i == "Closest" && $(i+1) == "Gene:") print $(i+2) }' RS_genes3.txt > RS_genes4.txt

Se borran duplicados

~$ awk '!seen[$1, $2]++' RS_genes4.txt > RS_genes5.txt

Con este archivo se aplica el script

~$ python fetchlocustag3.py RS_genes5.txt RSgenes7.txt

Así se obtienen los aliases para hacer un mapa de calor.

Extraer 3ra columna con nombres de alias

~$ awk '{print $3}' RSgenes7.txt > RS_genes8.txt

En excel se elimino v4 del final de las secuencias y se reemplazo _ por -

Borrar duplicados

~$ awk '!seen[$0]++' RS_genes9.txt > RS_genes10.txt

Con esta información se usó un script Extracted_SV_ZHWM82_RS.rmd para extraer la información correspondiente de *Tabula Glycine.* Este script es similar a Extracted_SV_ZHWM82_3.rmd

Con esta información se hace un mapa de calor usando el script
R notebook heatmap total_RS10.rmd

```
---
title: "Heatmap total_RS10"
output: pdf_notebook
author: Mélida Rojas
---

Libraries

```{r}
library(pheatmap)
library(RColorBrewer)
```

Establecer carpeta de trabajo

```{r}
```
```

```
setwd("C:/Users/LT816/OneDrive - Valleflor/Documentos/Generales/Tesis/Genes variation and expression")
```

Leer el archivo

```{r}

sv<- read.csv("matching_rows4.csv", header = TRUE)



# Extraer data sín títulos, (Empieza desde columna 3 y fila 3)
sv_numeric <- sv[1:nrow(sv), 2:ncol(sv)]


#Convert all columns to numeric
sv_numeric[] <- lapply(sv_numeric, as.numeric)

sv_numeric[is.na(sv_numeric)] <- 0  # o cualquier otro valor


# Check if all columns are numeric
sapply(sv_numeric, is.numeric)

```

Remover ceros

```{r}
sv_numeric<- as.matrix(sv_numeric[rowSums(sv_numeric)>0,])
```

Z-score function

```{r}
cal_z_score <- function(x){
  (x - mean(x)) / sd(x)
}
```

##Z-score calculation using the function described above

```{r}
sv_numeric <- t(apply(sv_numeric, 1, cal_z_score))
# Mantener los nombres de las filas

# Supongamos que sv es tu data.frame original
# y que la primera columna contiene nombres o IDs
```

```
# Filtrar filas donde NO todas las columnas (excepto la primera) son 0 o NA
sv_numeric <- sv[!apply(sv[ , -1], 1, function(row) all(is.na(row) | row == 0)), ]

# Asignar nombres de fila desde la primera columna
rownames(sv_numeric) <- sv_numeric[[1]]

# Eliminar la primera columna si ya no la necesitas
sv_numeric <- sv_numeric[ , -1]


#rownames(sv_numeric) <- sv[[1]]  # Nombres de filas


# Mantener los nombres de las columnas
colnames(sv_numeric) <- colnames(sv_numeric)  # Suponiendo que top_values y
bottom_values tienen las mismas columnas
# Elimina la columna de nombres
sv_numeric <- sv_numeric[ , -1]
```

```

##Heatmap color

```{r}


mycol = colorRampPalette(brewer.pal(n=7, name = "YlGnBu"))(255)
##Saving the heatmap plot
png("sv_genesWM82_RS10.png", 12000,35000, res=900)
pheatmap(sv_numeric,color = mycol,cluster_cols = FALSE, cluster_rows = TRUE,
fontsize = 4, cluster_col=FALSE,fontsize_col = 6)
dev.off()
```
Top15 vallues

```{r}
# Extract the top 100 values (for each column)
top_values <- sapply(1:ncol(sv_numeric), function(i) {
  sorted_indices <- order(sv_numeric[, i], decreasing = TRUE)[1:15]
  sv_numeric[sorted_indices, i]
})

# Set row and column names for top_values
rownames(top_values) <- rownames(sv_numeric)[order(sv_numeric[,1], decreasing =
TRUE)[1:15]]  # Top 100 rows
colnames(top_values) <- colnames(sv_numeric)  # Column names from the original
matrix

# Extract the bottom 100 values (for each column)
bottom_values <- sapply(1:ncol(sv_numeric), function(i) {
```
```

```
  sorted_indices <- order(sv_numeric[, i])[1:15]
  sv_numeric[sorted_indices, i]
})

# Set row and column names for bottom_values
rownames(bottom_values) <- rownames(sv_numeric)[order(sv_numeric[,1])[1:15]]  # Bottom 100 rows
colnames(bottom_values) <- colnames(sv_numeric)  # Column names from the original matrix
```

```
```

Top15 genes

```
```{r}
# Combinar top_values y bottom_values
combined_values <- rbind(top_values, bottom_values)

mycol = colorRampPalette(brewer.pal(n=7, name = "YlGnBu"))(255)
##Saving the heatmap plot
png("sv_topS.png", 12000,35000, res=900)
pheatmap(combined_values,color = mycol,cluster_cols = FALSE, cluster_rows = TRUE, fontsize = 4, cluster_col=FALSE,fontsize_col = 6)
dev.off()
```

24. Anexo 24

Para extraer genes homólogos a arabidopsis se descargaron las secuencias completas de los genes modelos del sitio web TAIR:

TAIR10_seq_20101214_representative_gene_model.fasta

Se usó el script

TairFetchSeq.py para extraer la lista de 10 genes de arabidopsis del archivo de genes modelos

Se hizo un blast:

```
$ blastn -query tair_sequences.fasta -db Wm82v4_db -out blast_results_tair.txt -outfmt 6 -e
```

value 1e-10 -num_threads 6
```

y se consiguió solamente 6 sequences ids correspondientes a dos genes de la lista de 10.

Con esta información se usó un script similar a Extracted_SV_ZHWM82_3 para extraer la información correspondiente de *Tabula Glycine*
Finalmente se usó un script similar a R notebook heatmap total para crear el heatmap de los genes de producción y resistencia.