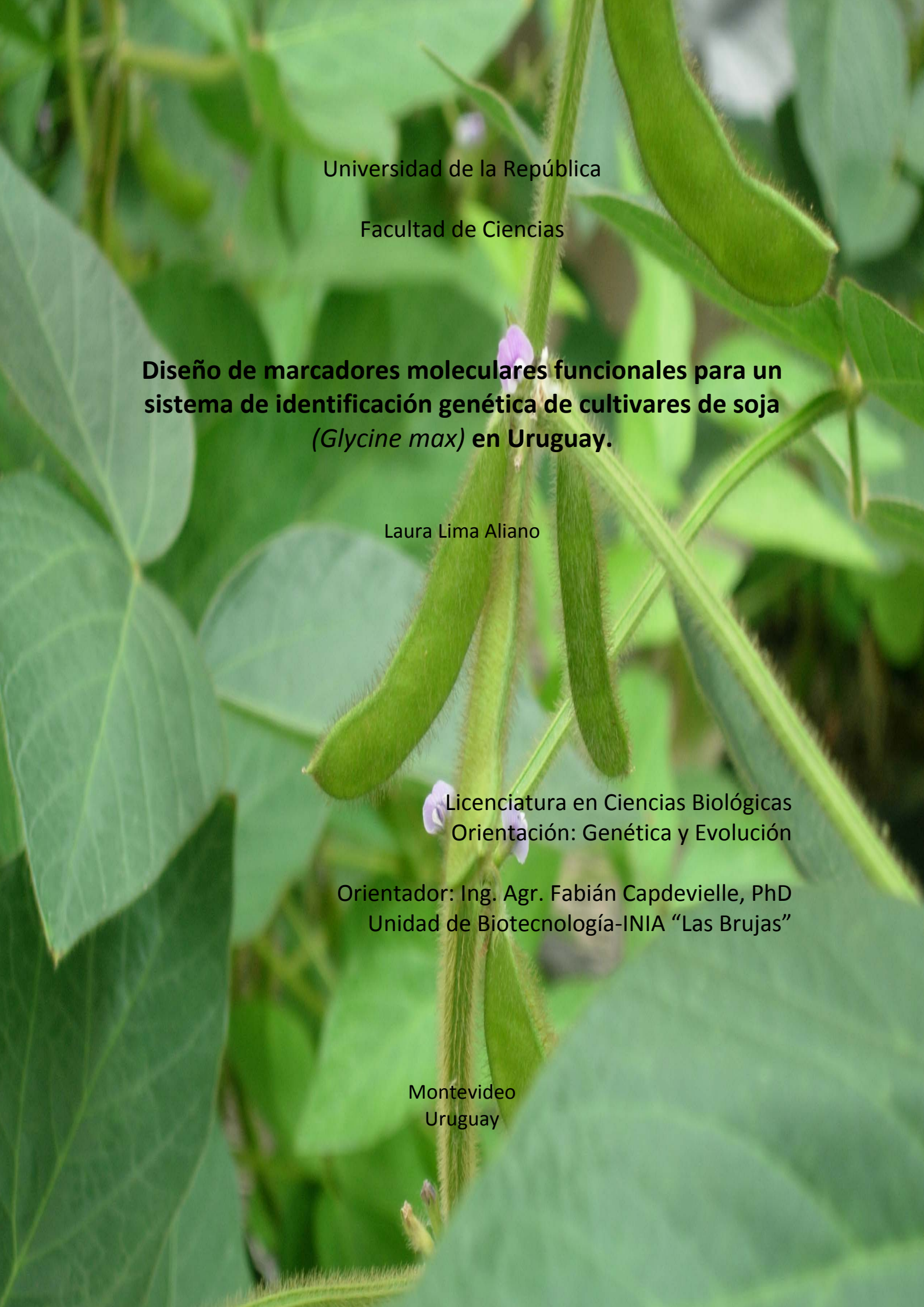


Universidad de la República

Facultad de Ciencias

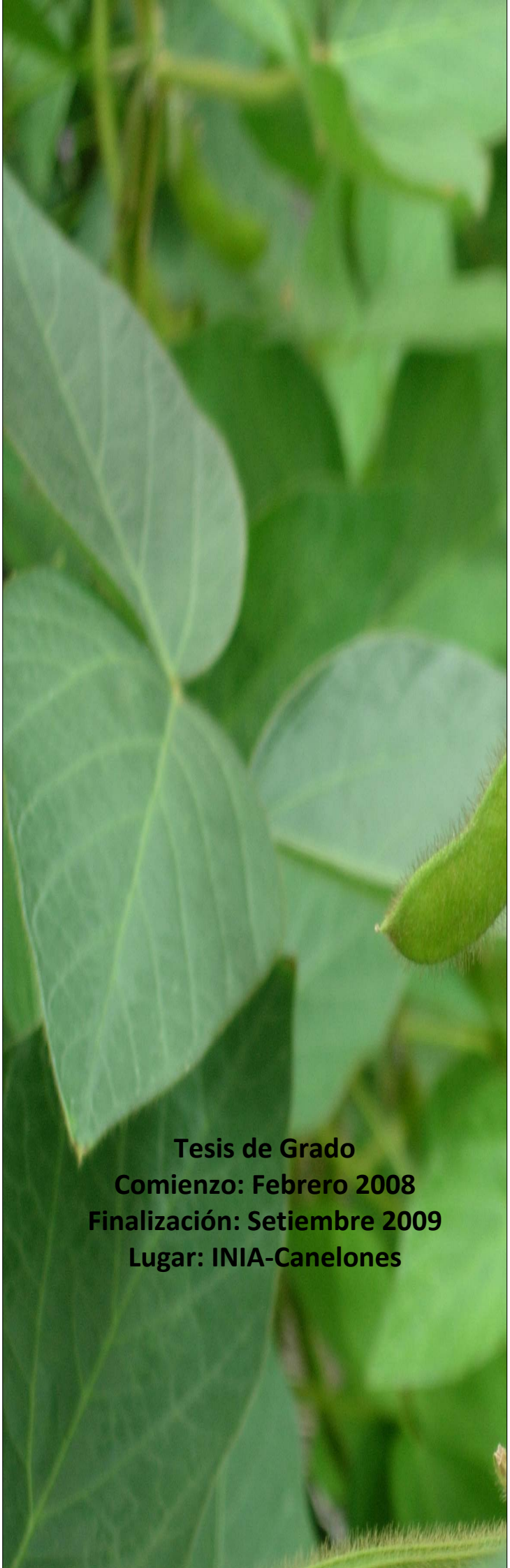
Diseño de marcadores moleculares funcionales para un sistema de identificación genética de cultivares de soja *(Glycine max)* en Uruguay.

Laura Lima Aliano

Licenciatura en Ciencias Biológicas
Orientación: Genética y Evolución

Orientador: Ing. Agr. Fabián Capdevielle, PhD
Unidad de Biotecnología-INIA “Las Brujas”

Montevideo
Uruguay



Tesis de Grado
Comienzo: Febrero 2008
Finalización: Setiembre 2009
Lugar: INIA-Canelones

Universidad de la República

Facultad de Ciencias

Diseño de marcadores moleculares funcionales para un sistema de identificación genética de cultivares de soja *(Glycine max)* **en Uruguay.**

Laura Lima Aliano

Licenciatura en Ciencias Biológicas
Orientación: Genética y Evolución

Orientador: Ing. Agr. Fabián Capdevielle, PhD
Unidad de Biotecnología-INIA “Las Brujas”

Montevideo
Uruguay

1. RESUMEN 3

2. INTRODUCCIÓN 5

2.1. El genoma vegetal - Glycine max 8

2.2. Análisis de SSR, contenido en GC y genes de soja 10

2.3. Organización del genoma de dicotiledóneas 11

2.4. Uso de microsatélites en plantas 12

2.5. Microsatélites y estrés ambiental 13

2.6. Respuesta a estrés abiótico en plantas 16

2.7. Objetivo general: 22

2.8. Objetivos específicos: 22

3. MATERIALES Y MÉTODOS 23

3.1. Marcadores moleculares anónimos 23

3.2. Marcadores funcionales: diseño de primers 24

3.3. Marcadores funcionales: DB-ICRISAT 27

3.4. Método automatizado (PLAN server) 27

3.5. Material vegetal 28

3.6. Extracción y cuantificación de ADN 29

3.7. Genotipado 30

3.8. Análisis de datos 32

3.9. Actualización de información 34

4. RESULTADOS 35

4.1. Marcadores funcionales: diseño de primers 35

4.2. Método automatizado: PLAN 37

4.3. Extracción y cuantificación de ADN 38

4.4. Amplificación de los microsatélites 39

4.5. Detección de los microsatélites 41

4.6. Análisis de datos: PIC 45

4.7. Análisis de datos: análisis discriminante entre variedades 46

4.8. Actualización de secuencias anotadas 50

4.9. Actualización de marcadores funcionales DB-ICRISAT 51

4.10. Actualización de marcadores anónimos (Satt) 51

5. DISCUSIÓN 54

5.1. Diseño de marcadores 54

5.2. Respaldo automático: PLAN Server 58

5.3. Detección de microsatélites 58

5.4. Análisis de PIC 59

5.5. Análisis discriminante 60

5.6. Actualización de secuencias funcionales y anónimas 61

6. CONCLUSIÓN 63

7. AGRADECIMIENTOS 64

8. BIBLIOGRAFÍA 64

9. ANEXOS 69

9.1. ANEXO 1: Matriz de tamaños alélicos para análisis de PIC 70

9.2. ANEXO 2: Matriz de 0 y 1 para análisis de discriminante 71

9.3. ANEXO 3: Actualización de anotación de marcadores funcionales diseñados manualmente .. 72

9.4. ANEXO 4: Actualización de marcadores funcionales DB-ICRISAT 78

9.5. ANEXO 5: Actualización de marcadores anónimos (Satt) 81

9.6. ANEXO 6: Planilla con los 44 marcadores funcionales anotados manualmente 91

9.7. ANEXO 7: Proceso de estrés abiótico en plantas 94

9.8. ANEXO 8: Resultado del análisis discriminante mediante uso de SAS 95

Diseño de marcadores moleculares funcionales para un sistema de identificación genética de cultivares de soja *(Glycine max)* en Uruguay.

1 RESUMEN

Los sistemas de marcadores moleculares aplicados en la identificación de variedades de cultivos han utilizado ampliamente microsatélites (SSR) localizados en regiones genómicas sin función conocida o asignada. Por esta condición, éstos SSR se denominan "marcadores anónimos" y en general no son aceptados como descriptores únicos para el reconocimiento de nuevas variedades a menos que se pueda demostrar su asociación con alguna característica fenotípica. Esta limitante puede ser superada a través del diseño de oligonucleótidos que permitan amplificar SSR asociados a una secuencia génica determinada (denominados SSR génicos ó funcionales). En el caso de especies de plantas donde se dispone de amplias colecciones de EST *(Expressed Sequences Tags)*, un gran número de genes están representados conjuntamente con diversos tipos de información (condiciones experimentales, orígenes, ensayos funcionales, etc.) para un conjunto de variantes genéticas dentro de la especie (variedades en el caso de especies cultivadas). Este tipo de información de secuencias generadas en forma aleatoria (a partir de secuenciación parcial de clones de cDNA y procesos de ensamblado (*in silico*) ha sido utilizada ampliamente como punto de partida para desarrollar sistemas de marcadores funcionales. **(1)(2)(3)**

En este trabajo se diseñaron marcadores funcionales a partir de secuencias expresadas, asociadas con respuesta de la planta a estrés

abiótico (frío, sequía, etc.) a través de la aplicación de diferentes estrategias de minería de datos. El conjunto de marcadores funcionales (EST-SSR) seleccionado, más un grupo de SSR anónimos fueron evaluados como herramienta de identificación varietal para cultivares de soja, con el objetivo de incorporarlos en una matriz de identificación de nuevas variedades de soja.

2 INTRODUCCIÓN

Las herramientas moleculares que se han utilizado ampliamente para el análisis individual de genes o pequeñas regiones cromosómicas vinculadas con la expresión de un carácter individual actualmente se aplican al análisis global de genomas completos, estudiando en conjunto los miles de genes, proteínas y metabolitos que constituyen un organismo, así como las complicadas redes de interacción que operan entre ellos. La información generada es enorme y es clave para la identificación y el aislamiento de genes de interés y permitirá interpretar, en términos moleculares, los procesos biológicos. Para ayudar en este proceso han surgido poderosas herramientas bioinformáticas que permiten almacenar e interpretar esta información. **(4)**

Por otro lado, el conocimiento de la secuencia de los genes posibilitara el desarrollo de marcadores "perfectos" -directamente basados en cambios detectables sobre la secuencia del mismo gen-, lo cual facilitaría la selección de los mismos en variedades de interés. La conjunción entre tecnologías de análisis molecular, marcadores moleculares, genómica y bioinformática tiene el potencial de revolucionar el mejoramiento genético vegetal, dando origen a lo que se denomina mejoramiento molecular, abriendo camino al desarrollo de nuevos y promisorios cultivares.

El objetivo de la ***genómica*** es la dilucidación completa y exacta de la secuencia de ADN de un genoma haploide representativo de una especie. La genómica se divide en dos grandes áreas; **genómica estructural**, que se ocupa de la caracterización física de genomas enteros y la **genómica funcional**, que se ocupa del estudio del transcriptoma, proteoma y metaboloma.

Por análisis computacional de la misma y utilizando principios conocidos de genética y el análisis molecular de los transcriptos y proteínas es posible:

- Comparar secuencias similares presentes en diferentes entidades biológicas y comprender el papel de dichas secuencias.
- Realizar predicciones acerca de todas las proteínas codificadas por una especie.
- Establecer las variaciones genéticas entre distintas poblaciones de una misma especie.
- Comparar secuencias de diferentes especies y entender procesos evolutivos.

Esto ha dado origen a la *genómica comparativa* y ha demostrado que existe considerable sintenia, es decir una localización conservada de los genes en posiciones equivalentes en especies relacionadas. También ha sido una excelente herramienta para identificar motivos o dominios de secuencias altamente conservados y por lo tanto, funcionalmente importantes en regiones codificantes y no codificantes del genoma. [(4)]

Los procesos celulares son controlados en varios niveles. La información básica es codificada por el genoma, que esencialmente es idéntico en cada célula del organismo, independientemente de la etapa de desarrollo o ambiente. El control del resultado de esta información invariante puede ocurrir en varios puntos, como se muestra en la Figura1. La información contenida en el ADN puede ser transcripta en el ARN, pero sólo un pedazo de la información en el ADN es transcripto alguna vez. La suma de todas las regiones de genoma que son transcriptas es colectivamente conocida como el **transcriptoma** y comprende todas las secuencias que arreglan el componente de ARN completo del organismo. Como no todo el ADN es transcripto, el

transcriptoma es menos complejo que el genoma. Los productos de la traducción del transcriptoma generan el **proteoma**. Sin embargo, dado que moléculas de ARN nunca son traducidas en proteínas, la complejidad del proteoma es menos que aquel de los transcriptomas. [(4)] Sin embargo, esta complejidad puede ser aumentada por modificaciones postraduccionales de las proteínas, potenciando los productos obtenidos de un mensaje dado. Estas proteínas entonces funcionan en la síntesis de metabolitos primario y secundario, el total de éstos constituyen el **metaboloma**. La combinación de ARN, proteínas, y metabolitos, funcionalmente integrados, causa la actividad biológica de una célula o tejido, y la información en todos estos procesos es necesaria para un entendimiento completo de forma de planta, función, y desarrollo.

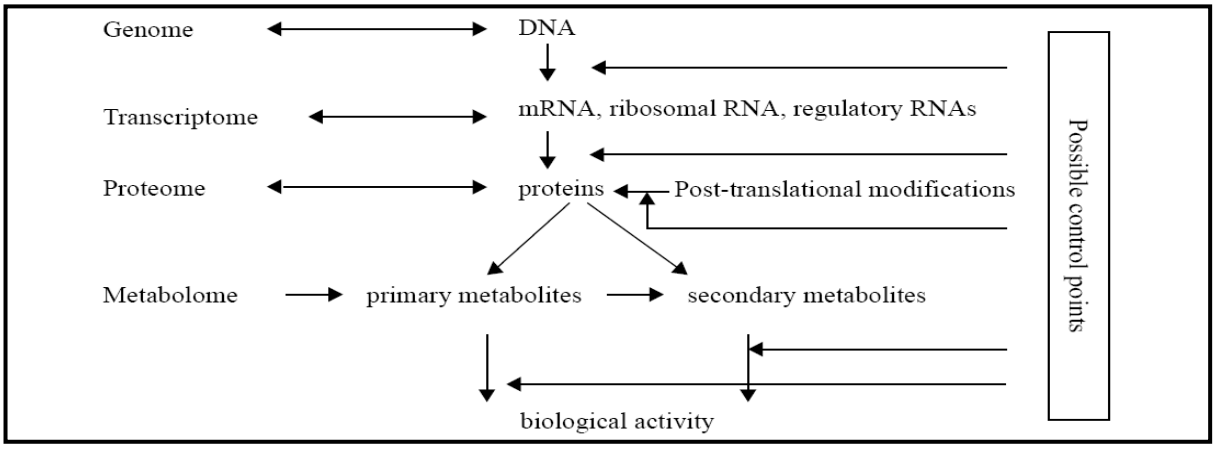


Fig.1- Niveles de control de los procesos celulares. (Adaptado de Jacobs et al., 2000) (4.1)

La información de secuencia de ADN ha sido, y sigue siendo, reunida tanto de ADN genómico como estudios de cDNA. Sin embargo, toda esta información de secuencia es sólo un primer pilar al entendimiento del control coordinado de la expresión de genes, así como entendiendo cuales de estos genes potenciales realmente son expresados en una manera funcional. La anotación de la secuencia genómica se refiere a los procesos (tanto manual como basada en un ordenador) por las cuales

varias regiones del genoma son clasificadas (regiones génicas, regiones no transcriptas, elementos transponibles, regiones de control conservadas). Los datos de ESTs y clones de cDNA de "largo completo" identifican las secuencias genómicas transcriptas a ARN. [4]

Se pueden identificar tres destinos posibles para una molécula de ARN:

- es traducido en una proteína
- funciona como un ARN sin cualquier modificación subsecuente
- es degradado sin realizar alguna vez realmente una función conocida específica

Por lo tanto, para entender el funcionamiento de la célula, la información adicional acerca de transcripción y traducción de cualquiera de los genes supuestos y la actividad subsecuente de algún producto proteico debe ser recopilada. [4]

2.1 El genoma vegetal - Glycine max

Soja (*Glycine max* L. Merr) es uno de los cultivos económicamente más importante en el mundo. Fue domesticada y cultivada en China por cientos de años, considerado hoy un caso de aplicación para desarrollar procedimientos de mejoramiento molecular porque tiene un mapa genético densamente saturado. [5]

Los trabajos realizados previamente en el estudio del genoma de la soja mediante ESTs, han servido como evidencia para demostrar que el procesamiento y ensamblado de estas secuencias en **unigenes** *(entendiendo como unigenes; una entrada en una base de datos que se corresponde a un set de secuencias transcriptas provenientes de un mismo locus -gen o pseudogen expresado- en conjunto con la información con similaridad de proteínas, genes expresados, clones de cADN y regiones genómicas)* son muy útiles para

hallar homología en otras especies como *Medicago truncatula* o *Arabidopsis thaliana.* Estudios realizados previamente por **Ai Guo** et *al.* (2003) [5] muestran que de el análisis de 629.000 ESTs (los cuales fueron ensamblados en 56.000 unigenes) cerca del 77 % de estos unigenes son homólogos con genes de Arabidopsis. Los productos hipotéticos de los unigenes fueron anotados de acuerdo con su homología a proteínas de Arabidopsis. Los genes de soja que no mostraban alta similaridad en el genoma de Arabidopsis fueron identificados como genes específicos de soja. [5]

El análisis de SSR mostró que el genoma de soja es más complejo que el de Arabidopsis y el de *Medicago truncatula*. Aunque el contenido GC en soja en secuencias de unigenes es similar a estas dos especies. [5]
Los unigenes de soja fueron buscados en contraste con el proteoma de Arabidopsis mediante Blastx y contra unigenes de *M.truncatula* mediante t-Blastx por análisis de homología. Los resultados en la **figura2** muestran que cuando el valor esperado (e-value) fue menor a 1.0E-1, 77 % de los unigenes de soja fueron homólogos a los genes de Arabidopsis y 82 % de ellos fueron homólogos a unigenes de *M. truncatula*. [5]

Cuando el e-value fue más bajo, el número de secuencias homologas decreció en ambas comparaciones. Cuando el e-value fue menor que 1.0E-180, solo el 1.8% de los unigenes de soja mostraron homología con el proteoma de *Arabidopsis* y 2.33% mostraron homología con unigenes de *M.truncatula*. [5]

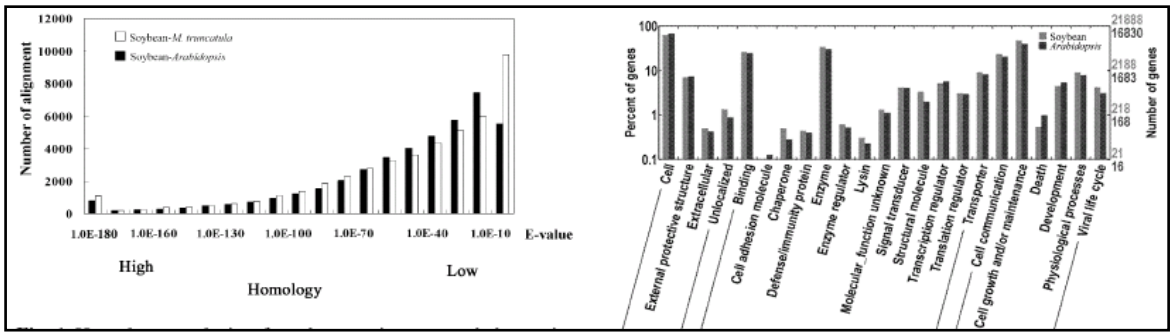


Fig.2 Se muestra un análisis de homología de unigenes de soja comparados con aquellos de *Arabidopsis* y *Medicago truncatula*. Las columnas oscuras representan el número de genes que muestran homología en secuencias aminoacídicas, con un E-value entre soja y *Arabidopsis.* Las columnas claras son entre soja y *M.truncatula*.

Fig.3 Clasificación funcional de unigenes de soja de acuerdo al Gene Ontology Consortium. La clasificación fue asignada por homología con genes categorizados de Arabidopsis. Solo el 61.7% de 27.290 genes predichos de Arabidopsis fueron clasificados. Para soja, 39.0% de 56.147 unigenes fueron clasificados.

2.2 Análisis de SSR, contenido en GC y genes de soja

El genoma de soja contiene entre un 40% y un 60% de secuencias repetidas y heterocromáticas, fueron analizados los SSRs *(simple sequence repeats)* en ambas secuencias, **genómicas** y **unigénicas**. En secuencias genómicas la distancia promedio entre dos SSR en soja fue de 3.96 Kb, la cual es la mitad de lo que se observo en *M.truncatula* y *Arabidopsis*, lo que quiere decir que en soja hay mas secuencias repetidas. **(5)**

Sin embargo, en unigenes la distancia promedio es de 10Kb en las tres especies, indicando que los unigenes contienen menos secuencias repetidas. Este análisis mostró que el tipo más común de repetido en unigenes son trinucleótidos en las tres especies. En secuencias genómicas el tipo más común también son trinucleótidos en soja pero a diferencia de *Arabidopsis* y *M.truncatula* que son dinucleótidos.

Además, la proporción de unidades de repetidos de SSRs en unigenes fue diferente. **(5)**

En unigenes, el repetido más común para *G.max* y *M.truncatula* fue AG, mientras que el más común en *A.thaliana* fue GAA. En secuencias genómicas el repetido más común fue AT en las tres especies. Este análisis también mostró que las secuencias genómicas de soja son más complejas que aquellas en *M.truncatula* y *Arabidopsis* y que el ensamble de secuencias será más dificultoso para el genoma de soja. **(5)**

Otros resultados mostraron que el contenido en GC en secuencias unigénicas en las tres plantas eran más altas que en secuencias genómicas y el contenido en GC fue similar en soja, *M.truncatula* y *Arabidopsis*.

De la comparación con arroz, el contenido en GC es de 0.51 para exones y 0.43 para secuencias genómicas **(5.1)** ambos valores son mucho más altos que para soja, los cuales son de 0.43 y 0.35 para unigenes y secuencias genómicas, respectivamente.

El contenido en GC permite reflejar la estabilidad relativa del genoma de una planta. Dicho análisis indicó que el número de genes en el genoma de soja pueda ser de 63.501 siendo mucho mayor que el de arroz o tomate. **(5)**

2.3 Organización del genoma de dicotiledóneas

El conteo de bandas usando sondas de RFLP (fragmentos de restricción polimórfico) indica que más del 90% de todas las secuencias de baja copia en el genoma de soja está presente en más de dos copias. Consistente con esto, la correlación genética detallada usando hibridización basado en marcadores RFLP y múltiples poblaciones identificó muchos casos de regiones genómicas duplicadas. **(6)**

La presencia de copias "anidadas" sugirió que al menos uno de los genomas originales hubiera sido duplicado antes del acontecimiento

poliploidización más reciente. Así se espera que la mayor parte de los genes de soja se agrupen en familias de genes que consisten en dos o más parálogos. En un estudio de Schlueter et *al.* se analizaron ESTs de genes duplicados y se llego a la conclusión de que el genoma de soja se sometió a un principal evento de duplicación aproximadamente hace 15 a 44 MA (millones de años). **(6)**

El acontecimiento de copia más reciente causaría muchos pares de parálogos de genes, lo cual dificultaría la identificación de genes mediante el uso de ESTs. Aunque estudios preliminares hayan examinado el nivel de la variación de secuencia entre genes seleccionados y sus alelos en soja ningún análisis sistemático ha sido hecho hasta ahora. **(6)**

2.4 Uso de microsatélites en plantas

Las secuencias de ADN de mini (VNTR) y microsatélites (SSR) son dos categorías de secuencias repetidas que se presentan en eucariotas. Ellas se encuentran repetidas en tandem y dispersas a través del genoma representando muchos *loci.* Cada *locus* tiene distinto número de repeticiones variable, asociándose de esta manera a alelos específicos de alta variabilidad. A través del uso de estos marcadores se han obtenido patrones complejos de variabilidad en el ADN en animales, plantas, y microorganismos. **(7)**

Los microsatélites son muy atractivos para los genetistas pues combinan varias ventajas como su codominancia, multialelismo y su alta heterocigosidad. El alto nivel de polimorfismo que detectan permite una discriminación precisa entre individuos altamente emparentados. Además de ser altamente polimórficos, el análisis de microsatélites requiere cantidades mínimas de ADN. Una de las desventajas de los microsatélites es el tiempo y costo involucrado en el proceso del diseño

de cada cebador, sin embargo existe la posibilidad de usar los mismos cebadores en más de una especie. Hoy en día la disponibilidad de genomas secuenciados y el bajo costo en las técnicas de secuenciación hacen que esta desventaja prácticamente no exista. [7]

Aunque el uso de marcadores bioquímicos y moleculares en estudios de germoplasma y diversidad genética es limitado por el costo de los reactivos y en algunos casos por el elevado costo de los equipos, los avances técnicos que estas tecnologías han alcanzado en los últimos años, los hacen más accesibles a los genetistas y a los programas de mejoramiento. Ambos tipos de análisis son complementarios en la caracterización morfológica y agronómica del germoplasma y en el entendimiento de la diversidad y estructura genética de las poblaciones, especies y taxas. [7]

Tabla 1. Características generales de los marcadores moleculares.

Característica	Isoenzima	RFLP	RAPD	VNTR	AFLP	SSR
Polimorfismo	Bajo	Bajo-alto	Medio-alto	Medio-alto	Medio-alto	Alto
Estabilidad ambiental	Moderada	Alta	Alta	Alta	Alta	Alta
Número de loci	Medio	Alto	Alto	Alto	Alto	Alto
Reproducibilidad	Moderada-alta	Alta	Moderada- alta	Alta	Alta	Alta
Aplicación	Rápida-barata	Lenta- cara	Rápida- cara	Intermedia	Lenta- cara	Lenta-cara

RFLP: Fragmentos de restricción polimórficos.
RAPD: Amplificación de ADN al azar.
VNTR: Número variable de repeticiones en tándem.
AFLP: Amplificación de fragmentos polimórficos.
SSR: Secuencias simples repetidas.

2.5 Microsatélites y estrés ambiental

Las plantas frecuentemente son sometidas a condiciones ambientales adversas que afectan negativamente su crecimiento y desarrollo. Los avances en el conocimiento básico sobre los mecanismos moleculares involucrados en la tolerancia al estrés por bajas

temperaturas, déficit hídrico o anegamiento y en la resistencia o tolerancia a enfermedades, permiten asociar la acción de ciertos grupos de genes candidatos al grado de tolerancia al estrés. Gran parte de la investigación básica en plantas se ha orientado a dilucidar los mecanismos moleculares que subyacen a las respuestas de las plantas al estrés ocasionado por estos factores, conocimiento que haría posible aumentar la tolerancia al estrés abiótico y biótico a través de programas de mejoramiento genético o enfoques biotecnológicos. [8]

Los sistemas de marcadores moleculares aplicables en identificación de variedades se han enfocado generalmente en regiones del genoma con alto grado de polimorfismo intervarietal, como por ejemplo secuencias conteniendo microsatélites (**SSR**=simple sequence repeat), sin considerar el grado de expresión fenotípica de estas secuencias (marcadores "anónimos"). Los marcadores moleculares SSR son secuencias cortas repetidas en *tandem* con motivos de di, tri o tetra nucleótidos, su variación en longitud es detectada por PCR al utilizar cebadores que amplifican las secuencias que limitan con la repetición. [8]

Los SSR son ampliamente utilizados debido a que ofrecen las siguientes ventajas: i) polimorfismo muy elevado (multialélicos); ii) enorme número de loci; iii) herencia mendeliana codominante; iv) alta reproducibilidad; v) interpretación sencilla de los resultados; vi) posible automatización. También pueden presentar algunas desventajas: i) requerimiento de conocimientos previos sobre el genoma de la especie en cuestión; ii) alto costo para el desarrollo de cebadores; iii) unilocus, aunque pueden hacerse multiplex; iv) alta tasa de mutación que puede afectar la reproducibilidad en estudios genealógicos. [7]

Diferentes trabajos de la última década han revelado que muchas respuestas fisiológicas y moleculares al estrés ocasionado por sequía, salinidad y bajas temperaturas están articuladas a través de vías compartidas. [9] Un alto número de SSR han sido localizados en regiones transcriptas del genoma, incluyendo genes que codifican para proteínas conocidas e identificadores de secuencias expresadas (EST= Expressed sequence tags), aún cuando en general el número de repetidos y longitud de los SSR en esas regiones es relativamente menor en comparación con regiones genómicas. [9]

Estos ESTs tienen un rango de funciones que incluye enzimas metabólicas, proteínas estructurales y de almacenamiento, señalización de interacciones planta-patógeno y factores de trascripción, lo que indica la posibilidad de asociar variantes alélicas con características funcionales. En el caso de especies de plantas donde se dispone de amplias colecciones de EST, un gran número de genes están representados conjuntamente con diversos tipos de información (condiciones experimentales, orígenes, ensayos funcionales, etc.) para un conjunto de variantes genéticas dentro de la especie (variedades en el caso de especies cultivadas). Este tipo de información de secuencias generadas en forma aleatoria (a partir de secuenciación parcial de clones de cDNA y procesos de ensamblado *in silico*) ha sido utilizada ampliamente como punto de partida para desarrollar sistemas de marcadores funcionales con mejores resultados en cuanto a su validación en comparación con los enfoques tradicionalmente utilizados. **(1) (2) (3)**

2.6 Respuesta a estrés abiótico en plantas

El estrés abiótico causado por sequía, salinidad, temperaturas extremas, toxicidad química y estrés oxidativo genera serios problemas para la agricultura. En general, representa una de las causas primarias en la pérdida de cosechas alrededor del mundo, reduciendo la producción promedio de cosecha en más de un 50%. [10] El estrés abiótico conduce a una seria de cambios morfológicos, fisiológicos, bioquímicos y moleculares que negativamente afectan el crecimiento de las plantas y la productividad de los cultivos. [11]

La sequía, salinidad, las temperaturas extremas y el estrés oxidativo a menudo están interconectadas, y pueden inducir un daño celular similar. Por ejemplo, la sequía y la salinización se manifiestan principalmente como estrés osmótico, causando la interrupción de los procesos de homeostasis y distribución de iones en la célula. [12][13] El estrés oxidativo, que con frecuencia acompaña las temperaturas altas, la salinidad, o el estrés por sequía, puede causar desnaturalización de proteínas funcionales y estructurales. [14] Como consecuencia, estos estreses ambientales activan en la célula diversos mecanismos de señalización similares [15] [16] y respuestas celulares, como la producción de proteínas de estrés, regulación de antioxidantes y acumulación de solutos. [17][18][19]

La compleja respuesta de las plantas al estrés ambiental, la cual envuelve mecanismos de control moleculares de tolerancia al estrés, está basada en cambios en la expresión de determinados genes causados por el estrés específico.

Estos genes incluyen tres categorías principales: **(i)** aquellos que están implicados en cascadas señaladas y en el control transcripcional, como MyC, MAP kinasas y SOS kinasa y factores transcripcionales como HSF y

familias de CBF/DREB y ABF/ABAE [20]; **(ii)** aquellos que funcionan directamente en la protección de membranas y proteínas, como "heat shock proteins" (Hsps) y chaperonas, proteínas LEA ("late embryogenesis abundant"), osmo-protectoras y proteínas supresoras de radicales libres [21]; **(iii)** aquellos que están implicados en el consumo de agua, iones y en el transporte, como acuaporinas y transportadores de iones. [22] Los estreses primarios, como sequía, salinidad, frío, calor y contaminación química a menudo están interconectados, causando daño celular y desencadenando estreses secundarios, como estrés osmótico y oxidativo.

Las señales de estrés iniciales (efectos como por ejemplo; osmóticos e iónicos, temperatura, cambios de fluidez de la membrana) provocan los controles de transcripción y procesos en cascada que activan mecanismos sensibles por el estrés para restablecer la homeostasis y proteger y reparar el daño de proteínas y membranas. La respuesta inadecuada en uno o varios pasos de la señalización y la activación de genes pueden causar cambios irreversibles en la homeostasis celular y en la destrucción de proteínas funcionales y estructurales de membranas, conduciendo a la muerte celular. (Ver **anexo 7**)

Factores de transcripción y su significado en la tolerancia de plantas al estrés

Los genomas de plantas contienen un número grande de factores de transcripción (TFs); por ejemplo, en *A. thaliana* aproximadamente el 5.9 % de su codificación de genoma es para más de 1.500 TFs. [23]

La mayor parte de estos pertenecen a unas familias multigénicas grandes, como MYB, AP2/EREBP, bZIP y WRKY. Factores de transcripción son aquellos factores que contribuyen con la polimerasa para promover el alargamiento del transcripto de ARN.

Los miembros individuales de la misma familia a menudo responden diferencialmente a varios estímulos desencadenantes de estrés, por otra

parte, algunos genes de respuesta a estrés pueden compartir el mismo TF. Los factores de transcripción sensibles por la deshidratación (DREB) y "C-repeat binding factors" (CBF) se unen a elementos DRE y CRT *cis*-acting que contienen el mismo motivo (CCGAC) en el promotor.

Los miembros de la familia CBF/DREB1, como CBF1, CBF2, y CBF3 (o DREB1B, DREB1C, y DREB1A, respectivamente) son inducibles bajo estrés. Las proteínas de DREB/CBF son codificadas por familias multigénicas AP2/EREBP y median la transcripción de varios genes como rd29A, rd17, cor6.6, cor15a, erd10, kin1, kin2 y otros en respuesta a frío y estrés hídrico. **(24)**

La señalización de ABA desempeña un papel vital en respuestas de estrés en plantas evidenciado por el hecho que muchos de los genes inducibles por sequía estudiados hasta ahora también son inducidos por ABA. Dos familias; TF bZIP y MYB, están implicados en la señalización de ABA y su activación de genes. Muchos genes inducibles por ABA comparten la secuencia, **(C/T) ACGTGGC**. Trabajos de **Abe et *al*.** (2003) mostraron que los factores de transcripción MYB, AtMYC2 y AtMYB2 funcionan como activadores transcripcionales en la expresión de genes inducibles por ABA, sugiriendo un sistema regulador nuevo para la expresión de genes en respuesta a ABA, además del sistema regulador ABRE-bZIP. **(24)**

Acumulación de solutos bajo condiciones de estrés osmótico

La función primaria de solutos u osmolitos es la de mantener la célula turgente y con esto manejar el gradiente de consumo de agua. Los estudios recientes indican que los solutos también pueden actuar como supresores de radicales libres y chaperonas químicas para estabilizar membranas y/o proteínas. **(25)** Los solutos caen en tres grupos principales: aminoácidos (por ej. prolina), aminas cuaternarias (por ej. glycina betaina, dimetilsulfoniopropionato) y poli-ol/azúcares (por ej.

manitol, trehalosa). La sobre expresión de solutos en plantas transgénicas puede causar mejoras en la tolerancia a estrés. [25]

Antioxidantes y genes de detoxificación

La salinidad, la sequía, temperaturas extrema y el estrés oxidativo son acompañados por la formación de especies reactivas del oxígeno (ROS) como O2, H2O2, y OH [26], que dañan membranas y macromoléculas. Las plantas han desarrollado varias estrategias de antioxidación pudiendo así aumentar la tolerancia a la respuesta frente a distintos tipos de estrés. Los antioxidantes (supresores de ROS) incluyen enzimas como catalasas, superóxido dismutasa (SOD), ascorbato peroxidasa (APX) y glutation reductasa, así como otras moléculas como ascorbato, glutation, carotenoides, y antocianinas. Los compuestos adicionales, como osmolitos, proteínas (por ej. peroxiredoxina) y moléculas amfifílicas (por ej. tocoferol), también pueden funcionar como supresores de ROS. [26]

El papel de HSPs y proteínas tipo LEA

Para enfrentarse con el estrés ambiental, las plantas activan un gran número de genes que conducen a la acumulación de proteínas asociadas por el estrés específico. Proteínas como Hsps ("heat shock proteins") y proteínas LEA ("late embryogenesis abundant") son dos tipos principales de proteínas inducidas que se acumulan a causa de déficits de agua, salinidad y temperaturas extremas, desempeñando el rol de protección celular durante el estrés. [27]

HSPs y chaperonas moleculares

El mantenimiento de proteínas en sus conformaciones funcionales y la prevención de la acumulación de proteínas no nativas son particularmente importantes para la supervivencia de la célula bajo estrés ambiental. Muchas proteínas inducidas por estrés, sobre todo Hsps, se han mostrado como chaperonas, siendo responsables de la

síntesis de proteínas, maduración y degradación en una amplia serie de procesos celulares normales. Además, las chaperonas moleculares funcionan en la estabilización de proteínas y membranas, colaborando con el plegamiento de proteínas en condiciones de estrés. Dentro de 5 familias conservadas de Hsps (Hsp100, Hsp90, Hsp70, Hsp60 y Hsps), estas pequeñas proteínas que varían en tamaño de 12 a 40 kDa son las más frecuentes en plantas. Se ha demostrado que las Hsps no se expresan solamente en respuesta al shock térmico, sino también bajo estrés hídrico, salinidad, estrés oxidativo, y a bajas temperaturas. **(28)**

Trabajos de **Hamilton y Heckathorn** (2001) sugirieron que la Hsp podría actuar como antioxidante en la protección del transporte de electrones mitocondriales del Complejo-I durante el estrés por NaCl. Además, las Hsps están implicadas en muchos procesos del desarrollo, como el desarrollo de embrión, germinación, embriogénesis, el desarrollo de polen y la maduración del fruto. Las plantas muestran menos semejanzas de secuencia de Hsps que las presentes en otros organismos. **(28)**

La superproducción de Hsps también puede proteger a las plantas del estrés oxidativo. Estudios de **Sol et *al.*** (2001) demostraron que la expresión de *Arabidopsis* AtHSP17.6A es regulada por shock térmico y por estrés osmótico. **(28)**

LEA

Las proteínas LEA han sido encontradas en una amplia variedad de especies de plantas en respuesta al déficit de agua que resulta de desecación, estrés por frío y osmótico. Estas caen dentro de varias familias, con funciones y estructuras diversas. **(29)**

Predicciones de su estructura secundaria sugieren que la mayor parte de estas proteínas existan enrolladas alrededor de una alfa hélice. Se propone por lo tanto que las LEA y las proteínas dehidrinas existen como

estructuras en gran parte desplegadas en su estado nativo, aunque algunos miembros existan como dímeros o tetrámeros. [30]

La hidrofobicidad es una característica común en las LEA y otras proteínas sensibles a estrés osmótico. La estabilidad bajo calor es otro rasgo notable de proteínas de LEA. Otra característica común de ellas es que, en la mayor parte de los casos, la expresión de sus genes es transcripcionalmente regulada y sensible a ABA. Se ha sugerido que las proteínas LEA tienen funciones similares a las chaperonas. [30]

2.7 Objetivo general:

Bajo este enfoque se propone identificar polimorfismos para marcadores candidatos para el estudio de la variabilidad presente en soja y desarrollar un sistema de identificación genética para nuevas variedades de cultivos agrícolas combinando marcadores moleculares anónimos, sin efecto directo sobre el fenotipo y marcadores moleculares funcionales (asociados con respuesta a estrés abiótico), aplicable al registro de cultivares y aseguramiento de la identidad genética para cultivos agrícolas.

2.8 Objetivos específicos:

1. Ajustar un sistema de genotipado para variedades de soja disponibles en Uruguay integrando microsatélites mapeados (SSR genómicos) y marcadores funcionales (EST-SSR y secuencias génicas) asociados con respuesta a estrés abiótico.
2. Generar una matriz de identificación aplicable para variedades de soja a partir de información para marcadores anónimos y marcadores funcionales asociados con respuesta a estrés abiótico en dicho cultivo.
3. Establecer materiales y protocolos de referencia para la utilización de marcadores moleculares como descriptores complementarios para el registro de nuevas variedades de cultivos agrícolas y para verificación de calidad genética en muestras de semillas.

3 MATERIALES Y MÉTODOS

3.1 Marcadores moleculares anónimos

En la tabla siguiente se muestran las características de los *primers* para diversos marcadores anónimos obtenidos de la bibliografía [31] que se utilizaran para el genotipado de los 7 cultivares de soja y que no han sido previamente asociados a ninguna característica fenotípica conocida pero que servirán como guía de mapeo en el genoma de soja.

Tabla 2- Característica de marcadores anónimos.

Nombre	Secuencia F	Secuencia R	Referencia
satt_173	TGCGCCATTTATTCTTCA	AAGCGAAATCACCTCCTCT	Roa et al. (2005)
satt_175	GACCTCGCTCTCTGTTTCTCAT	GGTGACCACCCCTATTCCTTAT	Roa et al. (2005)
satt_42	GACTTAATTGCTTGCTATGA	GTGGTGCACACTCACTT	Roa et al. (2005)
satt_577	CAAGCTTAAGTCTTGGTCTTCTCT	GGCCTGACCCAAAACTAAGGGAAGTG	Roa et al. (2005)
satt_177	CGTTTCATTCCCATGCCAATA	CCCGCATCTTTTTCAACCAC	Roa et al. (2005)
satt_226	GCGAAACAACTCACTTAAGCAATACAT	GCGTCCTCCTACCTTTCTTATC	Roa et al. (2005)
satt_231	GCGTGTGCAAAATGTTCATCATCT	GGCACGAATCAACATCAAAACTTC	Roa et al. (2005)
satt_324	GTTCCCAGGTCCCACCATCTATG	GCGTTTCTTTTATACCTTCAAG	Roa et al. (2005)
satt_534	TTCATGCATATACATCACGTATTATT	TGTAAAAACTAAAGAATGGACTGTGG	Roa et al. (2005)
satt_70	TAAAAATTAAAATACTAGAAGACAAC	TGGCATTAGAAAATGATATG	Roa et al. (2005)

Las **reacciones de amplificación de PCR** se llevaron a cabo en un volumen final de 11µl, en un termociclador Eppendorf. El ciclado fue tomado de Roa et *al.* y los programas variaron con los diferentes pares de oligonucleótidos. El ciclado consistió en 3 minutos a 94°C, un paso siguiente de 35 ciclos con 30 segundos a 94°C, 1 minuto a 54°C, 2 minutos a 72°C y por último 8 ciclos con 30 segundos a 94°C, 45 segundos a 53°C, 45 segundos a 72°C y terminados estos, 2 minutos a 72°C. El mismo fue modificado a partir de información de Schuelke [32]

para ajustar las condiciones del fluorocromo M13 (ver **figura 4**). Cada mezcla de reacción contenía en su concentración final; 200µM de dNTPs, 1.5mM de MgCl2, 100nM de cada oligo, 0.5 U de Taq polimerasa y 25 ng de ADN. Esta misma reacción de amplificación fue utilizada para microsatélites funcionales.

3.2 Marcadores funcionales: diseño de primers

Hay disponibles en el **GenBank** más de 395.000 secuencias de ESTs, de las cuales 14.486 están asociadas a estrés hídrico y son utilizados como perfiles de genes transcriptos. Estas secuencias están disponibles en el NCBI (http://www.ncbi.nlm.nih.gov/sites/entrez) y abaladas por la SoyBase (http://www.soybase.org).

En este trabajo, fueron seleccionadas 13.486 secuencias de ESTs filtradas por palabra clave; **"drought stress"**. Una vez obtenidas las mismas se realizó el ensamblado de las secuencias a través del software **CAP3** (http://pbil.univ-lyon1.fr/cap3.php), el cual genera secuencias consenso (contigs) y clusters de secuencia única (singletons) producto del solapamiento de ESTs.

El algoritmo de ensamblado tiene tres fases, una primera etapa en la que hay un recorte automático de regiones pobres en solapamiento en sentido 5´y 3´, las cuales son identificadas y eliminadas. Para el cálculo eficiente del algoritmo de programación dinámica Smith y Waterman (1981), es restringido a una banda en la diagonal en la matriz de programación dinámica (Pearson and Lipman, 1988; Green, 1996). En una segunda fase las zonas de lectura se han unido para formar contigs superponiéndose en orden decreciente de puntuación.

En una tercera etapa, se ha construido un alineamiento múltiple de secuencias y una secuencia consenso con un valor de calidad para cada

base, el cual es calculado para cada contig. Esto le permite al software CAP3 tomarse más tiempo en reconocer errores en la secuencia y producir con mayor precisión las secuencias consenso y hacer frente a errores de montaje debido a repeticiones. [33]

Al conjunto de secuencias consenso o contigs obtenidos como resultado del procesamiento de los EST a través del software CAP3, se les realizó búsqueda de microsatélites como segundo criterio de filtrado. Para ello se utilizó una pipeline de acceso público, **PBC** (http://hornbill.cspp.latrobe.edu.au/cgi-inpub/ssrprimer/indexssr.pl), implementada por PBC_ Bionformatics. [34]

En dicha pipeline se implementan dos algoritmos, uno de búsqueda de motivos SSR (SSRIT) y el segundo para el diseño de primers (Primer3). [35]

Esta permite fijar los parámetros de búsqueda de motivos SSR, así como para el diseño de primers, siendo requerido tener el archivo de entrada (input) en formato multifasta. Los parámetros utilizados para el diseño de primers de este trabajo se basó en La Rota et al., 2005 [36] y Zhang et al., 2005 [37] y fueron los siguientes:

- Tamaño: Min=18 / Opt=20 / Máx=25
- Primer Tm: Min=50 / Opt=55 / Máx=65 / Max Tm Difference: 0
- Primer GC%: Min=20 / Opt= nada / Máx=80
- Primer Max Complementarity: 5
- Primer Max 3' Complementarity: 3

Como resultado se obtiene una tabla con la información para cada secuencia de cada SSR encontrado, tipo y largo de repetido, y el *primer* correspondiente para la obtención del microsatélite (ver tablas en resultados).

El **Blastx** fue utilizado para buscar secuencias aminoacídicas, este algoritmo permite el ingreso de secuencias nucleotídicas, las que son traducidas en sus 6 marcos de lectura y comparadas contra una base de datos de proteínas. Fue elegido como estrategia esta alternativa del algoritmo BLAST, debido a su mayor sensibilidad en comparación con el Blastn, debido a que este último utiliza una base de datos de nucleótidos para buscar la secuencia nucleotídica problema. Con el objetivo de encontrar de esta manera, cuáles de ellas (las secuencias resultantes de la pipeline PBC), podrían estar vinculadas a proteínas o familias de proteínas relacionadas a estrés abiótico.

Como referencia, las funciones de proteínas fueron predichas mediante Blastx y anotadas en términos asociados a procesos biológicos, componente celular y función molecular usando el vocabulario controlado GeneOntology **(GO)**. Las secuencias contenidas en el set de secuencias expresadas consenso (unigen) han sido anotadas y clasificadas en categorías funcionales definidas de acuerdo a este vocabulario utilizado para describir características de genes y productos o procesos derivados en cualquier organismo, GO disponible públicamente vía; http://amigo.geneontology.org/cgi-bin/amigo/go.cgi.

El **Pfam,** es una colección de proteínas, dominios y familias, cuya versión actual de (Pfam 22.0) contiene 9318 familias de proteínas. Las comparación realizadas por le PFAM están basadas no solo en UniProtKB, sino también en NCBI GenePept y en las secuencias seleccionadas de proyectos de meta genómica. [38] Esto nos servirá para poder discriminar dentro de nuestros candidatos cuales se encuentran dentro de la misma familia y que criterio tomar al momento de elegir los candidatos para ser probados experimentalmente. Pfam esta disponible en la web a través de http:// pfam.sanger.ac.uk /.

3.3 Marcadores funcionales: DB-ICRISAT

También fueron obtenidos otros marcadores funcionales a partir de DB-ICRISAT (http://www.icrisat.org/), base de datos en la que estaban disponibles secuencias EST-SSR reportadas para soja y que fueron seleccionadas a partir de su anotación, esta anotación fue confirmada vía Blast (NCBI) y se eligieron las que tenían relación con estrés abiótico. Se detallan a continuación:

Tabla 3- Característica de microsatélites funcionales obtenidos de DB-ICRISAT.

Nombre	Secuencia F (5´-3´)	Secuencia R (5´- 3´)	Repetido	Anotación
Q93WZ6	GCAAACCAAGTGGT GGCTAT	GGTCTTTCTCTGCCT TGTGC	(TA)7	similar to UP|Q93WZ6 (Q93WZ6) Abscisic stress ripening-like protein, partial (61%)
Q43453	TCGAAGCCTTCAGG AGTGTT	TATCCCAAGTTTGCC TCGTC	(CTAT)6 ; (ACTC)3	UP|Q43453 (Q43453) G.max mRNA from stress-induced gene (H4), complete
PI27_ARATH	GTTGGTAGGGTTGC TCCTGA	GCCGTTTTCATGGTT CACTT	(CGC)4	homologue to UP|PI27_ARATH (P93004) Aquaporin PIP2.7 (Plasma membrane intrinsic protein 3) (Salt-stress induced major intrinsis protein), partial (97%)
Q6XPS8	GTTATGGACCAGAA GCCCAT	TTTGCCTCTCTGTTC ATTTCAA	(GCTC)3; (ATAAA)3; (TTTTA)3	homologue to UP|Q6XPS8 (Q6XPS8) Drought-induced protein 1, complete
Q41111	GCTCTGAGGTTGAG GTCCAG	TCCTCAGCATGTGCT GTTTC	(ATC)4 ; (GAA)5 ; (CTT)5 ; (TCACT)3; (GAA)5 ; (TGA)4 ; (GAT)4 ; (AGTGA)3	homologue to UP|Q41111 (Q41111) Dehydrin, partial (81%)

3.4 Método automatizado (PLAN server)

El navegador personal de Blast o PLAN Server es una plataforma versátil para ayudar a usuarios a hacer tareas personalizadas de Blast incluyendo búsqueda automatizada de resultados de alto rendimiento con el filtrado de resultados en línea, manejo de resultados de interés personal en categorías preferidas, anotación de secuencias automatizadas (tales como NCBI, NR y GOntology), etc. PLAN integra,

por defecto, el Decypher basado en algoritmo Blast, solución proporcionada por Active Motif Inc. Se encuentra disponible en: http://bioinfo.noble.org/plan/ . [39]

Los resultados de Blast se visualizan en hojas de cálculo y gráficos, la secuencias con la correspondiente anotación de posible función se pueden exportar en parte o en su totalidad, en diversos formatos, como excel o fasta. Además, todas las funciones analíticas se presentan a los usuarios sin necesidad de registro público.

PLAN, ha demostrado ser una herramienta de automatización de alto rendimiento y muy valiosa para el descubrimiento y manejo de los resultados de alineamiento de secuencias. Otra de las aplicaciones que ofrece, es la anotación funcional de las secuencias de consulta y proporciona una alta gama de opciones para este fin. Además, tiene aplicado el GO, para la clasificación de funciones, al igual que POntology y KEGG (Kyoto Encyclopedia of Genes and Genomes), y se están estudiando otras formas de aplicación y exportación de datos. [39]

Dicho algoritmo fue utilizado para el respaldo y validación de los resultados obtenidos por el método antes descrito.

3.5 Material vegetal

Se analizaron 7 cultivares de soja (***Glycine max* L.**): **DM4200, NM55R, NM70R, RA514, RA518, TJS2055 y TJS2178R** (origen de las semillas: **INASE**). La principal dificultad fue la obtención de material verde para la extracción de ADN, la germinación de las

semillas fue una etapa dificultosa, teniendo que recurrir a formas alternativas a las estandarizadas en el laboratorio de la Unidad.
Finalmente, las plantas fueron germinadas y crecidas en INASE mediante el método de rollo y arena, tomado de: La Asociación Internacional de Análisis de Semillas (ISTA) que establece los procedimientos estándar para la toma de muestras y el análisis de semillas, a fin de promover la uniformidad en el análisis de las semillas que son el objetivo del comercio internacional. Una vez obtenido el material verde, se llevó a cabo la extracción del ADN genómico.

3.6 Extracción y cuantificación de ADN

El aislamiento de ADN genómico foliar se hizo de acuerdo a la técnica propuesta por el protocolo de la FAO/IAEA Interregional Training Course on Mutant Germoplasm Characterization (International atomic energy agency, Vienna, 2002) con modificaciones menores.
Se molieron 200 mg de tejido vegetal en mortero con nitrógeno líquido, colocando el material finamente molido en tubos con 7 ml de buffer de extracción precalentado a 65°C (2% CTAB, 1,4 M NaCl, 20 mM ácido etilendiamintetracético [EDTA]. 0.125% de β–mercaptoetanol agregado en el momento). Luego de incubar las muestras a 65°C durante 20 minutos con agitación suave y periódica, se le agregó 1 volumen de Cloroformo-isoamílico (24:1) y se mezcló cuidadosamente. Posteriormente se centrifugaron a 12000 rpm durante 20 minutos. El sobrenadante fue transferido a otro tubo, donde se precipitó el ADN con 0.7 volúmenes de isopropanol. Las muestras fueron centrifugadas a 5000 rpm durante 20 minutos a temperatura ambiente (TA). Los precipitados fueron lavados con etanol 70% y centrifugados a 5000 rpm durante 5 minutos, luego secados en estufa a 37ºC.

La cantidad y calidad del ADN geonómico obtenido se evaluó mediante geles de agarosa 1%, TBE 0.5X con tinción de bromuro de etidio y marcador de peso molecular Mass Ruler (1Kb). La concentración del ADN genómico fue ajustada a 100 ng/μL como solución de trabajo.

3.7 Genotipado

Mediante este método se detectarán los fragmentos amplificados (alelos) con marcación fluorescente de un conjunto de marcadores funcionales y anónimos, utilizando una plataforma de secuenciación automática (ABI 310).

Antes de ingresar al secuenciador se precipitaron los productos de PCR, para ello, se agregó 80ul de isopropanol 75%, luego se hace un vortex corto y se dejan los tubos 15 minutos a temperatura ambiente. Se centrifugan 20 minutos a 13.000 rpm. Se descarta el sobrenadante y se agregan 150 ul de isopropanol 75%. Se centrifugan 5 minutos a 13.000 rpm. Se descarta el sobrenadante y se seca el pellet en estufa a 37ºC. Se resuspende el pellet en 15 ul de agua (mili-Q).

Para preparar las muestras para el secuenciador se utilizó información de Schuelke descrita anteriormente, luego son desnaturalizadas 4 minutos a 95ºC en termociclador Eppendorf y se pasan a hielo inmediatamente.

Los pares de primers seleccionados fueron sintetizados para su aplicación experimental. En el extremo 5’ de cada oligonucleótido “directo” (forward) se adicionaron 18pb (5’- TGTAAAACGACGGCCAGT-

3') complementarios al M13 (ver **figura 4**) a fin de incorporar un sistema de marcación fluorescente de los productos amplificados. Obteniendo ventajas económicas adicionales frente al uso de *primers* marcados de forma individual.

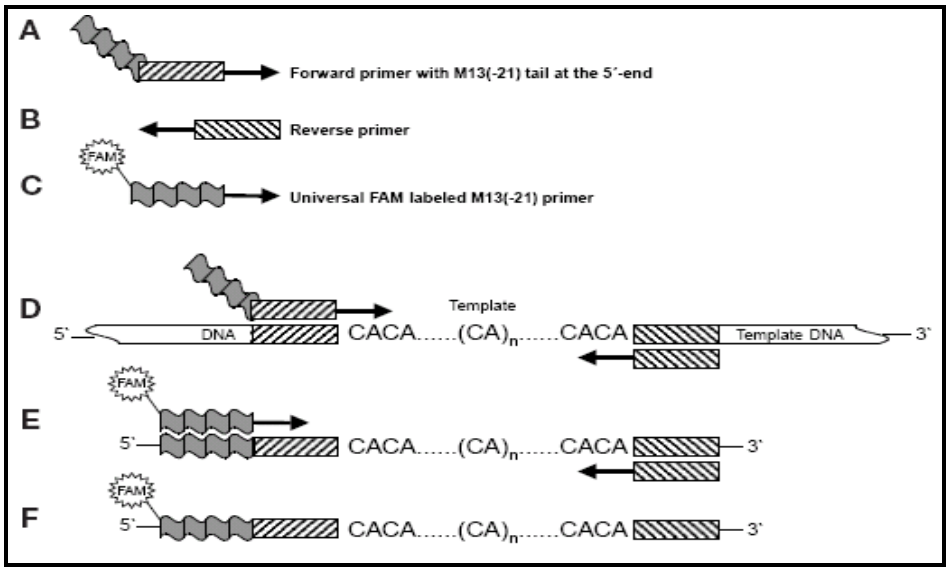


Fig.4- La figura muestra un esquema de amplificación para una reacción de "PCR anidada". (A-B) las cajitas a raya indican donde va el primer especifico para el microsatélite; (C) la colita gris ondulada es la secuencia universal M13 (-21) y la etiqueta FAM fluorescente; (D) en los primeros ciclos de PCR, el primer forward con el M13 (-21) es incorporado en los productos de PCR; (E) estos productos son entonces el objetivo para el FAM, el cual es incorporado durante ciclos subsecuentes a una temperatura "annealing" de 53ºC; (F) el producto final puede ser analizado en un sistema de detección láser. Tomado de **Schuelke, 2000**.

Para cada par de primers se ajustaron las condiciones de reacción de PCR y de corrida en el secuenciador automático ABI PRISM 310 Genetic Analyzer (PE Applied Biosystem, Foster City, CA) donde los productos amplificados son separados mediante electroforesis en capilar usando el módulo FAM-JOE-TAMRA. Para ello, 0.5 µl del producto amplificado fueron mezclados con 11.6 µl de Hi-DiTMformamide (PE Applied Biosystem, Foster City, CA) y 0.4 µl de marcador de peso molecular de ADN marcado con ROX (GeneScan®-500 [ROX] ™) (PE Applied Biosystem, Foster City, CA).

El análisis primario de los datos obtenidos se realiza con el programa GeneScan®, donde los productos de amplificación correspondientes a cada microsatélites son observados como picos de emisión del fluorocromo utilizado en el correspondiente electroferograma. Estos picos fueron analizados mediante el programa Genotyper® v 3.7 a efectos de establecer el tamaño de los alelos para cada SSR analizado en diferentes variedades. Mediante esta recolección de información (tamaño de los picos) será posible generar los archivos para el posterior análisis de datos.

3.8 Análisis de datos

Fue analizada la heterocigosidad esperada o **PIC** (contenido de información polimórfica) para marcadores anónimos y funcionales mediante el software ***Cervus 3.0.3***. **© Copyright Tristan Marshall 1998-2007.**

Este índice se define como:

$$\mathbf{PIC = 1- \Sigma p_i^2}$$

Siendo p_i la frecuencia del alelo i-ésimo.

El PIC es una medida de la diversidad genética y provee una estimación del poder discriminatorio de cada locus, teniendo en cuenta el número de alelos y sus frecuencias relativas en cada población.
Su valor varía de 0 (monomórfico) a 1 (altamente polimórfico con muchos alelos con frecuencias semejantes). El valor promedio de PIC para el total de loci representa la diversidad genética total (Ht), es decir, la diversidad dentro de cada población (Hs) más la diversidad entre poblaciones (Dst). **(40)**

Ver matriz de tamaños alélicas para el análisis en **anexo 1**.

Como segundo paso; se realizó un **análisis de discriminante**; a partir de los tamaños alélicos obtenidos (para marcadores anónimos y funcionales), se generó una matriz de datos *accesión x marcador*, donde los datos fueron codificados como ceros (ausencia) ó unos (presencia) para cada uno de los alelos identificados en los electroferogramas. (Ver matriz de 0 y 1 en **anexo 2**)

El algoritmo utilizado corresponde al tipo de análisis discriminante no paramétrico (K -nearest neighbor, k – NN) incluido dentro del procedimiento DISCRIM implementado en SAS/STAT® software 9.1) [SAS Institute, 2002-2003], utilizando una combinación de variables como predictores de la clase a la que pertenece cada observación (en este caso las variables predictivas son los marcadores moleculares analizados).

El algoritmo k-nearest neighbor clasifica cada observación del conjunto en estudio (sobre las que se quiere tener una predicción), en base al conjunto de referencia definido (aquellas observaciones para las que se conoce su clase) de acuerdo a los siguientes criterios: **1)** encuentra los k vecinos más cercanos (K=1 en este caso) dentro de las observaciones del conjunto de referencia, y **2)** predice la clase según la del vecino más cercano (si K>1, utiliza la clase más frecuente), por ejemplo, elige la clase que es más común entre los k - vecinos más cercanos. **(41)**

Para realizar las predicciones en clases se definió el vecindario por la distancia desde x a los kth puntos más cercanos del conjunto de referencia; luego se estima la probabilidad de que pertenezcan a la clase j, por la proporción de los puntos de referencia que pertenecen a dicha clase entre los k más cercanos a x.

3.9 Actualización de información

En última instancia; en Enero del 2006, el Departamento de Energía de Estados Unidos **(DOE)** y el Departamento de agricultura de Estados Unidos **(USDA)** anunciaron un acuerdo para coordinar el secuenciado de plantas y genomas microbianos. El primer genoma a ser secuenciado bajo este acuerdo sería el de soja (*Glycine max*) en coordinación con el **DOE Joint Genome Institute** (JGI) quien suministro la secuencia y el análisis del genoma y el USDA quien proporcionaría los recursos genómicos y la interacción con la comunidad de mejoradores de soja.**(42)**

En Diciembre del 2008 fue publicado en **Biology & Nature** que había finalizado la secuenciación del primer borrador del genoma de soja, por parte del **DOE-JGI** y que estaría disponible a la comunidad de investigación para renovar las estrategias de investigación en relación a este cultivo tan valioso internacionalmente; este estaría disponible a través de: **http://www.phytozome.net/soybean**. **(42)**

Teniendo en cuenta esto último y debido al reciente evento de la completa secuenciación del genoma de soja se realizó la **actualización de la información**, que consistió en localizar todas las secuencias de marcadores anónimos y funcionales utilizadas en este trabajo, en la web donde está disponible toda la información de secuencias de *Glycine max*. Para ello fue utilizada la web: www.phytozome.net, para detallar la ubicación de los marcadores moleculares funcionales en el genoma de soja. (Detalles de secuencias en **anexos 3, 4 y 5**)

4 RESULTADOS

4.1 Marcadores funcionales: diseño de primers

Como resultado del ensamblado utilizado CAP3 se obtuvieron **3882 contigs** y **1766 singletons**. Los contigs (secuencias consenso) fueron sometidos a estudios posteriores para elegir entre ellos los posibles genes candidatos a estar relacionados con estrés abiótico.

Luego del análisis realizado a través de la *pipeline* PBC, con parámetros fijados como se detallo en metodología, se obtuvieron **570 contigs** los que contenían SSR y el juego de *primer* correspondiente. Para algunos contigs se encontró más de un tipo de repetido, para los cuales también se diseño un primer específico.

Para estas 570 secuencias se realizó una anotación o asignación de funciones utilizando como estrategia la trasferencia horizontal de información con la utilización del programa BLASTx.

A partir de este procesamiento realizado manualmente se obtuvieron 44 secuencias que fueron encontradas y anotadas como mejores candidatos, (se detallan en **anexo 6**) tomando en cuenta la siguiente información: e-value, % de identidad con A. thaliana (planta dicotiledónea de la familia de las Brassicaceae, modelo completamente secuenciada) y GOprocess, siendo este último el que se vinculaba mejor a factores de estrés abiótico, en nuestro interés particular, estrés hídrico. A partir de una base de datos de familias de proteínas (**Pfam**) fue posible definir para cada candidato una familia o dominio al que pertenecía.

De estos 44 candidatos, los cuales están todos relacionados a respuesta a estrés abiótico, ya sea por respuestas universales (USP), respuesta al frío, al calor, a estrés por salinidad, a estrés oxidativo, respuesta a

congelamiento, a estímulos por ABA (acido abscísico), etc. fueron elegidos **6 marcadores** para ser probados experimentalmente. Para su elección se tuvo en cuenta también, características de los *primers*, largo de repetido, además de la anotación vinculada a estrés abiótico.

A continuación se detallan las características de los 6 contigs candidatos seleccionados que fueron probados experimentalmente:

Tabla 4- Muestra el número de acceso al NCBI, el contig correspondiente a cada secuencia y el tipo de repetido para cada una de ellas.

Número de acceso del NCBI	Nombre asignado	Sequence ID	Repeat Type	Repeat
ABQ81887.1	Lau1	Contig75	trinucleotide	AGAAGAAGAAGAAGGAGAA GAAGAAGAAG
NP188945.1	Lau2	Contig951	dinucleotide	ATATATATATATATATATATA
NP200414.1	Lau4	Contig2193	tetranucleotide	GAACGAACGAACGA
NP568755.1	Lau5	Contig2595	dinucleotide	TCTCTCTCTCTC
NP177745.1	Lau6	Contig2764	trinucleotide	TGATGATGATGAT
NP189283.1	Lau7	Contig3215	tetranucleotide	AGCAAGCAAGCAA

Tabla 5- Aquí se observa el juego de primer diseñado para cada contig con su correspondiente TM= temperatura de melting.

Left Sequence	Right Sequence	Left TM	Right TM
AAAGTCAAGGACAAGATCCA	TTAATCACTGTCACTGCTGC	54.727	54.91
CACCAGACACATTCAAGCTA	GGGCCTAATACAATGTTCAA	54.745	55.183
CTCTTGTCCGTCAAACTCTC	CGCCCTAAACGACTCTATC	54.965	55.064
GCAACCTAAATCAGAGAGGA	TCATCATCAGTTTCAAGCAG	54.573	54.792
TCCTTAATCTTGTCGAGGAA	AAGTGAGGAGGAAGGAGAAG	55.016	55.157
ATTATCCCAAGATCACCCTT	ATGTAAACCCTGCTGTCACT	55.013	54.706

Tabla 6- Resultados del vocabulario controlado GeneOntology (GOs) para cada secuencia y su e-value correspondiente.

GO process		e-value
1)Response to abscisic acid stimulus	KS-type dehydrin SLTI629) G. max	2,8E-2
2)Water deprivation	IAA7- auxin resistent 2-transcription factor	3,00E-82
3)response to heat, to water deprivation	HSP81-2	9,00E-108
4)response to cold	AP2 domain-containing transcriptionfactor	1,00E-16
5)early responseto dehydration 14	ERD14	1,00E-08
6)response to salt stress,aquoporin expression to environmental stress	GAMMA TIP2	4,00E-95

Tabla 7- Respaldo por Pfam, se muestran los dominios o familias a los que corresponde cada contig candidato y su correspondiente e-value.

Pf am	e-value
CDD|84648 pfam00257, Dehydrin, Dehydrin..	42 4e-05
gnl|CDD|66035 pfam02309, AUX_IAA, AUX/IAAfamily. Transcription ..	216 2e-57
gnl|CDD|84591 pfam00183, HSP90, Hsp90 protein..	38 6e-04
gnl|CDD|85066 pfam00847, AP2, AP2 domain. This 60 amino acid res...	68 8e-13
Gnl|CDD|84648 pfam00257, Dehydrin, Dehydrin..	42 3e-05
gnl|CDD|84627 pfam00230, MIP, Major intrinsic protein. MIP(Majo…	116 1e-27

4.2 Método automatizado: PLAN

Como ya se explico antes, el PLAN Server fue utilizado para respaldar las secuencias que fueron anotadas y procesadas en forma "manual". En el PLAN fueron introducidos aquellos contigs que contenían microsatélites y que fueron obtenidos a través del programa de ensamblado CAP3, con un total de **570 secuencias**.

Mediante el proceso manual de anotación, estas secuencias dieron como resultado a los 6 loci candidatos funcionales asociados a estrés hídrico ya nombrados en la **tabla 4** y que fueron evaluados experimentalmente; con el método automatizado se obtuvieron los siguientes resultados de respaldo para las secuencias escogidas:

Tabla 8- Muestra los resultados obtenidos por el método automatizado PLAN Server para los 6 marcadores moleculares funcionales elegidos a ser probados experimentalmente. Se detalla el Contig y la descripción proveniente del GOntology.

Lau1	*Contig75*	*No hits found*
Lau2	Contig951	IAA4/5: Auxin-induced protein IAA4-Pisum sativum
Lau4	**Contig2193**	**hsp90: heat shock protein Hsp90-response to stress**
Lau5	Contig2595	No hits found
Lau6	Contig2764	No hits found
Lau7	**Contig3215**	Water transport- water channel activity

Mediante la utilización del método automatizado se obtuvieron otras descripciones no obtenidas por el método manual. Un ejemplo interesante a destacar es la anotación encontrada para el **contig 1648**, la cual corresponde con una **MAPKK kinasa,** la cual interviene en la cascada de respuesta a estrés en las plantas; para los **contigs 3353**, **3603** y **807** tenían anotación vinculada a la familia de proteínas **USP** (*universal stress protein*) lo cual puede ser referida a plantas o bacterias.

4.3 Extracción y cuantificación de ADN

La extracción de ADN se realizó a partir de material verde de los 7 genotipos en estudio, mediante el protocolo de extracción antes descripto (FAO/IAEA). La cuantificación del ADN extraído se realizó mediante geles de agarosa 1% con tinción de bromuro de etidio. La concentración final del ADN genómico fue ajustada a 100 ng/μl.

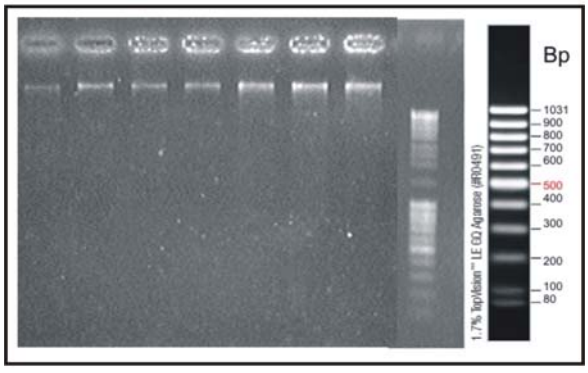


Fig.5- Gel de agarosa 1%, cuantificación de ADN genómico. Marcador de peso molecular: Mass Ruler de 1 Kb.

4.4 Amplificación de los microsatélites

La información molecular que permitirá identificar cada uno de los genotipos en estudio se obtuvo mediante la amplificación de los microsatélites seleccionados (funcionales y anónimos) como fue detallado anteriormente. EL ajuste de la reacción de PCR para soja se realizó a partir de las condiciones utilizadas por Roa et *al*. (2005), generando las modificaciones necesarias para la obtención de productos específicos para cada uno de los loci evaluados, tomando en cuenta la variante incluida para el marcado con fluorescencia (primer universal M13). En este caso fue posible llevar a cabo la reacción de PCR anidada y obtener productos específicos (Tabla 9).

Tabla 9- Reacción de PCR utilizada para amplificar microsatélites en soja.

REACCIÓN AJUSTADA SEGÚN ROA ET AL.(2005)			
MIX		CICLADO (*ROA ET AL*. 2005 + *SCHUELKE ET AL*.2000)	
	µL /TUBO		
H20	7,58	94ºC_3 min	
BUFFER (10X)	1,5	94ºC_30 seg	
MGCL2 (50 MM)	0,33	54ºC_1 min	35 CICLOS
DNTPS (10MM)	0,22	72ºC_2 min	
M13 (10 NM)	0,11		
REVERSE (10NM)	0,11	94ºC_30 seg	
FORWARD+M13 (10NM)	0,05	53ºC_45 seg	8 CICLOS (M13)
TAQ (5U)	0,1	72ºC_45 seg	
ADN	1	72ºC_2 min	
VOLUMEN FINAL	11		

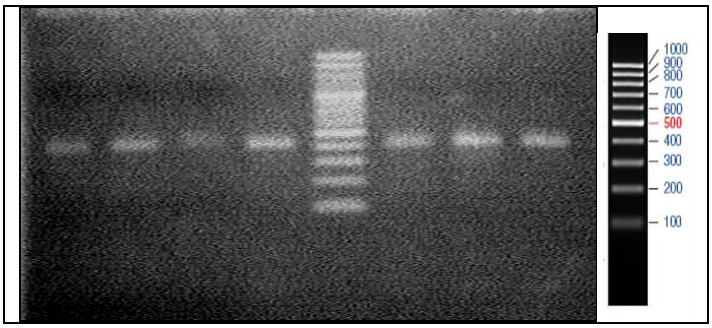


Fig.6- Gel agarosa 2% con tinción de bromuro de etidio para la separación de productos de PCR. Juego de primers: **Q6XPS8**, marcador molecular funcional de base de datos ICRISAT. Se utilizaron los 4 genotipos de soja antes mencionados. Marcador Gen Ruler: 100pb.

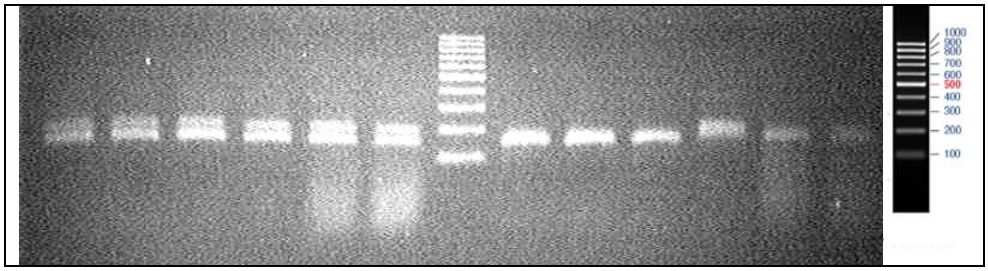


Fig.7- Gel agarosa 2% con tinción de bromuro de etidio para la separación de productos de PCR. Juego de primers: **LAU1** (izq. Del marcador) y **LAU2** (derecha del marcador) marcadores moleculares funcionales que fueron diseñados. Se utilizaron los 6 genotipos de soja antes mencionados. Marcador Gen Ruler: 100pb.

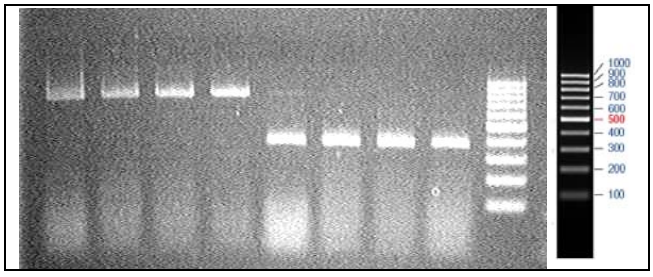


Fig.8- Gel agarosa 2% con tinción de bromuro de etidio para la separación de productos de PCR. Juego de primers: **LAU4** (primeros 4 posillos) y **LAU5**, marcadores moleculares funcionales que fueron diseñados. Se utilizaron los 4 genotipos de soja antes mencionados. Marcador Gen Ruler: 100pb.

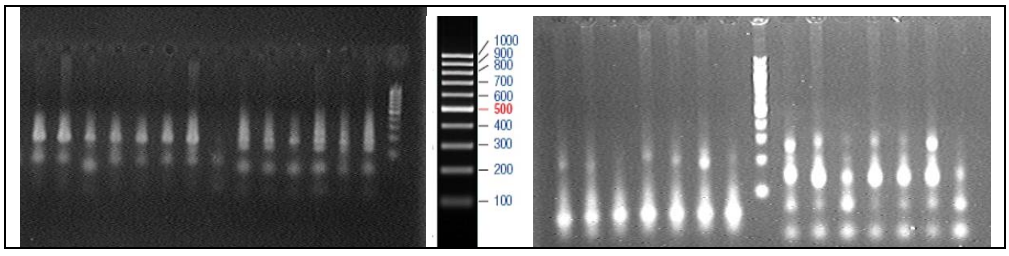


Fig.9- Gel agarosa 2% con tinción de bromuro de etidio para la separación de productos de PCR. Juego de primers: **Anónimos** (satt), marcadores moleculares obtenidos de bibliografía con los genotipos ya mencionados. Marcador Gen Ruler: 100pb.

4.5 Detección de los microsatélites

A partir de los resultados visualizados en los geles de agarosa (teniendo en cuenta la especificidad de los productos) se establecieron los parámetros de corrida en el secuenciador automático ABI PRISM 310 Genetic Analyzer. Los parámetros de corrida fueron los siguientes: tiempo de inyección 5", voltaje de la electroforesis 15 kv, tiempo de corrida 25 minutos (dependiendo del tamaño alélico) y temperatura de corrida 60ºC. Para cada SSR obtenido, el tamaño de los alelos fue determinado con el programa **Genotyper® v 3.7**.

En caso de existir diferentes perfiles en los picos identificados en un electroferograma se realizaron correcciones manuales en los valores que el software utiliza por defecto en comparación con estándares de peso molecular, a efectos de eliminar posibles errores en la determinación automática de los tamaños.

Luego de evaluados los resultados fueron descartados aquellos que superaban los 500 pb por una cuestión de resolución, por este motivo fueron descartados 3 marcadores funcionales (**LAU4**, **Q93WZ6** y **Q41111**). Si estos marcadores fueran imprescindibles se debería buscar una alternativa para la resolución de estos fragmentos, como por ejemplo el uso de geles de poliacrilamida y el genotipado de los mismos que de todas formas podrían dar problemas por la complejidad de dichos microsatélites.

En las tablas que se muestran a continuación se detallan los tamaños alélicos para los marcadores utilizados:

Tabla 10- Tamaño de los alelos en pares de bases determinados a través del programa Genotyper para marcadores anónimos y su genotipo correspondiente.

	satt_173	satt_175	satt_42	satt_577	satt_177	satt_231	satt_226	satt_324	satt_534	satt_70
Cultivar	(pb)	(pb)	(pb)	(pb)	(pb)	(pb)	(pb)	(pb)	(pb)	(pb)
DM4200	213	190	140	133	133	232	337	245	270	193
NM55R	213	181	140	136	125	259	343	242	270	163
NM70R	213	181	140	136	125	232	355	254	264	163
RA514	213	205	140	136	125	232	343	242	273	163
RA518	213	181/190	140	136	125	235	337/355	242	273	190
TJS2055	213	181	140	133	125	232	343	245	246	163
TJS2178R	222	190	140	127/130	125/133	232	343	245	273	163

Tabla 11- Tamaño de los alelos en pares de bases determinados a través del programa Genotyper para marcadores funcionales y su genotipo correspondiente.

	LAU_1	LAU_2	LAU_5	LAU_6	LAU_7	Q43453	PI27_ARATH	Q6XPS8
CULTIVAR	(pb)	(pb)	(pb)	(pb)	(pb)	(pb)	(pb)	(pb)
DM4200	140	135	369	176/218	374	285/289	326/329	405
NM55R	140	135	369	176/218	374	285/289	325/328	405
NM70R	140	135	369	176/218	374	285/289	325/328	405
RA514	139/142	135	319/369	176/218	374/398	285/289	325/328	405
RA518	139/142	135	319/369	176/218	374/398	285/289	325/328	405
TJS2055	139/142	135	319/369	176/218	374/398	285	325/328	405
TJS2178R	139/142	135/141	319/369	176/218	374/398	285/289	325/328	405

En las imágenes siguientes se muestran algunos electroferogramas obtenidos, para marcadores anónimos; funcionales diseñados manualmente y funcionales de bases de datos (ICRISAT). En la escala superior (horizontal) se señala el tamaño de los alelos en pares de bases (pb) y la escala vertical que se sitúa en la derecha indica la escala de emisión de fluorescencia presente, se considera que si la señal se encuentra entre 1000 y 5000 es un indicador confiable.

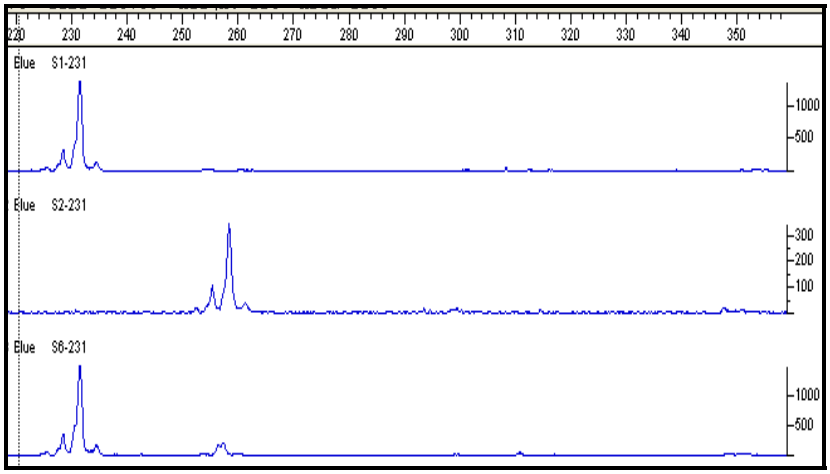


Fig.10- Electroferograma ABI software Genotyper para marcador anónimo satt_231.

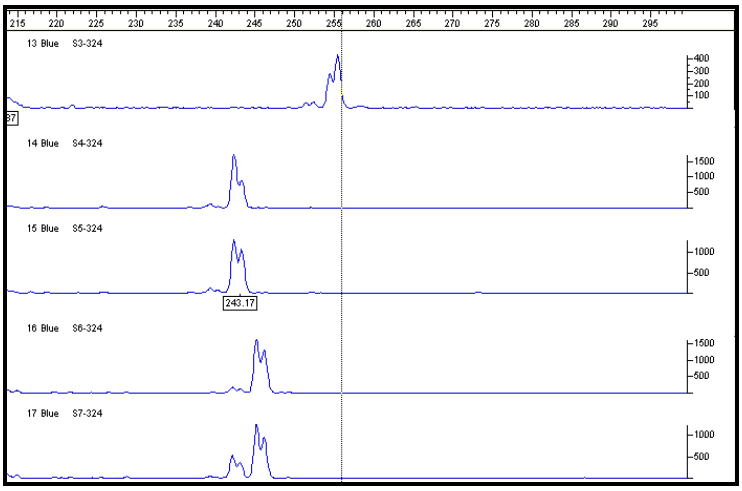


Fig.11- Electroferograma ABI software Genotyper para marcador anónimo satt_324.

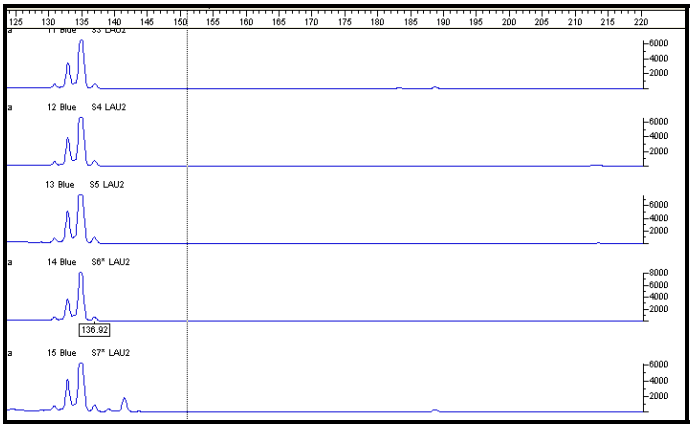


Fig.12- Electroferograma ABI software Genotyper para marcador funcional LAU 2.

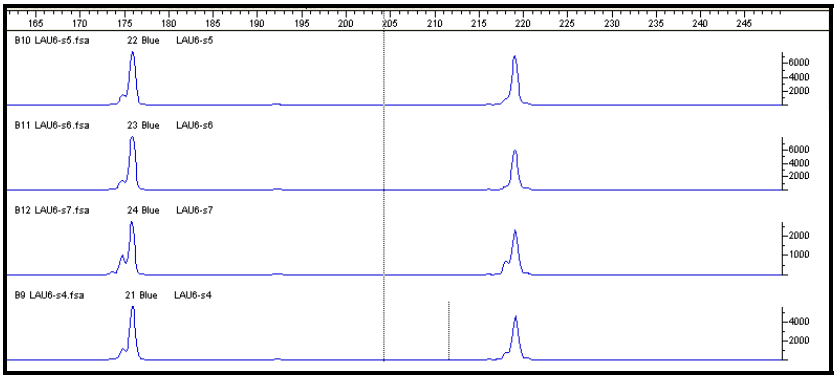


Fig.13- Electroferograma ABI software Genotyper para marcador funcional LAU 6.

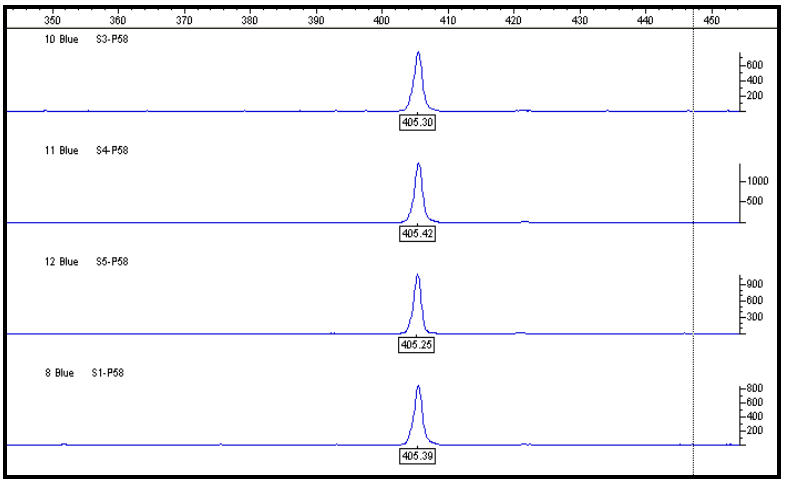


Fig.14- Electroferograma ABI software Genotyper para marcador funcional Q6XPS8 (ICRISAT).

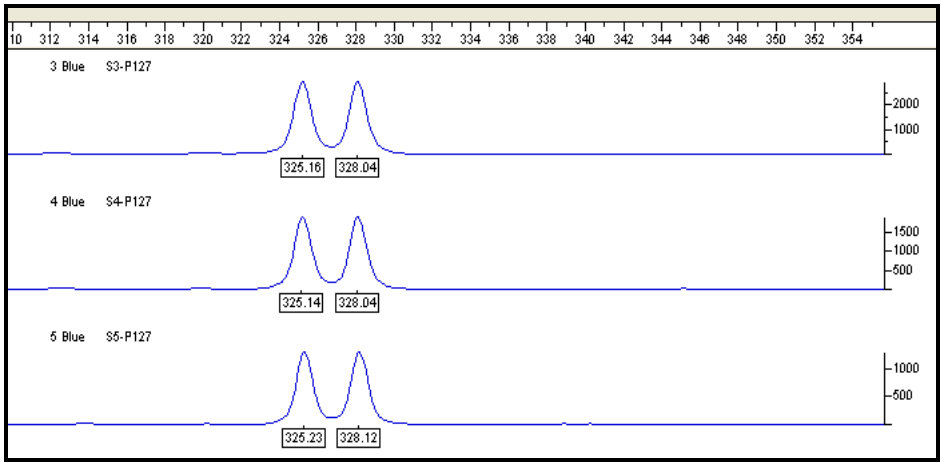


Fig.15- Electroferograma ABI software Genotyper para marcador funcional P127_ARATH (ICRISAT).

4.6 Análisis de datos: PIC

El análisis de PIC para marcadores anónimos y funcionales realizado mediante el ***Cervus 3.0.3***., el mismo que ya fue descrito anteriormente, fue utilizado para medir el nivel de polimorfismo entre variedades, valores de 0 fueron tomados como monomórficos y valores cercanos a 1 serán para aquellos alelos altamente polimórficos. A continuación se detallan los mismos.

Tabla 12- Resultado obtenido del análisis de PIC (contenido de información polimórfica), heterocigosidad observada y heterocigosidad esperada para todos los marcadores seleccionados mediante la aplicación de software Cervus 3.0.3. y para los 7 genotipos utilizados (N); k=variantes alélicas.

Locus	*k*	*N*	*H Obs.*	*H Esp.*	*PIC*
Satt173	2	7	**0.000**	0.264	0.215
Satt175	3	7	0.143	0.648	0.523
Satt42	1	7	**0.000**	**0.000**	**0.000**
Satt577	4	7	0.143	0.626	0.520
Satt177	2	7	0.143	0.363	0.280
Satt231	3	7	**0.000**	0.484	0.406
Satt226	3	7	0.143	0.626	0.517
Satt324	3	7	**0.000**	0.659	0.530
Satt534	4	7	**0.000**	0.747	0.641
Satt70	3	7	**0.000**	0.484	0.406
Lau1	3	7	0.571	0.703	0.580
Lau2	2	7	0.143	0.143	0.124
Lau5	2	7	0.571	0.440	0.325
Lau6	2	7	1.000	0.538	0.375
Lau7	2	7	0.571	0.440	0.325
P127	4	7	1.000	0.670	0.547
Q6XPS8	1	7	**0.000**	**0.000**	**0.000**
Q43453	2	7	0.857	0.527	0.370

La **H_O** es el número relativo de individuos heterocigotos para cada locus que se encuentra en la población estudiada y la **H_E** es la frecuencia relativa que se debería observar luego de apareamientos panmícticos con las mismas frecuencias génicas observadas en la población. **(43)**

Estos dos últimos parámetros son una función de la proporción de loci

polimórficos, el número de alelos por locus polimórfico, y las frecuencias alélicas. [44]

El valor de PIC varió entre 0.0 y 0.641, con un PIC promedio de 0.3714. El análisis se hizo para 7 genotipos diferentes, con un total de 18 alelos y un número promedio de alelos por locus de 2.56.

La heterocigosidad esperada (H_E) promedio fue de 0.4646, mientras que la heterocigosidad observada (H_O) se encuentra por debajo de la esperada para la mayoría de los casos. En especies autógamas como soja que se reproducen por autofecundación y donde el grado de alogamia es muy bajo la variación genética se distribuye en un gran número de genotipos homocigóticos y un solo ciclo de selección para un carácter específico agotará la libre variabilidad genética de la población.

4.7 Análisis de datos: análisis discriminante entre variedades

La obtención de una matriz de identificación varietal (ver matriz en **anexo 2**) requiere la comparación de diversas muestras de referencia y el ajuste de un procedimiento de análisis discriminante basado en información de los marcadores anónimos y funcionales para poder diferenciar entre los 7 cultivares de soja. Para dicho análisis fue utilizado el método K–NN (K nearest neighbor), implementado dentro del procedimiento **DISCRIM** de SAS/ STAT® 9.1 [SAS Institute 2002-2003].

Para la construcción del modelo de clasificación, una base de datos con información de microsatélites analizados para un grupo de muestras es tratada como un conjunto de entrenamiento para ajustar un modelo de clasificación basado en el algoritmo K-NN. Cada observación incluida en el conjunto de entrenamiento tiene información sobre el mismo conjunto

de variables y está asociada a una clase (identidad varietal en este caso). [45]

En este análisis fueron considerados la totalidad de los datos moleculares obtenidos para cada uno de los genotipos estudiados. Los tamaños alélicos hallados para cada loci estudiado fueron tomados como variables y los 7 genotipos como atributos para componer el conjunto de entrenamiento, con el cual fue ajustado el método de calificación antes mencionado.

A efectos de optimizar la clasificación desde el punto de vista de la identificación de las variedades se seleccionó un subconjunto de variables. La elección de las variables se realizó mediante el procedimiento **SAS PROC STEPDISC®**, el cual categoriza las variables según el grado de contribución a la clasificación. Este, asigna valores para el estadístico F según el grado de información aportado por la variable en cuanto a discriminar entre las clases predefinidas, es decir que selecciona aquellas variables que maximizan las diferencias entre las observaciones presentes en la población en estudio. El método usado dentro del sistema SAS (proc stepdisc) para calcular el valor F para las n variables seleccionadas se denomina "forward" (paso hacia delante), donde la variable de mayor valor F es seleccionada en cada iteración.

Una vez incorporada una nueva variable al modelo, el valor F es recalculado dentro del resto de las variables -esta vez sin la variable seleccionada anteriormente- quedando seleccionada de igual forma la segunda, repitiendo el procedimiento en sucesivas iteraciones hasta incorporar un total de n variables que contribuyan mayormente a la clasificación; de esta forma se genera un subgrupo de variables que maximizan la discriminación entre clases a la población.

El algoritmo K-NN fue corrido nuevamente, utilizando esta vez al subgrupo de variables seleccionado por SAS PROC STEPDISC, obteniendo los nuevos porcentajes de clasificación para cada una de las observaciones, basados en el conjunto de variables (marcadores) seleccionadas. Teniendo en cuenta la descripción anterior; los resultados del análisis de discriminante se detallen en el ***anexo 8.***

Los cuadros de dicha sección (anexo 8) se asignan con *a* y *b* debido a que se encuentran de a pares, son el resultado del análisis llevado a cabo y el resumen de dicho resultado. Los cuadros **1a y 1b** son el resultado del análisis SAS Proc stepdisc para los marcadores anónimos, *los más discriminantes fueron*: Satt 231-259, Satt 173-213, Satt 534-246, Satt 231-235, Satt 175-205, Satt 534-270.

Los cuadros **2a y 2b** presentan los resultados del estudio de clasificación para los datos calibrados, como resumen del método de resustitución usando el algoritmo K-NN.

Para el análisis discriminante con todas las variables, el resultado fue el "*overfitting*", un error que se produce cuando hace falta el ajuste del algoritmo utilizado en cuanto a su capacidad predictiva al ser aplicado a nuevas variables no incluidas en el conjunto de entrenamiento; se le llama "sobre ajuste" o "overfitting" y afecta a los algoritmos cuando el número de variables predictivas utilizadas es demasiado grande con respecto al tamaño de la población en estudio. Para evitar este problema se deben utilizar técnicas adicionales para seleccionar un número menor de variables de alto valor discriminante, a efectos de mejorar la capacidad del modelo ajustado para ser generalizable a futuros casos de uso. **(46)**

El paso siguiente fue el de ampliar el **set de datos** para poder hacer correr el programa y evitar los errores, este *"clonado"*, consiste simplemente en copiar los datos por duplicado o triplicado para aumentar dicho set y tener un resultado mas confiable.

Los cuadros **3a y 3b** son el resultado del análisis SAS Proc stepdisc para los funcionales, donde los más discriminantes fueron: Lau1-139, Lau2-141, P127-325, Q53-289.

En los cuadros **4a y 4b** se observan los resultados obtenidos a partir del análisis discriminante para funcionales, en el cual se observa el 57% de error.
Por último, en los cuadros **5a y 5b** también se muestra el error del 57% y además se muestra el error de un 100% para las cuatro variedades en las que se equivocó (NM55R, NM70R, RA514 y RA518) de un total de 7. La diferencia con el par anterior es que el último es la validación cruzada por vecino más cercano. Del uso combinado de marcadores anónimos y funcionales se obtuvo que los más discriminantes son los marcadores anónimos.

4.8 Actualización de secuencias anotadas

La localización de secuencias para los marcadores moleculares funcionales (Lau 1, 2, 5, 6 y 7) fue realizada mediante la web: **www.phytozome.net/soybean** (figura 20) en la cual se pueden ubicar los marcadores moleculares en el genoma de soja y permite así tener una noción de que es lo que esta amplificando ese marcador, este primer borrador de la secuenciación completa de *Glycine max* fue obtenida en el 2008 por el Departamento de energía estadounidense en conjunto con el **JGI** (Joint Genome Institute).

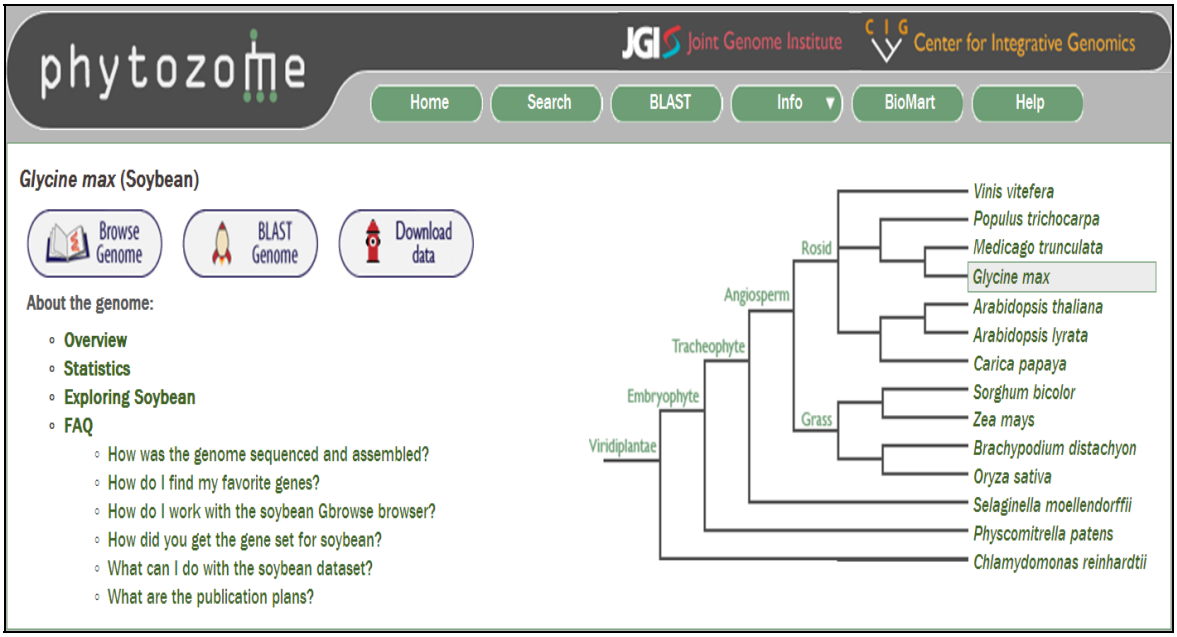


Fig.20- Imagen de la portada de www.phytozome.net/soybean.

Para el caso de **Lau1**; el *primer forward* de este marcador coincidió exactamente en la región codificante (CDS) para una dehidrina, proteína que se expresa en las plantas en situaciones de estrés hídrico. La región se puede observar pintada en amarillo entre dos secuencias cortas de 5´-UTR de 60 y 28 pares de bases respectivamente. Haber detectado la secuencia codificante es un paso muy importante en la anotación

funcional de este marcador, candidato asociado con un gen posiblemente vinculado con respuesta a estrés abiótico.

Para el caso de **Lau2**; el *primer reverse* coincidió justo sobre una región 3´-UTR de 375 pares de bases, la anotación de este marcador corresponde con un factor de transcripción. Para **Lau5**; el marcador molecular funcional corresponde a un factor de transcripción (AP domain) que esta vinculado a temperaturas extremas. El *primer forward* se ubicó sobre una región 3´-UTR de 973 pares de bases. Por último, para los marcadores moleculares funcionales, **Lau6 y Lau7**; los cuales se corresponden a una dehidrina y a un factor de transcripción respectivamente, ambos *primers reverse* se ubican en regiones codificantes del genoma de soja. Los detalles de las secuencias se pueden ver en el **anexo 3** para cada uno de los marcadores.

4.9 Actualización de marcadores funcionales DB-ICRISAT

Por el mismo procedimiento anterior fue realizada la localización de los marcadores funcionales obtenidos de la base de datos del I-CRISAT, el detalle de las secuencias se puede observar en el **anexo 4.**

A grandes rasgos se puede ver que los tres marcadores, los cuales corresponden a un ARNm de un gen inducido por estrés, una acuaporina y un gen inducido bajo estrés por sequía respectivamente, caen en regiones codificantes del genoma de soja.

4.10 Actualización de marcadores anónimos (satt)

Para el caso de los marcadores anónimos fue tomado en cuenta la ubicación de estos en el genoma de soja, esto tiene que ver con que

estos marcadores no estaban previamente relacionados con ninguna característica fenotípica, pero mediante la utilización del blast de estos primers en la base de datos del ***phytozome*** se pudo observar que de los 10 marcadores, solamente 2 de ellos no tenían ningún tipo de anotación funcional (**satt 42 y satt 226**), el resto, sorprendentemente se encuentran ubicados en regiones codificantes del genoma de soja, dentro de un gen o muy cerca de genes con alguna función específica para los que pudiera estudiarse a futuro su posible vinculación con la respuesta a estrés u otra función adaptativa conocida. Teniendo en cuenta esto; en un futuro pueden servir para ser utilizados como marcadores funcionales y posibles descriptores de variedades vinculados a alguna característica de interés particular, esto será discutido más adelante. En más detalle, se puede decir que el **satt 173** se encuentra muy cerca de una pectinesterasa (enzima responsable de la dimetilación de residuos galacturonil en pectinas y juega un rol importante en el metabolismo de la pared celular). El **satt 175**, se encuentra cerca de una *metiltransferasa*. Trabajos de **Vernon y Bohnert** muestran que mecanismos inducidos en plantas por estrés osmótico y salinidad revelan la presencia de esta enzima, la *metiltransferasa*, que cataliza el primer paso en la biosíntesis de alcohol pinitol de azúcar cíclico, este alcohol es muy abundante en plantas bajo estrés salino y osmotolerancia. **(46.1)**

Para el caso del **satt 577,** este se encuentra muy próximo a una HSP40 (*heat shock protein*) y a una proteína transmembrana rica en leucina. El **satt 177,** se encuentra muy próximo a una proteína de la superfamilia ascorbato oxidasa y de una proteína de la familia NPH3 (familias de genes relacionados a respuestas fototrópicas). El **satt 231** muy próximo a una ribonucleasa III, mientras que el **satt 324** se encuentra dentro de un gen con actividad de oxidoreductasa, CueO-multicopper oxidase (es

parte del operón de la regulación de cobre en *E.coli* y se expresa bajo condiciones de estrés por cobre, teniendo actividades oxidoreductasa). El **satt 534** se encuentra muy próximo a un factor de transcripción (WRKY) y por último; el **satt 70** se encuentra dentro de un gen que se corresponde con una proteína de transporte de membrana. Los detalles de secuencia se visualizan en el **anexo 5**.

5 DISCUSIÓN

5.1 Diseño de marcadores

Los ESTs son un producto de un tipo de secuenciación masiva aplicable en numerosos organismos, lo que ha permitido generar millones de secuencias relativamente cortas -pasada simple de secuenciación- integradas y explorables en bases de datos establecidas *ad hoc* (dbEST). Una vasta estrategia computacional ha sido desarrollada para organizar y analizar a pequeña y gran escala datos de ESTs para el descubrimiento de genes, transcriptos y análisis de SSR así como la anotación funcional de genes hipotéticos. **(47)**

Hay varios pasos a seguir para llevar a cabo el correcto análisis de grandes cantidades de ESTs. En nuestro caso fueron utilizadas las herramientas ya descriptas y ahora serán discutidos los resultados obtenidos y los posibles cuellos de botellas de ESTs.

Errores asociados con el procesamiento de ESTs:

Su alta posesividad los hace más susceptibles a errores. Vectores y secuencias repetidas raramente son suprimidas durante el pre procesamiento de ESTs, generalmente la calidad de lectura de bases en una secuencia individual de EST es pobre inicialmente (20% o 50-100pb), mejora gradualmente y disminuye otra vez hacia el final de la secuencia. **(47)**

La calidad de secuencia es significativamente mejor en el medio ("highly informative length" = largo sumamente informativo). Pero la sobre-representación y la sub-representación de transcriptos seleccionados son problemas inherentes con datos de ESTs debido a la variabilidad de protocolos usados en su generación.

Los artefactos del secuenciado, bases repetidas (específicamente G y T) y la baja calidad de secuencia, son los más frecuentes errores observados en ESTs. Esto puede deberse a una posible contaminación por vectores, adaptadores y secuencias quiméricas, como también de fragmentos de ADN genómico. [47]

Análisis de ESTs:

Las secuencias de ESTs extraídas de bases de datos son pre-procesados y ofrecen ESTs únicamente de alta calidad, habiendo sido filtradas y eliminadas las secuencias repetidas, contaminantes y secuencias complejas. Consiguientemente los ESTs de alta calidad son agrupados dentro de "clusters o conglomerados" basado en la similaridad de secuencias.

La secuencia consenso de máxima información es generada por el ensamblado de estos clusters, cada uno de los cuales puede representar un gen hipotético. Este paso sirve para elongar la secuencia, para elegir la información de varias secuencias cortas de ESTs simultáneamente. [47]

La búsqueda de similaridad en bases de datos son realizadas subsecuentemente contra bases de datos de ADN y la posible funcionalidad es asignada para cada "query" (término de búsqueda) correspondiente a una secuencia si se encontró una similaridad significativa. Adicionalmente la secuencia consenso puede ser traducida a un péptido hipotético y entonces comparada contra una base de datos de proteínas. La anotación funcional de proteínas incluye análisis de dominios y motivos. [47]

Conglomerado y ensamblado de ESTs:

La formación de cluster de ESTs tiene como objetivo seleccionar el traspaso de ESTs de un mismo transcripto a un solo gen dentro de un único cluster al disminuir la redundancia.

El cluster de ESTs es un dato fragmentado, el cual puede ser consolidado usando la información de secuencia del gen, de tal forma que todo lo expresado surgido de un solo gen, es agrupado en una sola clase indexada y cada clase contiene información para solo ese gen particular. [47]

Hay dos enfoques para llevar a cabo la formación de clusters de ESTs; uno "*riguroso*" y uno "*relajado*". El método de *clustering riguroso* es conservador, usa agrupamientos de ESTs, que resultan en un cluster relativamente exacto, que genera secuencias cortas con bajos cubrimientos de genes expresados. En contraste, el *clustering relajado* es un método más liberal y el alineamiento de ESTs incluye repetidos de baja calidad, genera menos exactitud pero secuencias consenso más largas. Consecuentemente, este da una mejor cobertura de genes expresados y alternativamente transcriptos procesados, pero se debe tener en cuenta que pueden ser incluidos en el análisis, genes parálogos no deseados para ser incluidos en el cluster. [47]

Los programas de computación Phrap y CAP3 son entre los más usados para el conglomerado y ensamblado de secuencias. CAP3 supera otros programas similares produciendo secuencias consenso de alta fidelidad y manteniendo un alto nivel de sensibilidad, los cuales efectivamente manejan errores de secuenciado. **Wang et *al*.** (2004) muestra dos tipos de posible fuente de error para el procesamiento por parte de CAP3, donde ESTs de un mismo gen no forman un cluster (tipo I) y ESTs de genes diferentes son incorrectamente agrupados juntos (tipo II). Ellos propusieron una aproximación estadística original para una más exacta estimación del perfil verdadero del cluster de genes.

Búsqueda de similaridad:

Las secuencias consenso (correspondientes a genes hipotéticos) pueden ser obtenidas a partir del ensamblado de ESTs, mientras que su posible función puede ser asignada a través de la anotación comparativa (transferencia de información) asignada en base a los resultados de una búsqueda en bases de datos por similaridad. Diferentes aplicaciones del algoritmo BLAST del NCBI sirven como una herramienta universal para la búsqueda de similaridad. BLASTx traduce la secuencia consenso de ESTs dentro de productos de proteínas en seis marcos de lectura seguido de la comparación con bases de datos de proteínas. En forma adicional, es posible investigar los dominios de proteínas al seleccionar el CDD (conserved domain database) y el COG (cluster of orthologous groups) usando RPS- Blast. **(47)**

Traducción conceptual de ESTs:

Los ESTs pueden ser correlacionados con términos de la anotación de proteínas, convirtiéndose estos en los mejores moldes para la identificación de dominios y motivos, al estudiar la localización de proteínas y asignar ontología al gen (GOs). **(47)**

Anotación funcional:

Una vez obtenido el polipéptido hipotético, su función puede ser predicha mediante el alineamiento contra secuencias de proteínas no redundantes. Las secuencias de proteínas son mejores modelos para la anotación funcional, particularmente para la construcción de alineamientos múltiples de secuencias, perfiles y generación de HMM, análisis filogenéticos y análisis de dominios y motivos usando Pfam y SMART. **(47)**

5.2 Respaldo automático: PLAN Server

El PLAN Server permitió respaldar la búsqueda manual de las secuencias de interés, pero en tres casos particulares no encontró similaridad de secuencia con la base de datos. Pero nos devolvió otros resultados que no habíamos tenido en cuenta para otras secuencias, también relacionadas a estrés (por ejemplo; el **contig 1648**, el cual corresponde a una **MAPKK kinasa**). En este trabajo no se tomó la opción de utilizar los resultados PLAN Server como posibles marcadores adicionales, este paso fue hecho para poder respaldar únicamente los marcadores elegidos, obtenidos del procesamiento manual, alcanzando así un resultado positivo.

5.3 Detección de microsatélites

Las muestras son separadas dentro del capilar electroforéticamente y la detección de los fragmentos se realiza cuando los mismos atraviesan una "ventana" existente en el capilar, donde se combinan un emisor de luz láser y un detector de fluorescencia. Este sistema permite detectar los fragmentos amplificados de ADN que fueron marcados por incorporación de dNTPs unidos a fluorocromos que fueron incluidos en la reacción de PCR. El monitor de fluorescencia de tipo CCD detecta los espectros de emisión correspondientes a cada fluorocromo utilizado, interpretando las señales en función del tiempo de corrida como electroferogramas [User Manual ABI 310 Genetic Analayzer].

Los errores en la determinación de los tamaños alélicos pueden deberse al deslizamiento ("**slippage**") de la Taq polimerasa y/o al agregado de bases extras en el fin de cada fragmento durante la reacción de PCR, observándose picos no definidos o los "llamados picos sombra". **(48)**

Estos tipos de errores son detectables al observar la morfología de los picos, pero en el análisis automático no serían detectados como errores, por lo que este tipo de artefactos experimentales podrían ser tomados como un tamaño válido de alelo en un análisis no supervisado.
En caso de existir diferentes perfiles en los picos identificados en un electroferograma se realizaron correcciones manuales en los valores que el software utiliza por defecto en comparación con estándares de peso molecular, a efectos de eliminar posibles errores en la determinación automática de los tamaños.

5.4 Análisis de PIC

El PIC es una medida de la diversidad genética y provee una estimación del poder discriminatorio de cada locus, teniendo en cuenta el número de alelos y sus frecuencias relativas en cada población. **(40)**
En este estudio los valores de PIC se encuentran desde valores de 0, o sea, monomórfico para 2 variedades (**satt 42** y **Q6XPS8**) y valores de 0.2 a 0.6 para las variedades restantes. El **PIC promedio** fue de 0.37. Se observan en la ***tabla 12*** valores de PIC más elevados en el caso de los marcadores anónimos que para los marcadores funcionales y también se observan más variantes alélicas en el caso de los marcadores anónimos. Los demás parámetros; la heterocigosidad observada y la esperada, son una función del número de alelos por locus polimórfico y las frecuencias alélicas. El análisis se hizo para 7 genotipos diferentes, con un total de 18 alelos y un número promedio de alelos por locus de 2.56.

5.5 Análisis discriminante

En este análisis fueron considerados la totalidad de los datos moleculares obtenidos (tanto para marcadores anónimos como funcionales) para cada uno de los genotipos estudiados. Los tamaños alélicos hallados para cada loci estudiado fueron tomados como variables y los 7 genotipos como atributos para componer el conjunto de entrenamiento, con el cual fue ajustado el método de calificación antes mencionado.

En este análisis (SAS Proc Stepdisc) los resultados revelaron que de las 46 variables utilizadas (obtenidas del clonado de datos para mejor resolución) las 6 más discriminantes se encontraron dentro de los marcadores anónimos, esto concuerda con los valores de PIC. Luego se llevo a cabo un análisis (DISCRIM) con todas las variables, en la que el resultado fue el ***overfitting***, después de esto se amplió el set de datos y dio un resultado correcto. En una segunda instancia, se realizó el mismo procedimiento (SAS Proc Stepdisc) para las 46 variables y el resultado fue el mismo que al principio, los más discriminantes siguen siendo los marcadores anónimos. Para el análisis hecho solamente con los satt, el resultado vuelve a ser el mismo, además el mismo análisis (SAS Proc Stepdisc) solo para los marcadores funcionales y el resultado fue poco favorable; con un 57% de error para dos marcadores elaborados manualmente (Lau) y dos de la base de datos del I-CRISAT. También hubo un error de un 100% para discriminar 4 variedades (NM55R, NM70R, RA514 y RA518) del total de 7. De esto se puede concluir que; si bien en algunos casos no discrimina bien, tampoco confunde el set de datos, o sea, no dice o identifica a un cultivar que en realidad no es. Si bien el cometido de este trabajo no es discutir el polimorfismo para tan pocas variedades, porque no tendría sentido, si es

importante ver que fue posible discriminar correctamente entre todas las variedades, aunque fueran muy pocas, por lo que se puede pensar que el aumento en el número de variedades no daría problemas para dicho análisis. Siempre utilizando marcadores anónimos y funcionales.

5.6 Actualización de secuencias funcionales y anónimas

La tarea dedicada a la localización de los marcadores en el borrador del genoma de soja es un paso clave en este estudio, debido a que cuando se comenzó con el diseño de estos, no estaba aun disponible el borrador de dicho genoma para trabajar. Teniendo en cuenta los resultados de la localización de las secuencias para todos los marcadores; se puede establecer que, para el caso de los marcadores funcionales **Lau** y para los obtenidos a partir de la base de datos **I-CRISAT** el resultado fue muy positivo ya que todos cayeron en regiones codificantes del genoma de soja lo cual era decisivo para que fueran buenos descriptores a pesar de las pocas variedades utilizadas, estos marcadores permitieran hacer un *screening* sobre el genoma de *Glycine max* abarcando una distribución muy variada en los cromosomas de dicho organismo (Lau1-c19, Lau2-c20, Lau5-c10, Lau6-c4, Lau7-c13; Q43453-c17, P127ARATH-c6, Q6XPS8-c18), lo cual aumenta la capacidad discriminatoria de los marcadores.

En el caso del **Lau4**, el cual fue descartado debido a su tamaño de más de 500pb y sus *primers* no pegan en el genoma de soja; esto se pudo deber a que el primer cae en una región en la que no hay modelo de soja o también puede explicarse esto por una inversión, lo que no quita que pueda ser utilizado como un marcador funcional. Se podría resolver con fragmentos de restricción para encontrar el polimorfismo, de todos

modos fue descartado porque no se podían obtener sus valores de PIC ni de discriminante como en los otros marcadores; al igual que este también fueron descartados **Q93WZ6** y **Q41111**.

Otra causa posible pudo ser el ensamblado del contig, debido a que pudo generarse un artefacto, pero esto se puede descartar ya que la secuencia consenso se corresponde con una HSP81-2, que indica que el contig está bien ensamblado y lo convierte en un marcador génico, si cayera en una región de *repeat masker* se convertiría en genómico.
Para el caso de los **marcadores anónimos** el resultado fue positivo teniendo en cuenta que siendo estos los más discriminantes dentro del total, el proceso reciente de actualización de secuencias revelo que muchos de ellos sirven como marcadores funcionales ya que se los puede asociar con alguna característica funcional relacionada a estrés, por ejemplo el caso del **satt 577**, el cual se encuentra a 4.8 Kbp de una HSP40 (proteínas que protegen membranas y estructuras celulares bajo condiciones de estrés); otro caso particular se da para el **satt 534** que se encuentra a 31Kbp de un factor de transcripción (WRKY), el cual es estimulado bajo condiciones de estrés hídrico por procesos de deshidratación.

Otros 2 casos novedosos fueron los de los **satt 70** y **satt 324;** los cuales dejaron de ser marcadores genómicos para ser marcadores génicos ya que se encuentran dentro de genes, el primero dentro de una proteína de transporte de membrana y el segundo dentro de un gen que codifica para una oxido reductasa. En su gran mayoría, estos microsatélites anónimos se encuentran muy próximos (de 2.1 a los 31 Kbp) a genes de soja que tienen alguna función, ya se relacionada a estrés o no, lo cual los vuelve firmes candidatos a convertirse en

marcadores funcionales. Para estos, el espectro de cromosomas cubiertos dentro del genoma soja fue también muy amplio (satt173-c10, satt175-c7, satt42-c5, satt577-c14, satt177-c8, satt226-c17, satt231-c5, satt324-c18, satt534-c14, satt70-c14).

6 CONCLUSIÓN

Teniendo en cuenta los objetivos de este trabajo fue posible ajustar un sistema de genotipado para variedades de soja disponibles en Uruguay integrando marcadores anónimos y funcionales asociados con respuesta a estrés abiótico. También se logró generar una matriz de identificación para variedades de soja a partir de información de ambos tipos de marcadores (anónimos y funcionales), alcanzándose así, la meta de establecer materiales y protocolos de referencia para la utilización de marcadores moleculares como descriptores complementarios para el registro de nuevas variedades de soja y para la verificación de calidad genética en muestras de semillas.

Se puede decir que a partir de los resultados obtenidos en la clasificación de variedades de soja, se puede inferir que los microsatélites utilizados permitirían discriminar entre muestras correspondientes a diferentes accesiones cultivadas, con posibles aplicaciones en el contexto de programas de manejo de recursos genéticos para identificación y selección de genes vinculados con la respuesta a estreses abióticos en soja.

7 AGRADECIMIENTOS

Esta tesis ha sido posible gracias a muchas personas. Cada una de ellas ha aportado algo; sus conocimientos, su trabajo o su apoyo moral. Es imposible nombrarlos a todos en estas líneas pero mi sentimiento de profunda gratitud va dirigido a todas y cada una de ellas. En especial, al Dr. Fabián Capdevielle por la oportunidad, a todo el personal del laboratorio de la Unidad de Biotecnología de INIA (estación: Las Brujas), a la Lic. en Bioquímica Silvia Garaycochea por su incansable paciencia en el laboratorio y todas sus enseñanzas. A Héctor "Yuyo" Romero de la sección de biomatemáticas de Facultad de Ciencias por sus aportes en los conocimientos bioinformáticas. A los correctores externos: Inés Ponce de león y Fernando Alvarez-Valín. A mi familia. ***Gracias*!!!**

8 BIBLIOGRAFÍA

[1] Fulton,T.M., Van der Hoeven,R., Eannetta,N.T. and Tanksley,S.D. (2002) Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. Plant Cell, 14, 1457–1467.

[2] Kota,R., Rudd,S., Facius,A., Kolesov,G., Thiel,T., Zhang,H.,Stein,N., Mayer,K. and Graner,A. (2003) Snipping polymorphisms from large EST collections in barley (Hordeum vulgare L.). Mol. Genet. Genomics, 270, 24–33.

[3] Thiel, T., W. Michalek, R. K. Varshney, and A. Graner. (2003). Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). Theor. Appl. Genet. 106:411–422.

[4] Cullis, Chapter 6, 2004- Functional genomics. In: Plant genomics and proteomics. Published by Jhon Wiley & Sons, Inc. – New Jersey- USA. (1-218)

[4.1] Jacobs, D. I., R. van der Heijden, and R. Verpoorte (2000) Proteomics in plant biotechnology and secondary metabolism research. *Phytochem. Anal.* 11, 277–287. Cit in: Cullis, Chapter 6, 2004- Functional genomics.

[5] Ai-Guo et al. (2003) Characterization of soybean genomic features by análisis of its expressed sequence tags. Springer- Verlag- Theor Appl Genet (2004) 108: 903-913.

[5.1] Yu et al. (2002) A draft sequence of the rice genome (Oryza sativa L. ssp. indica). Science 296:79–92.

[6] Randy C. Shoemaker, Jessica A. Schlueter, and Scott A. Jackson (2008).Chapter6-Soybean Genome Structure and Organization. In: G. Stacey (ed.), *Genetics and Genomics of Soybean*, C_ Springer Science+Business Media, LLC 2008. Vol. 2 (ed: Gary Stacey) 405 páginas.

[7] Viviana Becerra V. y Mario Paredes C. (2000) Uso de marcadores bioquímicas y moleculares en estudios de diversidad genética. AGRICULTURA TÉCNICA (CHILE) 60 (3):270-281.

[8] Rocha E., Matic I., Taddei F. (2002) Over-representation of repeats in stress response genes: a strategy to increase versatility under stressful conditions? 1886–1894 Nucleic Acids Research, Vol. 30, N° 9.

[9] Morgante, M., M. Hanafey, and W. Powell. (2002). Microsatellites are preferentially associated with non repetitive DNA in plant genomes. Nat. Genet. 30:194–200.

[10] Boyer JS (1982) Plant productivity and environment. Science 218:443–448

[11] Wang WX, Vinocur B, Shoseyov O, Altman A (2001a) Biotechnology of plant osmotic stress tolerance: physiological and molecular considerations. Acta Hort 560:285–292

[12] Serrano R, Mulet JM, Rios G, Marquez JA, de Larrinoa IF, Leube MP, Mendizabal I, Pascual-Ahuir A, Proft M, Ros R, Montesinos C (1999) A glimpse of the mechanisms of ion homeostasis during salt stress. J Exp Bot 50:1023–1036

[13] Zhu JK (2001a) Plant salt tolerance. Trends Plant Sci 6:66–71

[14] Smirnoff N (1998) Plant resistance to environmental stress. CurrOpin Biotech 9:214–219

[15] Shinozaki K, Yamaguchi-Shinozaki K (2000) Molecular responses to dehydration and low temperature: differences and cross-talk between two stress signalling pathways. Curr Opin Plant Biol 3:217–223

[16] Zhu JK (2001b) Cell signaling under salt, water and cold stresses.
CurrOpin Plant Biol 4:401–406

[17] Vierling E, Kimpel JA (1992) Plant responses to environmental stress. Curr Opin Biotech 3:164–170

[18] Zhu JK, Hasegawa PM, Bressan RA (1997) Molecular aspects of osmotic stress in plants. Crit Rev Plant Sci 16:253–277

[19] Cushman JC, Bohnert HJ (2000) Genomic approaches to plant stress tolerance. Curr Opin Plant Biol 3:117–124

[20] Stockinger EJ, Gilmour SJ, Thomashow MF (1997) Arabidopsis thaliana CBF1 encodes an AP2 domain-containing transcriptional activator that binds to the C-

repeat/DRE, a cis-acting DNA regulatory element that stimulates transcription in response to low temperature and water deficit. Proc Natl Acad Sci USA 94:1035–40

[21] Bohnert HJ, Sheveleva E (1998) Plant stress adaptations—making metabolism move. Curr Opin Plant Biol 1:267–274

[22] Blumwald E (2000) Sodium transport and salt tolerance in plants. Curr Opin Cell Biol 12:431–434.

[23] Riechmann JL, Heard J, Martin G, Reuber L, Jiang C, Keddie J, Adam L, Pineda O, Ratcliffe OJ, Samaha RR. (2000) Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. Science 290:2105–2110.

[24] Ingram J, Bartels D (1996) The molecular basis of dehydration tolerance in plants. Annu Rev Plant Biol 47:377–403.

[25] Diamant S, Eliahu N, Rosenthal D, Goloubinoff P (2001) Chemical chaperones regulate molecular chaperones in vitro and in cells under combined salt and heat stresses. J Biol Chem 276:39586–39591.

[26] Mittler R (2002) Oxidative stress, antioxidants and stress tolerance. Trends Plant Sci 7:405–410.

[27] Vierling E (1991) The roles of heat-shock proteins in plants. Annu Rev Plant Biol 42:579–620.

[28] Hamilton III EW, Heckathorn SA (2001) Mitochondrial adaptations to NaCl. Complex I is protected by anti-oxidants and small heat shock proteins, whereas complex II is protected by proline and betaine. Plant Physiol 126:1266–1274.

[29] Thomashow MF (1998) Role of cold-responsive genes in plant freezing tolerance. Plant Physiol 118:1–7.

[30] Ceccardi TL, Meyer NC, Close TJ (1994) Purification of a maize dehydrin. Protein Express Purif 5:266–269.

[31] Roa M, Schlatter AR, Suárez EY, Acevedo A. (2005) Caracterización molecular de cultivares de soja (Glycine max L.) mediante microsatélites.

[32] Schuelke M. (2000) An economic method for the fluorescent labeling of PCR fragments. Nature Biotechnology 18233-234. DOI:10.1038/72708.

[33] Huang X. and A. Madan (1999) CAP3: A DNA secuence assembly program. Genome Res. 9: 868-877.

[34] Robinson A., Christopher G. Love, Jacqueline Batley, Gary Barker and David Edwards (2004).Simple sequence repeat marker loci discovery using SSR primer. Vol. 20 no. 9, pages 1475–1476.

[35] Rozen S. and Helen J. Skaletsky (2000) Primer3 on the www for general users and for biologist programmers. In: Krawetz S, Misener S (eds) Bioinformatics Methods and Protocols: Methods in Molecular Biology. Humana Press, Totowa, NJ, pp 365-386.

[36] La Rota M, Kantety R, Yu J, Sorrells M. (2005). Nonrandom distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat, and barley. BMC Genomics 6(1):23. DOI:10.1186/1471-2164-6-23.

[37] Zhang LY, Bernard M, Leroy P, Feuillet C, Sourdille P. (2005). High transferability of bread wheat EST-derived SSRs to other cereals. TAG Theoretical and Applied Genetics 111(4):677-687. DOI 10.1007/s00122-005-2041-5.

[38] Finn et al. (2007).The Pfam protein families database. Nucleic Acids Research, 2008, Vol. 36, Database issue D281–D288.

[39] He, Dai and Zhao. (2007) PLAN: a web platform for automating high-throughput BLAST searches and for managing and mining results. BMC Bioinformatics, **8**:53.

[40] Barcaccia G., 2003. Genet. Res. and Crop Evol. 50: 253-271.

[41] Capdevielle, F. (2001) Evaluation of a discriminant analysis procedure combining agronomic and molecular marker information for germplasm improvement in rice. MSc Thesis Series, Louisiana State University.

[42] Jeremy Schmutz, Jarrod Chapman, Uffe Hellsten, and Daniel Rokhsar, 2008, Chapter 7- Sequence and assembly of the soybean genome. **In:** Genetics and genomics of soybean. (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA). Vol. 2 (ed: Gary Stacey) 405 páginas.

[43] Yenen Villasmil-Ontiveros1, (2008) Genetic Diversity of Limonero Creole Breed Using Microsatellites Molecular Markers. Revista Científica, FCV-LUZ / Vol. XVIII, Nº 4, 415 – 423.

[44] Labate J.A., 2003. Crop Science. 43: 80-91.

[45] Capdevielle, F., Pinson, S., Oard, J. (2003b). ´Marker – assisted classification of Lemont x Tequing RILs into disease response groups: comparison of discriminate analysis and neural network algorithms, In Proceedings 3rd International Temperate Rice Conference, Ururguay.

[46] Tetko, I. Livingstone, D., Luik, A. (1995).Comparison of overfitting and overtraining.Journal of Chemical Information and Computer Sciences. 35(5): 826-833.

[46.1] Daniel M.Vernon' and Hans J.Bohnert. A novel methyl transferase induced by osmotic stress in the facultative halophyte Mesembryanthemum crystallinum. The EMBO Journal vol. 1 1 no.6 pp.2077 - 2085, 1992

[47] Shivashankar H. et al., 2006. A hitchhiker's guide to expressed sequence tag (EST) analysis. Briefings in Bioinformatics Advance Access published online on May 23, 2006.

[48] Ginot, F., Bordelais, I., Nguyen, S., Gyapay, G. (1996).´Correction of some genotyping errors in automated fluorescent microstellite analysis by enzymatic removal of one base overhangs.´ Nucleic Acid Research 24: 540-541.

9 ANEXOS

9.1 ANEXO 1: Matriz de tamaños alélicos para análisis de PIC

ID	*SATT173_A*	*SATT173_B*	*SATT175_A*	*SATT175_B*	*SATT42_A*	*SATT42_B*	*SATT577_A*	*SATT577_B*
DM4200	213	213	190	190	140	140	133	133
NM55R	213	213	181	181	140	140	136	136
NM70R	213	213	181	181	140	140	136	136
RA514	213	213	205	205	140	140	136	136
RA518	213	213	181	190	140	140	136	136
TJS2055	213	213	181	181	140	140	133	133
TJS2178R	222	222	190	190	140	140	127	130

SATT177_A	*SATT177_B*	*SATT231_A*	*SATT231_B*	*SATT226_A*	*SATT226_B*	*SATT324_A*	*SATT324_B*	*SATT534_A*
133	133	232	232	337	337	245	245	270
125	125	259	259	343	343	242	242	270
125	125	232	232	355	355	254	254	264
125	125	232	232	343	343	242	242	273
125	125	235	235	337	355	242	242	273
125	125	232	232	343	343	245	245	246
125	133	232	232	343	343	245	245	273

SATT534_B	*SATT70_A*	*SATT70_B*	*LAU1_A*	*LAU1_B*	*LAU2_A*	*LAU2_B*	*LAU5_A*	*LAU5_B*
270	193	193	140	140	135	135	369	369
270	163	163	140	140	135	135	369	369
264	163	163	140	140	135	135	369	369
273	163	163	139	142	135	135	319	369
273	190	190	139	142	135	135	319	369
246	163	163	139	142	135	135	319	369
273	163	163	139	142	135	141	319	369

LAU6_A	*LAU6_B*	*LAU7_A*	*LAU7_B*	*P127_A*	*P127_B*	*Q6XPS8_A*	*Q6XPS8_B*	*Q43453_A*	*Q43453_B*
176	218	374	374	326	329	405	405	285	289
176	218	374	374	325	328	405	405	285	289
176	218	374	374	325	328	405	405	285	289
176	218	374	398	325	328	405	405	285	289
176	218	374	398	325	328	405	405	285	289
176	218	374	398	325	328	405	405	285	285
176	218	374	398	325	328	405	405	285	289

9.2 ANEXO 2: Matriz de 0 y 1 para análisis de discriminante

CULTIVAR	*SATT173_213*	*SATT173_222*	*SATT175_181*	*SATT175_190*	*SATT175_205*	*SATT42_140*	*SATT577_127*	*SATT577_130*	*SATT577_133*	*SATT577_136*	*SATT177_125*	*SATT177_133*	*SATT231_232*	*SATT231_235*	*SATT231_259*	*SATT226_337*	*SATT226_343*
DM4200	1	0	0	1	0	1	0	0	1	0	0	1	1	0	0	1	0
NM55R	1	0	1	0	0	1	0	0	0	1	1	0	0	0	1	0	1
NM70R	1	0	1	0	0	1	0	0	0	1	1	0	1	0	0	0	0
RA514	1	0	0	0	1	1	0	0	0	1	1	0	1	0	0	0	1
RA518	1	0	1	1	0	1	0	0	0	1	1	0	0	1	0	1	0
TJS2055	1	0	1	0	0	1	0	0	1	0	1	0	1	0	0	0	1
TJS2178R	0	1	0	1	0	1	1	1	0	0	1	1	1	0	0	0	1

SATT226_355	*SATT324_242*	*SATT324_245*	*SATT324_254*	*SATT534_246*	*SATT534_264*	*SATT534_270*	*SATT534_273*	*SATT70_163*	*SATT70_190*	*SATT70_193*	*LAU1_139*	*LAU1_140*	*LAU1_142*	*LAU2_135*	*LAU2_141*	*LAU5_319*	*LAU5_369*
0	0	1	0	0	0	1	0	0	0	1	0	1	0	1	0	0	1
0	1	0	0	0	0	1	0	1	0	0	0	1	0	1	0	0	1
1	0	0	1	0	1	0	0	1	0	0	0	1	0	1	0	0	1
0	1	0	0	0	0	0	1	1	0	0	1	0	1	1	0	1	1
1	1	0	0	0	0	0	1	0	1	0	1	0	1	1	0	1	1
0	0	1	0	1	0	0	0	1	0	0	1	0	1	1	0	1	1
0	0	1	0	0	0	0	1	1	0	0	1	0	1	1	1	1	1

LAU6_176	*LAU6_218*	*LAU7_374*	*LAU7_398*	*P127_325*	*P127_326*	*P127_328*	*P127_329*	*Q6XPS8_405*	*Q43453_285*	*Q43453_289*
1	1	1	0	0	1	0	1	1	1	1
1	1	1	0	1	0	1	0	1	1	1
1	1	1	0	1	0	1	0	1	1	1
1	1	1	1	1	0	1	0	1	1	1
1	1	1	1	1	0	1	0	1	1	1
1	1	1	1	1	0	1	0	1	1	0
1	1	1	1	1	0	1	0	1	1	1

9.3 ANEXO 3: Actualización de anotación de marcadores funcionales diseñados manualmente

LAU1

phytozome
Glycine max

Name:	Glyma19g29210.1
Cluster:	Glyma19g29210.1
Source:	JGI
Genomic Span:	1658 bp
Position:	Gm19:36802487..36804144 (- strand)
Protein Sequence:	>Glyma19g29210.1:peptide MSGIIHKIGETLHVGGHKKEEEHKGEHHAGEYKGEHHGEHSSEYKGEHHGEHKPEHKEGFLDKVKDKIHGEGGAAEGEKK KKEKKKKEHGHEHGHDSSSSSDSD* NCBI BLAST Phytozome BLAST
Protein Length:	105 aa
Full CDS Sequence:	>Glyma19g29210.1:cds atgtcagggatcattcacaagattggggagacccttcatgtgggagggcacaagaaggaggaggagcataagggtgagca ccatgctggagaatacaaaggggagcaccatggtgaacacagtagtgagtacaaaggagagcatcatggtgaacacaaac cagagcacaaagaggggttcctagacaaagtcaaggacaagatccatggtgagggcggcgccgccgagggcgagaagaag aagaaggagaagaagaagaaggagcatggccatgagcatggccatgacagcagcagcagcagtgacagtgattaa
Full CDS Length:	315 bp
Genomic Sequence:	>position=Gm19:36802487..36804144 (- strand) ATACCAAAAACAAAACACAAGCCAAGTGAAAGTGAGAAAAAATAAAAATCAAAGAGCATAATGTCAGGGATCATTCACAA GATTGGGGAGACCCTTCATGTGGGAGGGCACAAGAAGGAGGAGGAGCATAAGGGTGAGCACCATGCTGGAGAATACAAAG GGGAGCACCATGGTGAACACAGTAGTGAGTACAAAGGAGAGCATCATGGTGAACACAAAGCTGGAGAGTACCATGGTGAG CACAAGCCAGAGCACAAAGAGGGGTTCCTAGACAAAGTCAAGGACAAGATCCATGGTGAGGGCGGCGCCGCCGAGGGCGA GAAGAAGAAGAAGGAGAAGAAGAAGAAGGAGCATGGCCATGAGCATGGCCATGACAGCAGCAGCAGCAGTGACAGTGATT AAACTGCTGCATCACCTATATATTTAAAAGGTATACTTTTGTTTTGCATTTTTATGTTTGGCTCTGTCAAAATTAAATCT TTCCACATTGGTCTATTACAACAACATGGATGATGAATTAGGAAATTTAATTATGATCTTATTTTCTTCTATCACTTTTG AGCTAACATTAGTTAGGGAAATTATACATTCGGGCACGATGACACCCTGTGTGTTAGATCTCATTTTTAAACTTCATATA ATTCTTTTTCGGGCGATCTAAGATGTTATAACCATGGAATTATTGGTCAAATTTTAGTTTCTATATAGGTATATTGGAAA AAACTATGATCGTCAATTTCTGCTCCCGCATGTATCTCCTATTTTTACATGTGTAATGATTAGTGTACACTGCATTCTAG AGACATAATAAGTGAAAGGAAATGAAAGGAAAAAGATTAAAAACATGAAAATAAGTAAATATAAGAAAGAGATATTTAGG ATTTTTTTACAATAATGAACAATGACAAACGTTTGAAAACATTTGTTCTGATTTTTTTTCCTTCGATCTAATGCAGGTTT ATGGGACCATAGTAGCGAGGGCTGTGCCTTAATTTCTCTCTCTCTCTCTCTCTCTCTCTCTCTACTCTCTTTTACC TTTTTCCCTATCTTGTGATCTATTTATGAGTGAGAAAAACAAAAGTGAGAAATTGTGCGTGTATCTGGTTGCTAATTCGA GTGCTATTAGTCTGTGGTTTGGTTGGTTTGTAACCCCGCTACTTGGGTTATGTTTGATTGTATATTATCCTAGAGAAATA AAAATTTTAAATTATCAGTACGTATTGGTATTACATGCTAGTGAGCTGTGGTTACATTTTTCTTTTTTTTTAAAGGAGCT ATTGTGGTTACACTGTATTCCAACATAATGCAATTGATGTGATTATAATTGCCACAGCAATAATTCTAAAAACTTGATAT TGTGGCTACAATTACATATTATTTTTTAGAATCATGATCATTCGCACCAATAAATGTTGATCAATTTTATAATTGTTATA AATGTTTAAAAGAAAAAAAAGGATATGGACAATATAAGATAATTTAATGTTCTCATTTAATTATTATCATATGTTAAGTT TTTTACATTTTATAATAATTATTTCAAAAGTCATACTAGTGATTTACTTTTAGATAATTATATAAATTTTTTTACAGTAT TATTAAATGATTTAAAAATGAAAATATTATGAAGCATCAAATTGTATGTTTTTTTTTT
Parts:	**Type:** CDS **Length:** 159 bp **Position:** Gm19:36803926..36804084 (- strand) **Type:** CDS **Length:** 156 bp **Position:** Gm19:36803743..36803898 (- strand) **Type:** 5'-UTR **Length:** 60 bp **Position:** Gm19:36804085..36804144 (- strand) **Type:** 5'-UTR **Length:** 28 bp **Position:** Gm19:36803715..36803742 (- strand) **Type:** 5'-UTR **Length:** 702 bp **Position:** Gm19:36802487..36803188 (- strand)

LAU2

phytozome
Glycine max

Name:	**Glyma20g35270.1**
Cluster:	Glyma20g35270.1
Source:	JGI
Genomic Span:	2963 bp
Position:	Gm20:43516095..43519057 (- strand)
Protein Sequence:	>Glyma20g35270.1:peptide MTAKQTKQHYNRKRTQSEREGSYSHKEKNFQRGESKRVLLFLFYIYLGAKLCEFIDKPEMATMLTKEHGLNLKETELCLG LPGGGGGGGGGGGGGGGEVETPRATGKRGFSETVDLKLNLHSKEDLNENLKNVSKEKTLLKDPAKPPAKAQVVGWPPVRS YRKNMMAVQKVSTEDVAEKTTSSTANPGAFVKVSMDGAPYLRKVDLTMYKSYKELSDALAKMFSSFTMGNYGAQGMIDFM NESKLMDLLNSSEYVPSYEDKDGDWMLVGDVPWEMFVESCKRLRIMKGSEAIGLAPRAMEKCKSRS* NCBI BLAST Phytozome BLAST
Protein Length:	307 aa
Full CDS Sequence:	>Glyma20g35270.1:cds atgactgccaaacaaaccaagcagcattataatagaaagaggacacaaagtgaaagagaagggagctatagccataagga gaaaaactttcaaagaggtgaaagcaaaagagttcttcttttccttttttatatatacttgggtgctaagctgtgtgagt tcattgacaagcctgaaatggcaactatgctgacaaaggagcatggtctgaacctcaaggagaccgagctttgcctcggt ttgcctggtgggggaggcggcggcggcggcggcggcggcggcggtggtgaggtggaaactccaagggccactgggaa gagagggttctctgagactgttgatctgaaacttaatcttcattccaaggaagatctgaatgagaatctgaagaatgtct caaaggagaagaccctccttaaggatcctgccaagccaccggctaaggctcaagtggttggttggccaccagtgaggtca tacaggaagaacatgatggcagtacaaaaggttagcactgaggatgtggcagagaagacaacaagcagcactgctaatcc tggggcatttgtcaaggtttccatggatggagcaccttacctgcgcaaggtggacctcacaatgtacaaaagctacaaag agttatctgatgccttggccaaaatgttcagctccttcaccatgggtaactatggggcccaaggaatgatagacttcatg aatgagagcaagttgatggatcttcttaacagctctgagtatgtgccaagctatgaagataaggatggtgactggatgct cgtgggtgatgtcccatgggagatgtttgttgagtcatgcaagcgtctgcgaataatgaagggatcagaagcgattgggc ttgcgccaagagcaatggaaaaatgcaaaagcagaagctga
Full CDS Length:	921 bp
Genomic Sequence:	>position=Gm20:43516095..43519057 (- strand) ATGACTGGTTAGAATATTGGGGAGACATATGGAAGGTCAATCCATTTTGGTAGGTACCCAAAATCCCGTTTGCACATGAA AATGATAAATCATATATATTAATTTCGATGTATAATTTCTTTGTACTACAAAATACTACATTAAATAATTGATACGATAA TAAAAAATGAAAAAATAATCACATAAATAAAAATAATATAAGTTTAGGGAAAATTATGAGAGGCGGTTAAGCCTACATAC CCTCCTTCTGCAAAGGATCAACGTTCCACATACCTCCATGTGTCAGCATGCAATTGGTGGCAAACCTGATATGACATACG CATGCAGCCACAAGCACAAGTTCCTCAAAGCTATATTATTGAGGGCTATCACACACACTCTCTCTCTCTTGTAATCTATA TAGCCAAACAAACCAAGCAGCATTATAATAGAAAGAGGACACAAAGTGAAAGAGAAGGGAGCTATAGCCATAAGGAGAAA AACTTTCAAAGAGGTGAAAGCAAAAGAGTTCTTCTTTTCCTTTTTTATATATACTTGGGTGCTAAGCTGTGTGAGTTCAT TGACAAGCCTGAAATGGCAACTATGCTGACAAAGGAGCATGGTCTGAACCTCAAGGAGACCGAGCTTTGCCTCGGTTTGC CTGGTGGGGGAGGCGGCGGCGGCGGCGGCGGCGGCGGCGGTGGTGAGGTGGAAACTCCAAGGGCCACTGGGAAGAGA GGGTTCTCTGAGACTGTTGATCTGAAACTTAATCTTCATTCCAAGGAAGATCTGAATGAGAATCTGAAGAATGTCTCAAA GGAGAAGACCCTCCTTAAGGATCCTGCCAAGCCACCGGCTAAGTAAGTCATTTTCTCACTTTTCTTTTATAATTAAAGGC CTTTGCATGCTTCTGTTCATCTTTTCTTTCTTGGACCTAGCTTCAATAGTTTGTTTGATTCCATATCTCATACAAGATTT TTCACGAGAAGTTAATAGTTATAACTTTCATATTCCAAACGTATTTTTTTTAATTAATTTCTCAACGAGTTTCTAAGCAC TACTTAGTGCTTATGATCTATTTTTCTTATAGTTTAGCTATTCAATGTTAGAATATACGGTTTGGATAAGGTGTAAATTA ATTTGCAACATTTTGTACCGAAAAAAGGGAGATATATTAGAAAGTGAAGTGAAGATTGATGGGGACAATATTTCAGGGCT CAAGTGGTTGGTTGGCCACCAGTGAGGTCATACAGGAAGAACATGATGGCAGTACAAAAGGTTAGCACTGAGGATGTGGC AGAGAAGACAACAAGCAGCACTGCTAATCCTGGGGCATTTGTCAAGGTTTCCATGGATGGAGCACCTTACCTGCGCAAGG TGGACCTCACAATGTACAAAAGCTACAAAGAGTTATCTGATGCCTTGGCCAAAATGTTCAGCTCCTTCACCATGGGTTAG TTACCACAAAAAACCTTCCTACGTATCATATTCCCCTCTCTTTAAACTACTTGTTTACCACAACTATGTCACTGAAAGGA TAACCACGCATACACACACAAAAATAAACACCAACTTAAACATCTAACCTTAATTTTTTTCATAATCTATTTTCGATCTG CACCTATTATACACCCATTTCTTTTTCCCTCTTTTTGATTACATTTACTTCTCTCTCTTTTTTTTATTTCTTTCTTCTCT AACCATCTTTAACCACTTGCTCAAGAGTGTCGAATCATCACTATTATATATATAAAGAAGAGATCGATAAAAACAAGTGA CATGAAATGAATATTGTTTATTATATCATATACTAGTAAAGTCTTAGATCCTTGATGCATGGGAATTGTTAATGATGAGT GTACGGGTTGGTTAATCACCTTGGGATTGTTACATTCAGCTATATGATTGTGGGTTATTAAACTAGTATTTTCTAGATAT TTATTTCATATATTGAAACTGGTGAAGGTAACTATGGGGCCCAAGGAATGATAGACTTCATGAATGAGAGCAAGTTGATG GATCTTCTTAACAGCTCTGAGTATGTGCCAAGCTATGAAGATAAGGATGGTGACTGGATGCTCGTGGGTGATGTCCCATG GGAGTAAGTTCTTCATCCAACAACAACCTCTGCAGCCATTATATACAATAATTTCCTACATACTTGATTTATACATCCAT ATATTTTACAAATTCAGTAATCTTTTTTAATCCAAAAAAAGAAAATTATGACTTTTGAGGAGCTCTTTATCAGAAACGTA TATACTTAGGCATTAATTATTCTTATACTGTATTTTAATTCTGTAAACCCACGAAATAAACCTTAATTAACTGTTTAATT TTCTGAAGGATGTTTGTTGAGTCATGCAAGCGTCTGCGAATAATGAAGGGATCAGAAGCGATTGGGCTTGGTATGCATCA TTATATCTTAAGCTTAATTAATTACATCTAGCTACTTTTTCTTTTCACTCATGGAAATCATTGTTTTGAAAACTAATCAA TAATACATACATGCCATTAATTTGTTGGCTATTCTAATGTGTTGGCGCGCATCTTTTTATGTACTTTCAGCGCCAAGAGC AATGGAAAAATGCAAAAGCAGAAGCTGAAGCGGGCCTAATACAATGTTCAATGAAATCCCGTACCAAAGTGGACCTATAT ATATATATATATATATGAAGCACTTCAGCGAGACTATGGAGCGGAGTGTTGTTAGCTTGAATGTGTCTGGTGTTTTTTGT TTTGTTCGTATATAGATTCAACTTTTAATTTTAAGGAGCCGATCGATGGGGCATTTGTAATGGACAAGAAGCAATAAACT TAACCTATATAGTCTGGCTTGTTTATTATGTGATCCTTCCATGCTGTCATGATTGTTGTGTGCTAAAATACCTTGAATTT TAATTCATACTTTATTCCCTCAAGAGATTCTTCATATATCATTTATTTTGATTTTATGTTTTAATTAATTGTGTATGCCC TGC

Parts:		
	Type:	CDS
	Length:	7 bp
	Position:	Gm20:43519051..43519057 (- strand)
	Type:	CDS
	Length:	439 bp
	Position:	Gm20:43518216..43518654 (- strand)
	Type:	CDS
	Length:	239 bp
	Position:	Gm20:43517623..43517861 (- strand)
	Type:	CDS
	Length:	136 bp
	Position:	Gm20:43516975..43517110 (- strand)
	Type:	CDS
	Length:	62 bp
	Position:	Gm20:43516668..43516729 (- strand)
	Type:	CDS
	Length:	38 bp
	Position:	Gm20:43516470..43516507 (- strand)
	Type:	3'-UTR
	Length:	375 bp
	Position:	Gm20:43516095..43516469 (- strand)

LAU5

phytozome
Glycine max

JGI Joint Genome Institut
C I G Center for Integrative G

Name:	**Glyma10g23450.1**
Cluster:	Glyma10g23450.1
Source:	JGI
Genomic Span:	1779 bp
Position:	Gm10:29588814..29590592 (- strand)
Protein Sequence:	>Glyma10g23450.1:peptide MVVAIVVVVAVVCDIAVARPEVMEGIVMFFKEGLRLLRFDEDLGGLEKDVI* NCBI BLAST Phytozome BLAST
Protein Length:	52 aa
Full CDS Sequence:	>Glyma10g23450.1:cds atggttgtagctattgttgttgttgttgctgttgtttgtgacatagctgttgcaagaccagaggttatggaagggattgt gatgtttttcaaggagggtttgcggttgcttagatttgatgaggaccttggaggcttggagaaagatgtgatttga
Full CDS Length:	156 bp
Genomic Sequence:	>position=Gm10:29588814..29590592 (- strand) GTATTGGTTAAAGGAATAGTGAAGATTAAATCCAATAAATAATCTCACCAGCTATTTTCAACTTTCAATTGCGGCAATTC CAATGTACAACATATTTTCTCCAATAAACGAATAAACAAATTTCAATATTGCTTCAAGTACAGTACATTTTTTTACCTAA AGGAAGAAGATAAGCATGGGTTTTTTGTTTTTTTTTTCAACAAACAAAAAGATCTTGCTTCATAGTACTATCCTATATTT ACATTATACAAATCCATATATATATTCACTTGCAAATTAATCTATTAGTAGATTAATTAGTACTACCATTTTGATTTAGG CAAATTAGCAACCTAAATCAGAGAGGACAATCACCAGCATCATCTCTCTCTCTCCCCTTAGTTTGGTTACCTCAAAACAT CATTGAAAGAGCAAAAGTTCCACAAAATCTCAAACAACCATTAGTGGAGATAGTGGTGACAAAGGAGGCACACTAAAGAT TGTTCCCATAACATCAGTGTCCCAAAAGCCTTTCCAACATGATGGAGTCAAAGGGCACTCTTCTTTCTTCACATGCTTAT TACTATTACTATTGTTGTTAAACTCACTTTTCTCATAGTTGTTCTTCCTTTCACTCTCTTCCTCTCTTTCTCTCTTCACA CCAACCTTGATGACGATGCTGCTGCTTGAAACTGATGATGATGATGATGATGACTCTTCTTTAGACTGATCTGATCCAAT CTCCAAAGGGAAGTTCAATATGGCTTTGCTCCCTCTCATCTTGAAAGCAGCTTTGTCATAAGCCTTAGCAGCTTCTATGG CTGTGTCAAAGGTTCCAAGCCACACCCTTGAACCTTTTCTATTAGGGTCACGGATTTCTGCAGCATACTTCCCCCATGGC CTTCTTCTAACCCCTCTATAGTGCTTGTCGTCTTTGGTATTGTTGGTTTCATCTGAAGAAAAAGAAGAAGAAGAAACTTG GTTGTTGTTATACATGGTTGTAGCTATTGTTGTTGTTGTTGCTGTTGTTTGTGACATAGCTGTTGCAAGACCAGAGGTTA TGGAAGGGATTGTGATGTTTTTCAAGGAGGGTTTGCGGTTGCTTAGATTTGATGAGGACCTTGGAGGCTTGGAGAAAGAT GTGATTTGAGGCATTTGAAGGTTATTATTGGTATGGTTCATTGGCATGTTGTTGTCATTCTTGCTCATATCTTGAAATTG AAATTGAAATGGAACTTCTTGGGGTACAAGGTAATCAGTGATAAAAATATCAGATTGTGAAATTGAACTTTGCATTTTTG GTTGTGGAATTGAGACTCAATTGGGACTTCTAAGATGTTTGAGGTGCAAACAGTGAAATTGGGTTCAAAGCTTTGAATGC TTTTCTAATTCAAGGTGAGAGATAATTAAGGGACTGAACTTCTTAGATGGTTGCTAAAGTTAGGAAAGTTGAGAATTTGG AATTGGTTGAGTGGAAAATGGGGAGTGATGAGGTGCATATAAATATGAAGATAAGTATCAAAGAATTAAAAAGGTGGGTT TCTAGAGGGGTTCCAAGATGGTAAGAGGAAGCATCGTTTTGTTACTATCAATAATTGGTGTTAGAAGAAATGTCACGTCT TTTATTTTTTGTCGTTTTAGAATTTTTCTAATGTCAACTTGACTTGTTTTGTTTGTAAGGGACCCCATTACTTGCACGAA ATGACCCTTCGAGGGTACTTCCTTGCACTGGCTTTAAACTTTAAGGTTTATTTAAATTCAAGTTGAGGGTTTTGGTGTAT TTGTGAAAGTATTTCGAGT
Parts:	**Type:** CDS **Length:** 156 bp **Position:** Gm10:29589464..29589619 (- strand) **Type:** 5'-UTR **Length:** 650 bp **Position:** Gm10:29588814..29589463 (- strand) **Type:** 3'-UTR **Length:** 973 bp **Position:** Gm10:29589620..29590592 (- strand)

LAU6

phytozome
Glycine max

JGI Joint Genome Institut
C I G Center for Integrative Ge

Name:	**Glyma04g01130.1**
Cluster:	Glyma04g01130.1
Source:	JGI
Genomic Span:	1371 bp
Position:	Gm04:710039..711409 (- strand)
Protein Sequence:	>Glyma04g01130.1:peptide MADETQNKYESSEVEVQDRGVFDFLGKKKDEEDKPHPQEEVIATEFQKVTVSDQGENKHSLLEKLHRSDSSSSSSSEEEG EDGEKRKKKKKEKKGLKEKIEEKIEGDHHHHKDEDTSVPVEKVEVVETAHAEEKKGFLDKIKEKLPGHKKTEEATATTPP PPPPVASLEHGEGAHHEGEAKEKKGILEKIKEKLPGYHSKTEEEKEKEKESGAH* NCBI BLAST Phytozome BLAST
Protein Length:	215 aa
Full CDS Sequence:	>Glyma04g01130.1:cds atggcagacgagacccagaacaagtatgagagctctgaggttgaggtccaggatcgtggtgtttttgactttctcggtaa gaaaaaggatgaagaagacaagcctcatcctcaggaggaggtcatcgccaccgagtttcaaaaagtcactgtctcagacc aaggagagaacaagcacagcctcttagaaaagcttcaccgatctgacagcagctctagctcttcaagtgaggaggaagga gaagatggagagaaaaggaagaagaagaagaaggaaaagaaagggctgaaagagaagatcgaggagaaaatagagggtga tcatcatcatcacaaggatgaggacacaagtgtccctgttgagaaagttgaggttgttgaaacagcacatgctgaggaaa agaaggggttcctcgacaagattaaggagaagctaccagggcacaagaagacagaggaggccacagctactactcctcct ccaccaccacctgttgcatcattggagcatggtgaaggtgctcatcatgaaggagaggccaaggagaagaaaggtatatt agaaaagataaaagagaagcttcctggttatcactccaagacagaggaggaaaaggaaaaggaaaaggagagtggtgctc attga
Full CDS Length:	645 bp
Genomic Sequence:	>position=Gm04:710039..711409 (- strand) GGGTTTTGGTATTTCTTAATCGCGGGAAGTGCGCGTTATGTGAAGTATAAATGGTGTCCCAGTCTTGCACCTTAAAACCA TCCAATCCAATTGAATTCTCCAGAGAGAAGAAGAAACTTAATCGATCATCACTTCACTTCACTAATACATCATGGCAGAC GAGACCCAGAACAAGTATGAGAGCTCTGAGGTTGAGGTCCAGGATCGTGGTGTTTTTGACTTTCTCGGTAAGAAAAAGGA TGAAGAAGACAAGCCTCATCCTCAGGAGGAGGTCATCGCCACCGAGTTTCAAAAAGTCACTGTCTCAGACCAAGGAGAGA ACAAGCACAGCCTCTTAGAAAAGCTTCACCGATCTGACAGCAGCTCTAGCTCTGTAAGCTTCTTTTCCTTTATTTTCTTT GCATGCATAAACATTATATATATATAACTTGTTCATCCTGAAAGGAATAACTTTATATTAACTTGTCTTTTTTTTTTAAT ATTTATTTTATAATTTAAACAAAACTTAATTTTATACGAGATTGTTCTAGTTGAATCAGCGATCTTGAGTTTGGTTCTAG GATTGTCAACCCTTGATGAATGATTTCGGATCTAAAAAATAAAAAATAAAAAATGTTACTTCAAGGTGGGTTTTGATTGG TATATATATGTTTGCAGTCAAGTGAGGAGGAAGGAGAAGATGGAGAGAAAAGGAAGAAGAAGAAGAAGGAAAAGAAAGGG CTGAAAGAGAAGATCGAGGAGAAAATAGAGGGTGATCATCATCATCACAAGGATGAGGACACAAGTGTCCCTGTTGAGAA AGTTGAGGTTGTTGAAACAGCACATGCTGAGGAAAAGAAGGGGTTCCTCGACAAGATTAAGGAGAAGCTACCAGGGCACA AGAAGACAGAGGAGGCCACAGCTACTACTCCTCCTCCACCACCACCTGTTGCATCATTGGAGCATGGTGAAGGTGCTCAT CATGAAGGAGAGGCCAAGGAGAAGAAAGGTATATTAGAAAAGATAAAAGAGAAGCTTCCTGGTTATCACTCCAAGACAGA GGAGGAAAAGGAAAAGGAAAAGGAGAGTGGTGCTCATTGATATTTAAGATTGAGACAAGGCTTTTTTGTTGGTTGGAGAT GCATGATTGTGGTGTGCTTTGTTTTCATTTCATCATATGACAGCCTTGTGGGTTTCTTTGGTGTTTCTTGATGCATACTT AATTTACTTTATTATTGTGTATTTATCTCTTCCATTTACAAAAAAAAAAAGTTTGTGTAGGTTTCCTTCTTTGAAGTAAAA AGAAATTAAATATGTACTGAAAATGGAATAATAGTGTGTTTTATTTGTGTTTTATAATTTCTTTTTTAAACAGTTGTCGT CTTTACATTTC
Parts:	**Type:** CDS **Length:** 222 bp **Position:** Gm04:711037..711258 (- strand) **Type:** CDS **Length:** 423 bp **Position:** Gm04:710330..710752 (- strand) **Type:** 5'-UTR **Length:** 151 bp **Position:** Gm04:711259..711409 (- strand) **Type:** 5'-UTR **Length:** 291 bp **Position:** Gm04:710039..710329 (- strand)

LAU7

phytozome
Glycine max

Name:	**Glyma13g40820.1**
Cluster:	Glyma13g40820.1
Source:	JGI
Genomic Span:	1414 bp
Position:	Gm13:41270585..41271998 (+ strand)
Protein Sequence:	>Glyma13g40820.1:peptide MPISRIAIGNSSELNQSDALKAALAEFISMLIFVFAGEGSGMAYNKLTNNGSATPAGLVAASLSHAFALFVAVSVGANIS GGHVNPAVTFGAFVGGHITLFRSILYWIAQLLGSVVACLLLKFATGGLETSAFALSPGVEAGNALVFEIVMTFGLVYTVY ATAVDPKKGDLGIIAPIAIGFIVGANILAGGAFDGASMNPAVSFGPAVVSWTWSNHWVYWVGPFAGAAIAAVVYEIFFIS PNTHEQLPVTDY* NCBI BLAST Phytozome BLAST
Protein Length:	253 aa
Full CDS Sequence:	>Glyma13g40820.1:cds atgccgatttctagaattgccattggaaattcttcggagttgaaccaatctgatgcacttaaggctgcactagctgagtt catctcaatgctcatctttgtttttgccggggaaggctctggaatggcttataataagctgacaaacaatggttcagcaa caccagcagggttagtggcagcatcactgtcacatgcctttgcacttttgttgcggtctctgttggagccaacatttct ggcggtcatgtaaaccctgctgtcacttcggtgcctttgttggtggccacattaccctctttagaagcatcttgtactg gattgctcagttactcggctctgtcgttgcttgcttgctccttaaatttgccactggtggactggaaacatctgcatttg cactatctcctggggtggaagcagggaacgctctagtgtttgagattgtgatgacttttgggttggtttacacggtgtac gcaactgcagtggatccaaagaagggtgatcttgggataattgctccaattgcaattggtttcatcgttggtgcgaacat cttggcgggggggtgcatttgatggtgcatccatgaaccctgcagtgtcgtttgggcctgctgttgtcagttggacctggt ctaaccactgggtttattgggtcggcccatttgctggtgctgccattgctgctgttgtctacgagattttcttcattagc ccaaacactcatgaacagctccccgtcacagattattag
Full CDS Length:	759 bp
Genomic Sequence:	>position=Gm13:41270585..41271998 (+ strand) CCTAACAACGACTTAAGGCATTCTCTCTTCTATTCTATTCTAAACTCGAAACAATCTTAGAGAAAGAAGCAGAAGAAAAT GCCGATTTCTAGAATTGCCATTGGAAATTCTTCGGAGTTGAACCAATCTGATGCACTTAAGGCTGCACTAGCTGAGTTCA TCTCAATGCTCATCTTTGTTTTTGCCGGGGAAGGCTCTGGAATGGCTTATAGTAAAGTTCTTCCTATTTATTTATTCAAC TTTATGTTGGTTGGAACATGCATGTAGTAATATGTATCAAGAGTTTTATACACATCCAACTATAAATAAGTTTACTGACT TCTACGATAACACATAACTACGTTAAAAGTCACGGGAACAATGATTTCTAATTAGTAAACATAGAAATTAAACTCGTACT AATAACTTGTGTTATGCTTGATTGGACAGATAAGCTGACAAACAATGGTTCAGCAACACCAGCAGGGTTAGTGGCAGCAT CACTGTCACATGCCTTTGCACTTTTTGTTGCGGTCTCTGTTGGAGCCAACATTTCTGGCGGTCATGTAAACCCTGCTGTC ACTTTCGGTGCCTTTGTTGGTGGCCACATTACCCTCTTTAGAAGCATCTTGTACTGGATTGCTCAGTTACTCGGCTCTGT CGTTGCTTGCTTGCTCCTTAAATTTGCCACTGGTGGACTGGTACACACCGCAACATCATTTTTAATTGCTCTCTAATTTC TAGTTTTGTGTTGAGTCTTAAGTATGAATATATGTTTTGTGTGTTAGGAAACATCTGCATTTGCACTATCTCCTGGGGTG GAAGCAGGGAACGCTCTAGTGTTTGAGATTGTGATGACTTTTGGGTTGGTTTACACGGTGTACGCAACTGCAGTGGATCC AAAGAAGGGTGATCTTGGGATAATTGCTCCAATTGCAATTGGTTTCATCGTTGGTGCGAACATCTTGGCGGGGGGTGCAT TTGATGGTGCATCCATGAACCCTGCAGTGTCGTTTGGGCCTGCTGTTGTCAGTTGGACCTGGTCTAACCACTGGGTTTAT TGGGTCGGCCCATTTGCTGGTGCTGCCATTGCTGCTGTTGTCTACGAGATTTTCTTCATTAGCCCAAACACTCATGAACA GCTCCCCGTCACAGATTATTAGAGCTTAACTCTCTGCCTTCCTTTGTTGTCACTGCTCTGTGTTTTTTTGTTTTGCATTT GCTTGTTTTCAAATGTTGGTTCTCTGAACTTTGTAACATGAACATCTTTTTCTTTTTTTTTCATTATGAAATGTTCATTT CATCTAAACTAGAATTATTGTGTTAGCCAAGCTTCATTGCAAGTTAAATTTAATCATTCGTGAGAAGAGAATTGCTGGAG TCGACAGAGGCATAAACCTGTGTTAGATTACCAAGGAACATCAAGCATCAATTG
Parts:	**Type:** CDS **Length:** 133 bp **Position:** Gm13:41270663..41270795 (+ strand) **Type:** CDS **Length:** 251 bp **Position:** Gm13:41271014..41271264 (+ strand) **Type:** CDS **Length:** 375 bp **Position:** Gm13:41271352..41271726 (+ strand) **Type:** 5'-UTR **Length:** 78 bp **Position:** Gm13:41270585..41270662 (+ strand) **Type:** 3'-UTR **Length:** 272 bp **Position:** Gm13:41271727..41271998 (+ strand)

9.4 ANEXO 4: Actualización de marcadores funcionales DB-ICRISAT

Q43453

phytozome

Glycine max

JGI Joint

C I G Center f

Name:	**Glyma17g03350.1**
Cluster:	Glyma17g03350.1
Source:	JGI
Genomic Span:	1074 bp
Position:	Gm17:2222082..2223155 (- strand)
Protein Sequence:	>Glyma17g03350.1:peptide MGIFTFEDETTSPVAPATLYKALVTDADNVIPKAVEAFRSVENLEGNGGPGTIKKITFVEDGESKFVLHKIESVDEANLG YSYSVVGGVGLPDTVEKITFECKLAAGANGGSAGKLTVKYQTKGDAQPNPDDLKIGKVKSDALFKAVEAYLLANPHYN* NCBI BLAST Phytozome BLAST
Protein Length:	159 aa
Full CDS Sequence:	>Glyma17g03350.1:cds atgggtattttcacatttgaggatgaaaccacctcccctgtggctcctgctacccttacaaagctctagttacagatgc cgacaacgtcatcccaaaggctgtcgaagccttcaggagtgttgaaaaccttgaggggaacggtggccctggaaccatca agaagatcactttcgttgaggatggagaaagcaagtttgtgttgcacaaaatagaatcagttgatgaggcaaacttggga tacagctatagcgtagttggtggagttgggttgccagacacagtggagaagatcacattcgaatgcaaattggctgctgg cgccaacggagggtctgctgggaagctaactgtcaaataccaaaccaaaggagatgctcagcccaacccagacgacctca aaattggcaaagtcaagtctgatgctcttttcaaggccgttgaggcctacctttttggccaatcctcattacaactga
Full CDS Length:	477 bp
Genomic Sequence:	>position=Gm17:2222082..2223155 (- strand) GTCCACACTATGGTATTGGGGCCTACATAGAACTAGAGTGACCCCAGGGCTACGCTAGTTCTTATAAATAGAGGACACTC CTCTGCTAAAGAAACACACAGCAACAAACATCTTCACAAACTAGTAATATTATTCTTCGAATCCATTTTATTATTCATAA TGGGTATTTTCACATTTGAGGATGAAACCACCTCCCCTGTGGCTCCTGCTACCCTTTACAAAGCTCTAGTTACAGATGCC GACAACGTCATCCCAAAGGCTGTCGAAGCCTTCAGGAGTGTTGAAAACCTTGAGGGGAACGGTGGCCCTGGAACCATCAA GAAGATCACTTTCGTTGAGGGTACTCACTCACTCACTATCTATCTATCTATCTATCTATGTTCATTTTATATTCTACCAT ATAGTATAACTAAGTTAGTTAATGAAATAAATTATGATAATGAGGCTGATCATTGGATGGATGGACAGATGGAGAAAGCA AGTTTGTGTTGCACAAAATAGAATCAGTTGATGAGGCAAACTTGGGATACAGCTATAGCGTAGTTGGTGGAGTTGGGTTG CCAGACACAGTGGAGAAGATCACATTCGAATGCAAATTGGCTGCTGGCGCCAACGGAGGGTCTGCTGGGAAGCTAACTGT CAAATACCAAACCAAAGGAGATGCTCAGCCCAACCCAGACGACCTCAAAATTGGCAAAGTCAAGTCTGATGCTCTTTTCA AGGCCGTTGAGGCCTACCTTTTGGCCAATCCTCATTACAACTGATCCAATTCGATCTTCAGTATTCAGTGATCTGCAAAG ACCTTGTTTTATATTATATACAAGAGTTTTCTTGCCTTGTGTGGAGTGTTATTCAATCACTTTGAGTGTGTGGTGTGGCT TCCAATTGATGTAACGAGTGTTTCCTTTCTTATTTCCTTTTTCCACAGAATTGTGAGAGCCAGTTGATGTCTTTGTATGT CACTTCATCAATAAATCAATATGTTATAATAATAAAAGAAAGCATCATTTATACTTCTGCTTGTTTTATGTTTACTATGG TGACTTTATATAGAATTTAACCCAGTTTTTGAGG
Parts:	**Type:** CDS **Length:** 181 bp **Position:** Gm17:2222816..2222996 (- strand) **Type:** CDS **Length:** 296 bp **Position:** Gm17:2222392..2222687 (- strand) **Type:** 5'-UTR **Length:** 159 bp **Position:** Gm17:2222997..2223155 (- strand) **Type:** 5'-UTR **Length:** 310 bp **Position:** Gm17:2222082..2222391 (- strand)

P127_ARATH

phytozome
Glycine max

JGI Joint Genome Institute US Depart
C I G Center for Integrative Genomics UC

Name:	**Glyma06g00550.1**
Cluster:	Glyma06g00550.1
Source:	JGI
Genomic Span:	1515 bp
Position:	Gm06:264336..265850 (- strand)
Protein Sequence:	>Glyma06g00550.1:peptide MSKEVSQEGLQRKDYVDPPPAPLFDLAEIKLWSFYRALIAEFIASLLFLYVTVATIIGHKKQTGPCDGVGLLGIAWSFGG MIFVLVYCTAGISGGHINPAVTFGLFLARKVSLIRAVFYMVAQCLGAICGVGLVKAFMKHSYNSLGGGANSVSAGYNKGS ALGAEIIGTFVLVYTVFSATDPKRSARDSHVPVLAPLPIGFAVFMVHLATIPITGTGINPARSLGAAVIYNNGKVWDEHW IFWVGPLVGALAAAAYHQYILRAGAIKALGSFRSNPTN* NCBI BLAST Phytozome BLAST
Protein Length:	279 aa
Full CDS Sequence:	>Glyma06g00550.1:cds atgtcgaaggaagtgagccaagaaggcctgcagaggaaggactacgtagaccctcctccagcccctcttttcgaccttgc cgagatcaagctctggtccttctacagagccctcatcgccgagttcattgcctcactcctcttcctctacgtcaccgtcg ccaccatcatcggccacaagaaacagaccggaccatgcgacggcgttggccttctcggcatcgcatggtccttcggtggc atgatcttcgtcctcgtctactgcaccgccggcatttctggtggacacatcaaccctgcggtgacttttggcctgttcct ggcccgcaaggtgtctctcattcgcgccgtgttctacatggtagcacagtgtcttggtgccatctgcggtgttgggttgg tgaaggccttcatgaagcactcctacaactccctcggcggcggtgctaactccgtcagcgcaggctacaacaaaggcagt gctctgggtgctgagattatcggcactttcgtccttgtctacaccgttttctcagctaccgaccccaagagaagcgcgcg tgactcccatgtcccgtgttggccccattgcccattgggtttgccgttttcatggttcacttggccaccattcccatca ccggtaccgggatcaacccagccaggagcttgggtgcagctgttatctacaacaatggcaaagtttgggacgagcattgg atattctgggttgggccattggtgggagctttggcggcggcggcataccaccagtacatccttagagcaggggctattaa ggccttgggttcattcaggagcaaccctaccaactag
Full CDS Length:	837 bp
Genomic Sequence:	>position=Gm06:264336..265850 (- strand) TCCCAGTCTCTTTCCAACTCTACTAGTCTAAGTTTGTAACTCTTGTTGATTCCAAAGTTGCACACCACACTCAACTTACC GTTCCTTTTCCTCTCCCATAAAACCAGCAAGCAACTCACTCCCACACTCACCCCAGTCAAGAGAGAGAATGTCGAAGGAA GTGAGCCAAGAAGGCCTGCAGAGGAAGGACTACGTAGACCCTCCTCCAGCCCCTCTTTTCGACCTTGCCGAGATCAAGCT CTGGTCCTTCTACAGAGCCCTCATCGCCGAGTTCATTGCCTCACTCCTCTTCCTCTACGTCACCGTCGCCACCATCATCG GCCACAAGAAACAGACCGGACCATGCGACGGCGTTGGCCTTCTCGGCATCGCATGGTCCTTCGGTGGCATGATCTTCGTC CTCGTCTACTGCACCGCCGGCATTTCTGGTAACTTTACCCACCTCTTCGCTTCTCGTTCATTTTTTTTGTCGTTAACACA AGTACACAACACAACACAGGTGGACACATCAACCCTGCGGTGACTTTTGGCCTGTTCCTGGCCCGCAAGGTGTCTCTCAT TCGCGCCGTGTTCTACATGGTAGCACAGTGTCTTGGTGCCATCTGCGGTGTTGGGTTGGTGAAGGCCTTCATGAAGCACT CCTACAACTCCCTCGGCGGCGGTGCTAACTCCGTCAGCGCAGGCTACAACAAAGGCAGTGCTCTGGGTGCTGAGATTATC GGCACTTTCGTCCTTGTCTACACCGTTTTCTCAGCTACCGACCCCAAGAGAAGCGCGCGTGACTCCCATGTCCCCGTATG TTTATTTCTTTTCATTCCCCGTTTCCCCGCATCTAAATTAGTTAATAAATATTCTTCATTTTTTGCAGGTGTTGGCCCCA TTGCCCATTGGGTTTGCCGTTTTCATGGTTCACTTGGCCACCATTCCCATCACCGGTACCGGGATCAACCCAGCCAGGAG CTTGGGTGCAGCTGTTATCTACAACAATGGCAAAGTTTGGGACGAGCATGTATGTGGATCTATTATTGTCCACGTCATAT GACTTGTTTGATTTATTAAATTAAATGTAATCTGTGTTTGGTTGCAGTGGATATTCTGGGTTGGGCCATTGGTGGGAGCT TTGGCGGCGGCGGCATACCACCAGTACATCCTTAGAGCAGGGGCTATTAAGGCCTTGGGTTCATTCAGGAGCAACCCTAC CAACTAGTTCCCAAGACAATGTCATCATCACCAATTCGCCTCTTCTTTTCTATTTATTTCTTTTGTTTGTATGAGAGATC ATCATGCCCTTGTGTATGGGGATAATTCGTCTTTTTTTCTTCTTCCTTCCTTTTATCTCCCCGCCTGTCTCATTTGTCCG TTCGCTTACACTTCTAACCAAAATTCCCGTGAACACAACCCACAGGGTTTGCTGTCGGAATGTGAACCTTTTTCGCTTGC GTTCAATTATGCCAAGAATTTTCTAGCTAACATGGGTATATAACTCCTATATTTATGCTGCAAAATAGATAAATA
Parts:	**Type:** CDS **Length:** 280 bp **Position:** Gm06:265423..265702 (- strand) **Type:** CDS **Length:** 296 bp **Position:** Gm06:265056..265351 (- strand) **Type:** CDS **Length:** 141 bp **Position:** Gm06:264842..264982 (- strand) **Type:** CDS **Length:** 120 bp **Position:** Gm06:264644..264763 (- strand) **Type:** 5'-UTR **Length:** 148 bp **Position:** Gm06:265703..265850 (- strand) **Type:** 3'-UTR **Length:** 308 bp **Position:** Gm06:264336..264643 (- strand)

Q6XPS8

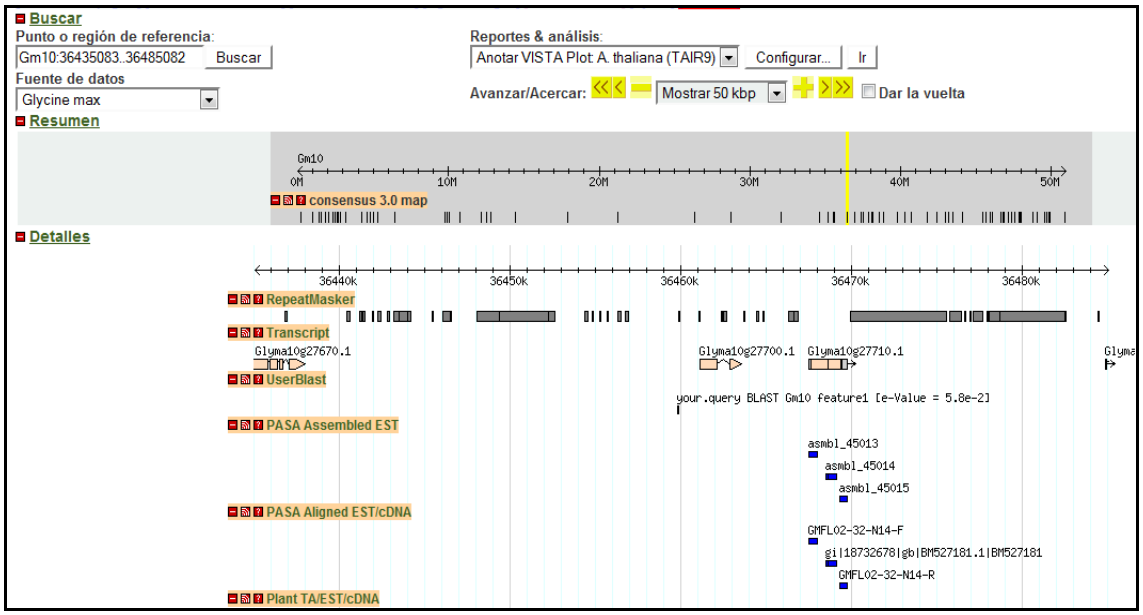
phytozome
Glycine max

JGI Joint Genome Institute US Departr
C I G Center for Integrative Genomics UC

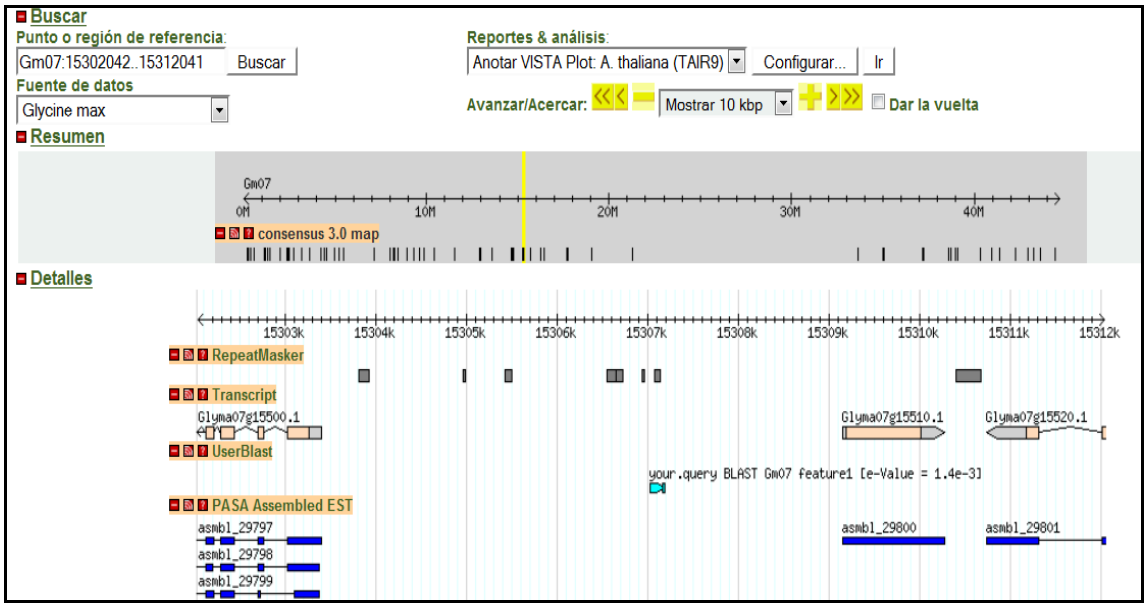
Name:	**Glyma18g00980.1**
Cluster:	Glyma18g00980.1
Source:	JGI
Genomic Span:	720 bp
Position:	Gm18:461960..462679 (+ strand)
Protein Sequence:	>Glyma18g00980.1:peptide MGPMVLSQLATGLSVLAGAVLVKSVMDQKPMAGPFTRCPTCNGTGRVTCLCSRWSDGDVGCSTCSGSGRMACSSCGGSGT GRPLPAKIAIRPPNRPIN* NCBI BLAST Phytozome BLAST
Protein Length:	99 aa
Full CDS Sequence:	>Glyma18g00980.1:cds atgggtccgatggtgctgagccaactcgccaccggtcttagcgttctagccggagcggttctggtgaaatcggttatgga ccagaagcccatggcaggcccattcactcgctgccccacgtgcaacggaaccggtcgagtcacgtgcctctgctcgcgtt ggtccgacggcgacgtcggatgctccacgtgctccgggtcgggtcgcatggcctgcagcagttgcggcggctccggtacc ggtcgacccttgccggcgaaaatcgcgattcgcccgccgaaccgcccaattaattag
Full CDS Length:	297 bp
Genomic Sequence:	>position=Gm18:461960..462679 (+ strand) ACAATTCCACAACCACTTTTTATTTATTTACACAAATTTTTCTCTCCTTACCTATTATTCATTCCACACTTCACATACAT AACATAAAATCTTCCTCCGTTCAATCTTCATCTCTCTCTCTCTATTCATCTCCGTTAAGCAAAAATCATGGGTCCGATGG TGCTGAGCCAACTCGCCACCGGTCTTAGCGTTCTAGCCGGAGCGGTTCTGGTGAAATCGGTTATGGACCAGAAGCCCATG GCAGGCCCATTCACTCGCTGCCCCACGTGCAACGGAACCGGTCGAGTCACGTGCCTCTGCTCGCGTTGGTCCGACGGCGA CGTCGGATGCTCCACGTGCTCCGGGTCGGGTCGCATGGCCTGCAGCAGTTGCGGCGGCTCCGGTACCGGTCGACCCTTGC CGGCGAAAATCGCGATTCGCCCGCCGAACCGCCCAATTAATTAGTTCACCAATTCGTTGGGCTCGCTTGCTCGCTTGCTA CTACTCCTTTTATTTTAGCGTCAGTATAACATTTTAATTTTATATAGCTTTTTTATTTTATTTTATATCGATCAATTATT ATGTAATATAAAGTGCACGTTGACACTTGTATTGAAATGAACAGAGAGGCAAATTAATTATTAGGTTCCAATCATATACT TAGTATTACTGTAGTAATTCAGATTCTTGATTTCTTATACAAGCATATATATAATTTTGCTCCTCTTTTCGTTTGAAGGC
Parts:	**Type:** CDS **Length:** 297 bp **Position:** Gm18:462107..462403 (+ strand) **Type:** 5'-UTR **Length:** 147 bp **Position:** Gm18:461960..462106 (+ strand) **Type:** 3'-UTR **Length:** 276 bp **Position:** Gm18:462404..462679 (+ strand)

9.5 ANEXO 5: Actualización de marcadores anónimos (satt)

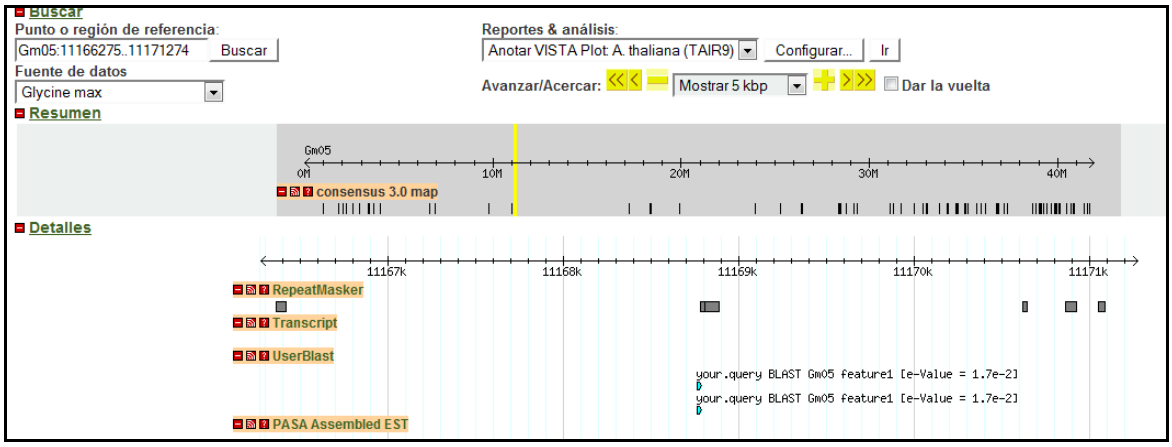
satt173



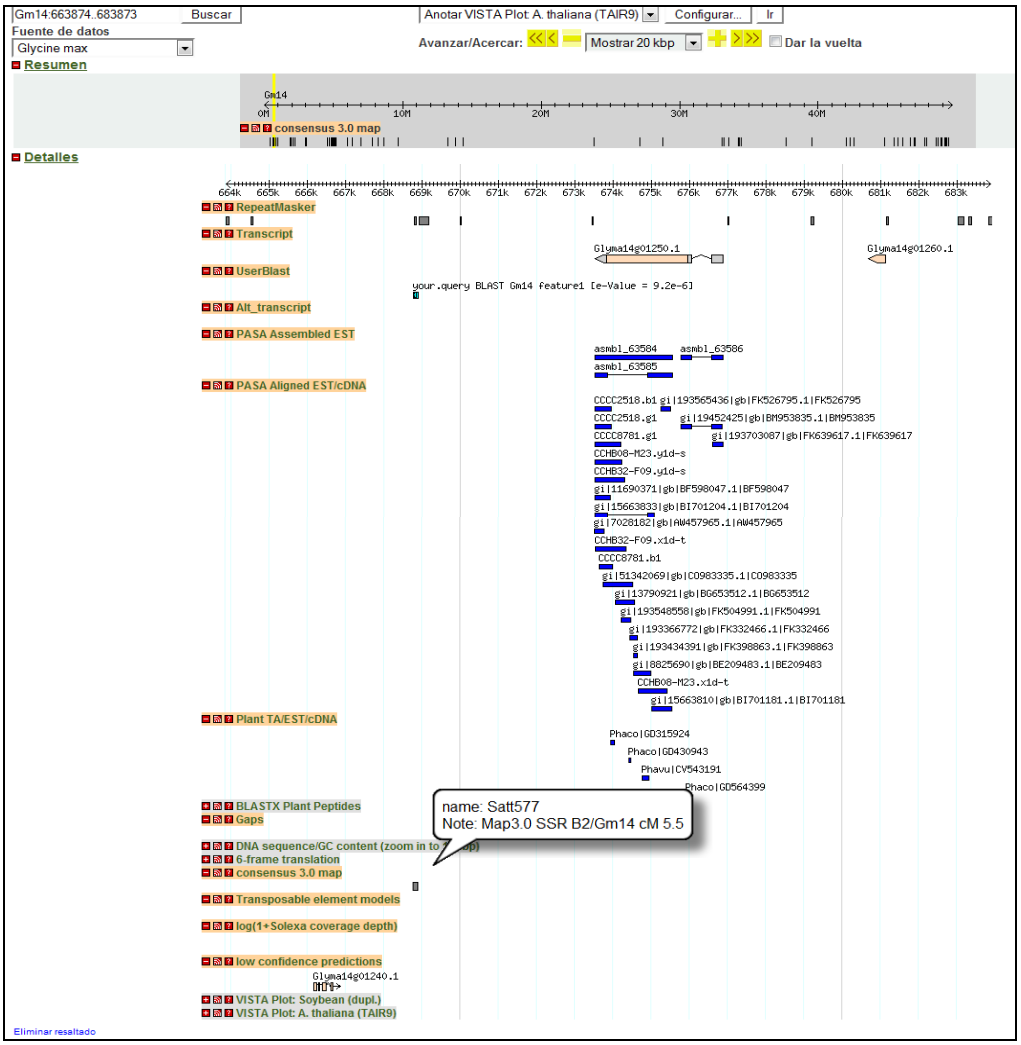
satt175

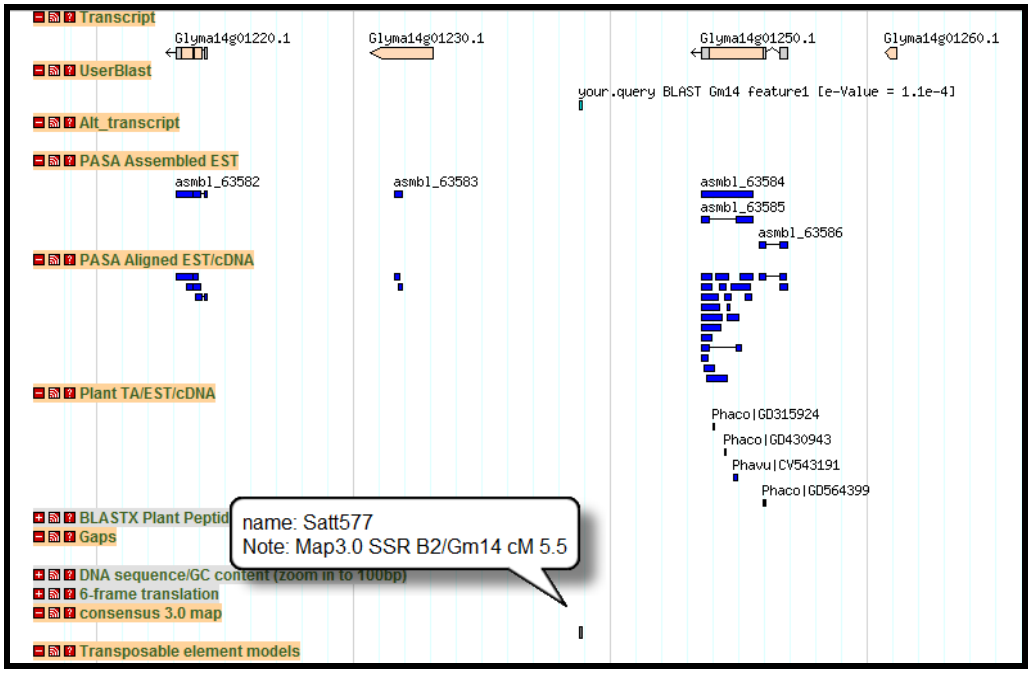


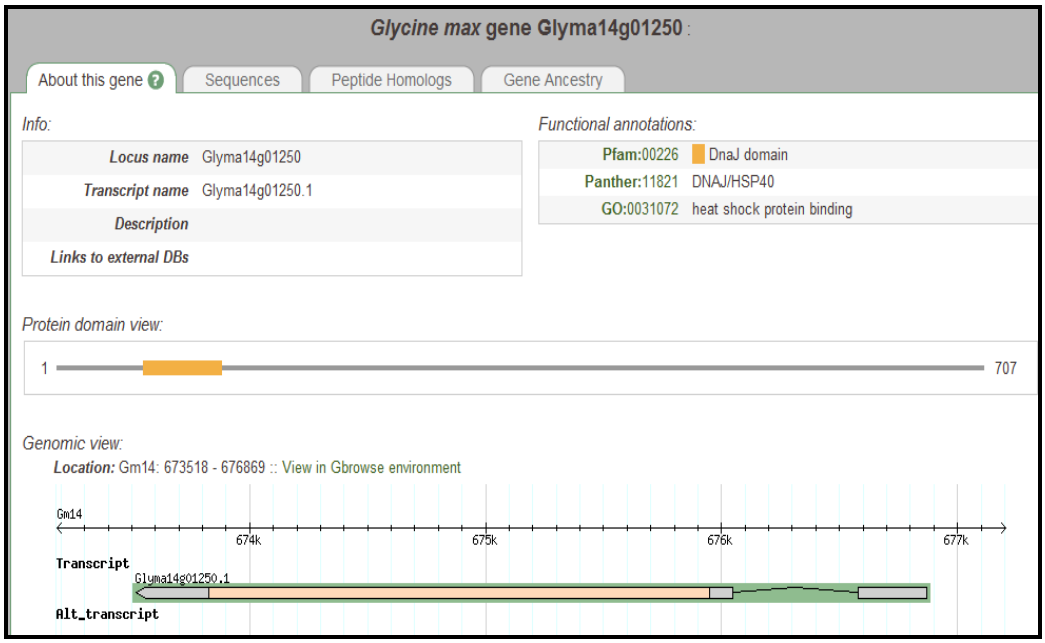
satt42

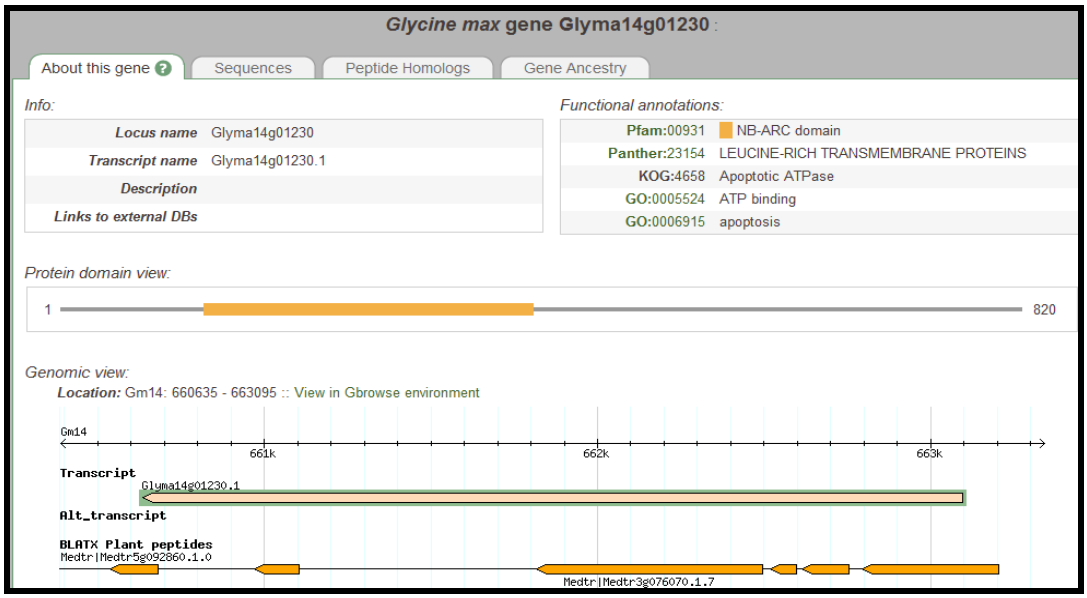


satt577

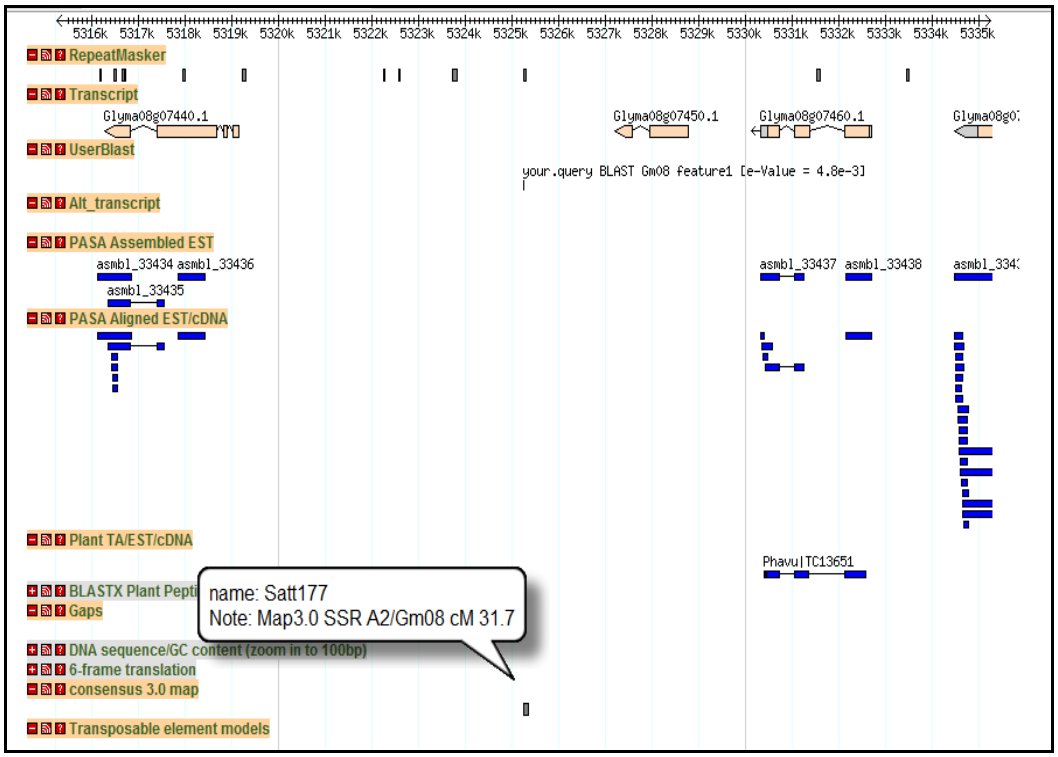


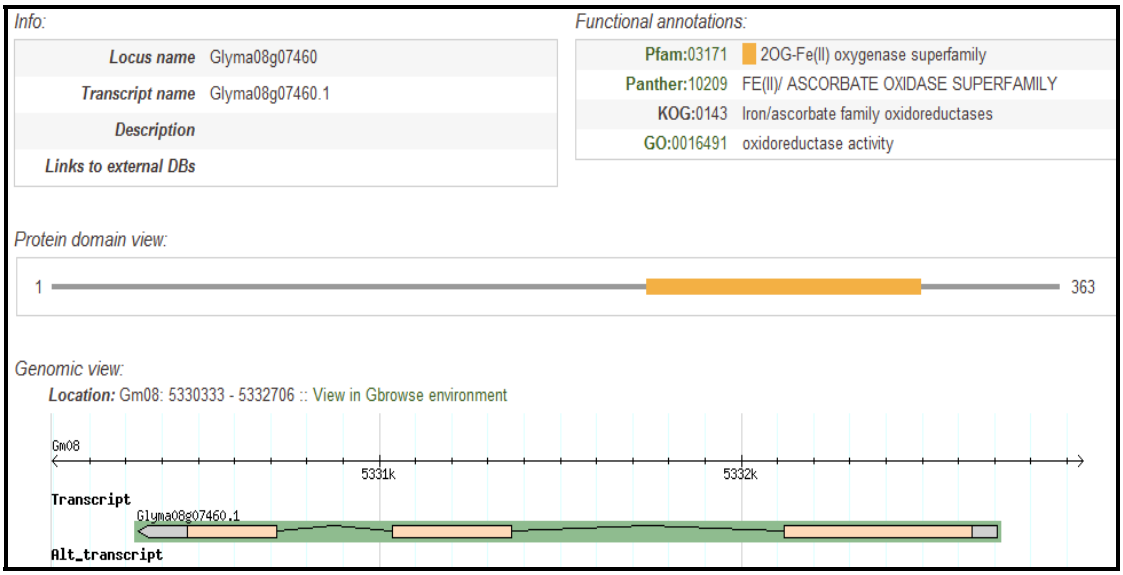

Transcript
Glyma14g01220.1
Glyma14g01230.1
Glyma14g01250.1
Glyma14g01260.1
UserBlast
your.query BLAST Gm14 feature1 [e-Value = 1.1e-4]
Alt_transcript
PASA Assembled EST
asmbl_63582
asmbl_63583
asmbl_63584
asmbl_63585
asmbl_63586
PASA Aligned EST/cDNA
Plant TA/EST/cDNA
Phaco|GD315924
Phaco|GD430943
Phavu|CV543191
Phaco|GD564399
BLASTX Plant Peptid
Gaps
name: Satt577
Note: Map3.0 SSR B2/Gm14 cM 5.5
DNA sequence/GC content (zoom in to 100bp)
6-frame translation
consensus 3.0 map
Transposable element models

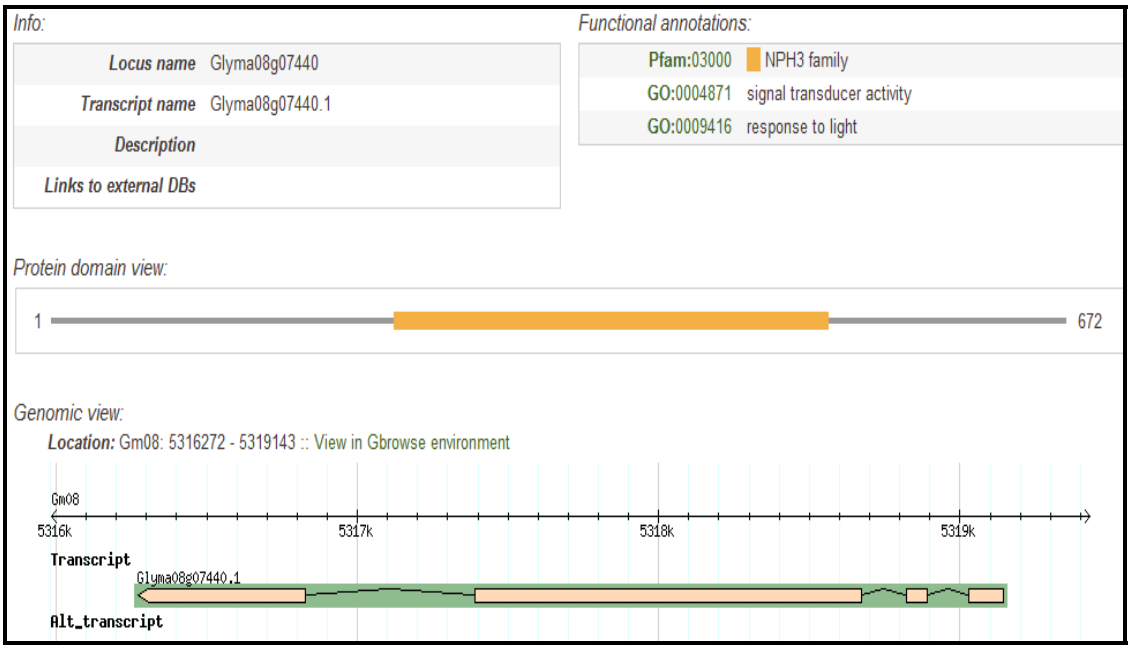

Glycine max gene Glyma14g01250 :
About this gene
Sequences
Peptide Homologs
Gene Ancestry
Info:
Locus name Glyma14g01250
Transcript name Glyma14g01250.1
Description
Links to external DBs
Functional annotations:
Pfam:00226 DnaJ domain
Panther:11821 DNAJ/HSP40
GO:0031072 heat shock protein binding
Protein domain view:
1
707
Genomic view:
Location: Gm14: 673518 - 676869 :: View in Gbrowse environment
Gm14
674k
675k
676k
677k
Transcript
Glyma14g01250.1
Alt_transcript


Glycine max gene Glyma14g01230 :
About this gene
Sequences
Peptide Homologs
Gene Ancestry
Info:
Locus name Glyma14g01230
Transcript name Glyma14g01230.1
Description
Links to external DBs
Functional annotations:
Pfam:00931 NB-ARC domain
Panther:23154 LEUCINE-RICH TRANSMEMBRANE PROTEINS
KOG:4658 Apoptotic ATPase
GO:0005524 ATP binding
GO:0006915 apoptosis
Protein domain view:
1
820
Genomic view:
Location: Gm14: 660635 - 663095 :: View in Gbrowse environment
Gm14
661k
662k
663k
Transcript
Glyma14g01230.1
Alt_transcript
BLATX Plant peptides
Medtr|Medtr5g092860.1.0
Medtr|Medtr3g076070.1.7

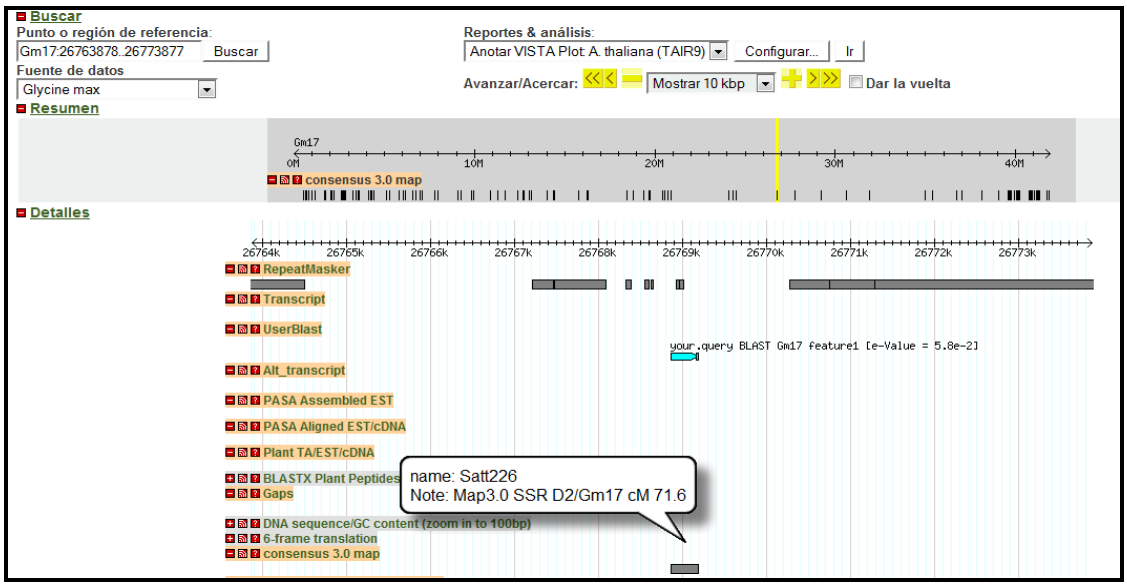
satt177


5316k 5317k 5318k 5319k 5320k 5321k 5322k 5323k 5324k 5325k 5326k 5327k 5328k 5329k 5330k 5331k 5332k 5333k 5334k 5335k
RepeatMasker
Transcript
Glyma08g07440.1
Glyma08g07450.1
Glyma08g07460.1
UserBlast
your.query BLAST Gm08 feature1 [e-Value = 4.8e-3]
Alt_transcript
PASA Assembled EST
asmbl_33434 asmbl_33436
asmbl_33435
asmbl_33437 asmbl_33438
PASA Aligned EST/cDNA
Plant TA/EST/cDNA
Phavu|TC13651
BLASTX Plant Pepti
Gaps
name: Satt177
Note: Map3.0 SSR A2/Gm08 cM 31.7
DNA sequence/GC content (zoom in to 100bp)
6-frame translation
consensus 3.0 map
Transposable element models

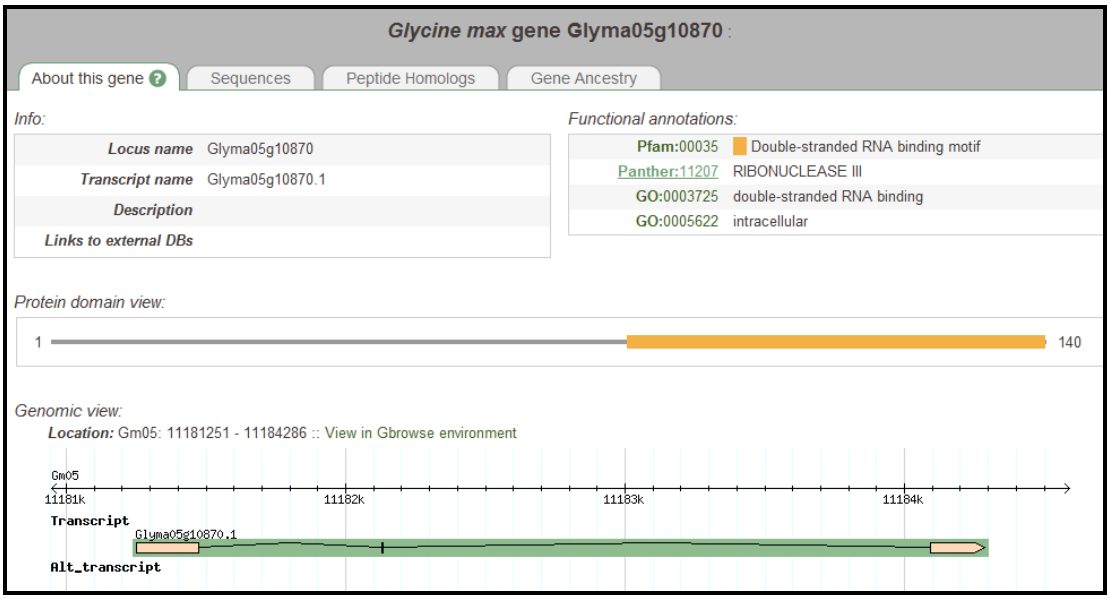

Info:
Locus name Glyma08g07460
Transcript name Glyma08g07460.1
Description
Links to external DBs
Functional annotations:
Pfam:03171 2OG-Fe(II) oxygenase superfamily
Panther:10209 FE(II)/ ASCORBATE OXIDASE SUPERFAMILY
KOG:0143 Iron/ascorbate family oxidoreductases
GO:0016491 oxidoreductase activity
Protein domain view:
1
363
Genomic view:
Location: Gm08: 5330333 - 5332706 :: View in Gbrowse environment
Gm08
5331k
5332k
Transcript
Glyma08g07460.1
Alt_transcript


Info:
Locus name Glyma08g07440
Transcript name Glyma08g07440.1
Description
Links to external DBs
Functional annotations:
Pfam:03000 NPH3 family
GO:0004871 signal transducer activity
GO:0009416 response to light
Protein domain view:
1
672
Genomic view:
Location: Gm08: 5316272 - 5319143 :: View in Gbrowse environment
Gm08
5316k
5317k
5318k
5319k
Transcript
Glyma08g07440.1
Alt_transcript

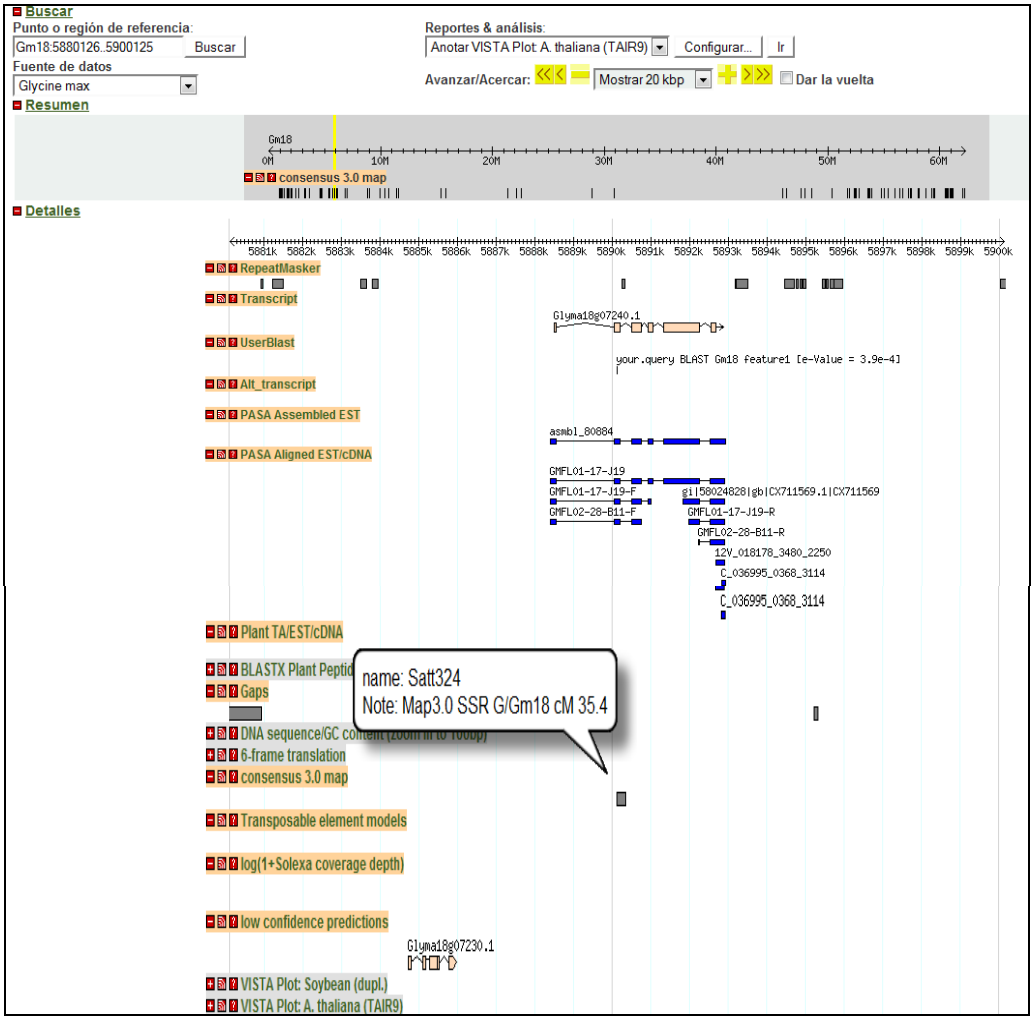
satt226

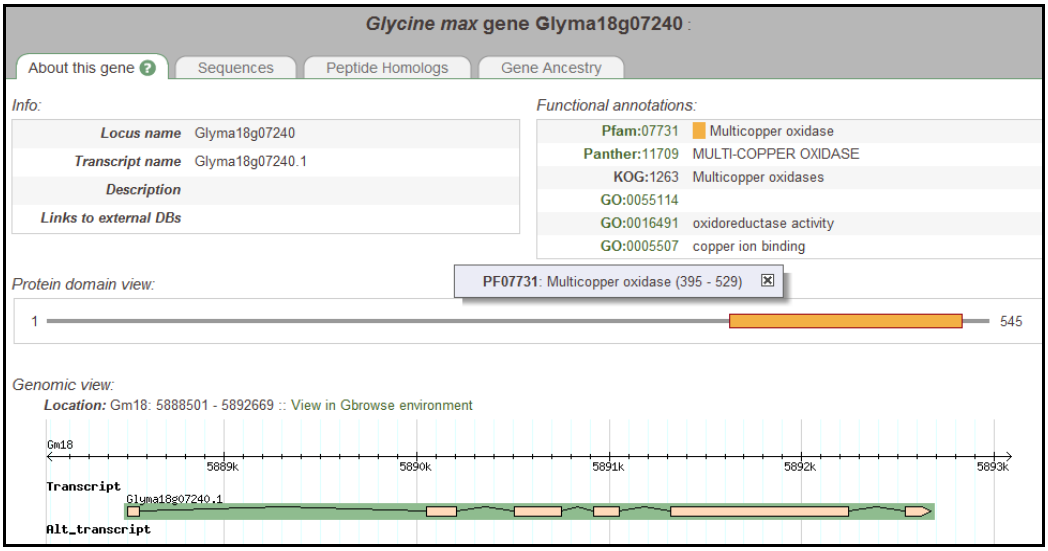


satt231


Glycine max gene Glyma05g10870 :
About this gene
Sequences
Peptide Homologs
Gene Ancestry
Info:
Locus name Glyma05g10870
Transcript name Glyma05g10870.1
Description
Links to external DBs
Functional annotations:
Pfam:00035 Double-stranded RNA binding motif
Panther:11207 RIBONUCLEASE III
GO:0003725 double-stranded RNA binding
GO:0005622 intracellular
Protein domain view:
1
140
Genomic view:
Location: Gm05: 11181251 - 11184286 :: View in Gbrowse environment
Gm05
11181k
11182k
11183k
11184k
Transcript
Glyma05g10870.1
Alt_transcript

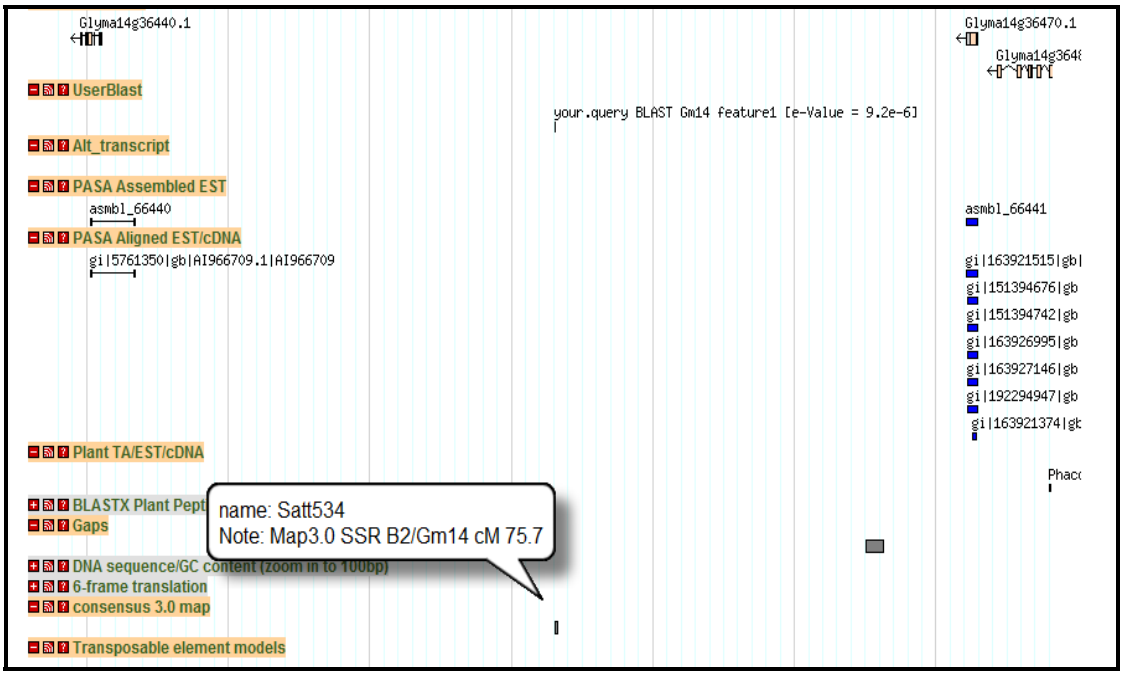
satt324

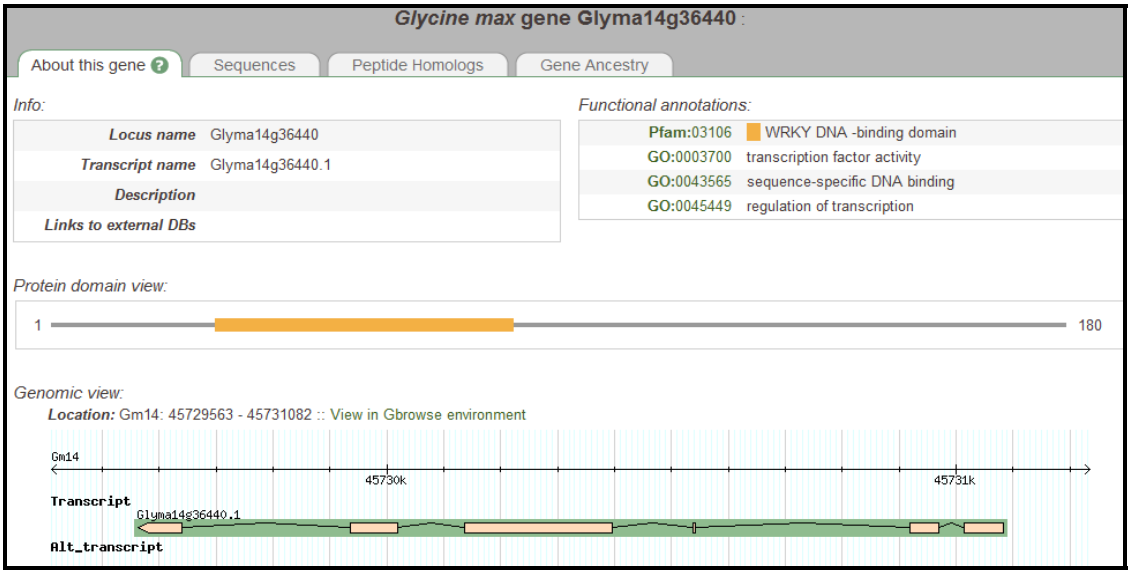

Buscar
Punto o región de referencia:
Gm18:5880126..5900125
Buscar
Fuente de datos
Glycine max
Reportes & análisis:
Anotar VISTA Plot A. thaliana (TAIR9)
Configurar...
Ir
Avanzar/Acercar:
Mostrar 20 kbp
Dar la vuelta
Resumen
Gm18
consensus 3.0 map
Detalles
RepeatMasker
Transcript
Glyma18g07240.1
UserBlast
your.query BLAST Gm18 feature1 [e-Value = 3.9e-4]
Alt_transcript
PASA Assembled EST
asmbl_80884
PASA Aligned EST/cDNA
GMFL01-17-J19
GMFL01-17-J19-F
GMFL02-28-B11-F
gi|58024828|gb|CX711569.1|CX711569
GMFL01-17-J19-R
GMFL02-28-B11-R
12V_018178_3480_2250
C_036995_0368_3114
C_036995_0368_3114
Plant TA/EST/cDNA
BLASTX Plant Peptid
Gaps
name: Satt324
Note: Map3.0 SSR G/Gm18 cM 35.4
DNA sequence/GC content (200nt in to 100bp)
6-frame translation
consensus 3.0 map
Transposable element models
log(1+Solexa coverage depth)
low confidence predictions
Glyma18g07230.1
VISTA Plot: Soybean (dupl.)
VISTA Plot: A. thaliana (TAIR9)


Glycine max gene Glyma18g07240 :
About this gene
Sequences
Peptide Homologs
Gene Ancestry
Info:
Locus name Glyma18g07240
Transcript name Glyma18g07240.1
Description
Links to external DBs
Functional annotations:
Pfam:07731 Multicopper oxidase
Panther:11709 MULTI-COPPER OXIDASE
KOG:1263 Multicopper oxidases
GO:0055114
GO:0016491 oxidoreductase activity
GO:0005507 copper ion binding
Protein domain view:
PF07731: Multicopper oxidase (395 - 529)
1
545
Genomic view:
Location: Gm18: 5888501 - 5892669 :: View in Gbrowse environment
Gm18
5889k
5890k
5891k
5892k
5893k
Transcript
Glyma18g07240.1
Alt_transcript

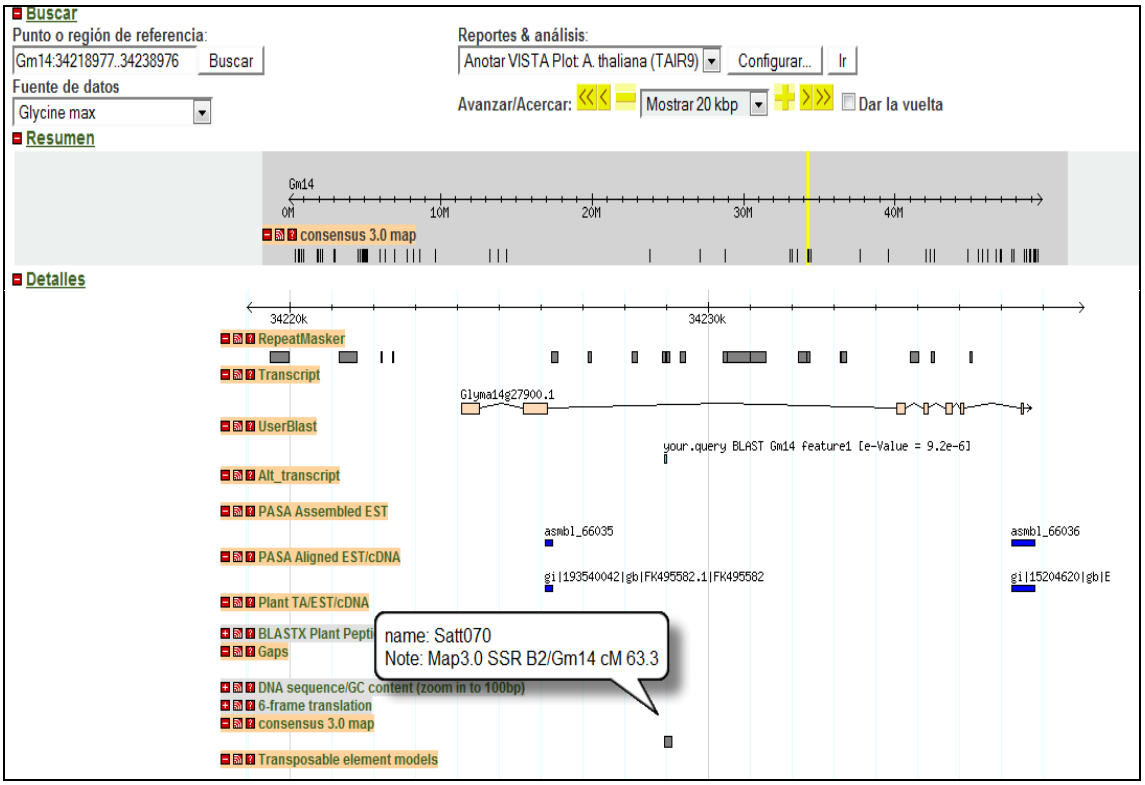
satt534

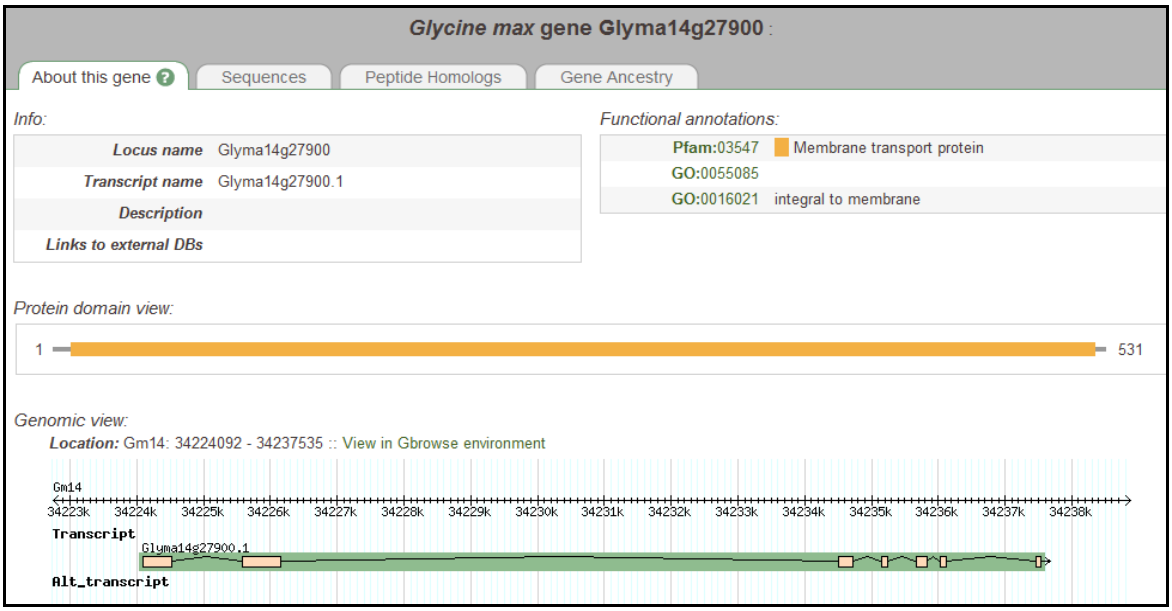

Buscar
Punto o región de referencia:
Gm14:45753095..45773094
Buscar
Fuente de datos
Glycine max
Reportes & análisis:
Anotar VISTA Plot A. thaliana (TAIR9)
Configurar...
Ir
Avanzar/Acercar:
Mostrar 20 kbp
Dar la vuelta
Resumen
Gm14
0M
10M
20M
30M
40M
consensus 3.0 map
Detalles
45760k
45770k
RepeatMasker
Transcript
UserBlast
your.query BLAST Gm14 feature1 [e-Value = 9.2e-6]
Alt_transcript
PASA Assembled EST
PASA Aligned EST/cDNA
Plant TA/EST/cDNA
BLASTX Plant Pepti
Gaps
DNA sequence/GC content (zoom in to 100bp)
6-frame translation
consensus 3.0 map
name: Satt534
Note: Map3.0 SSR B2/Gm14 cM 75.7


Glyma14g36440.1
Glyma14g36470.1
UserBlast
your.query BLAST Gm14 feature1 [e-Value = 9.2e-6]
Alt_transcript
PASA Assembled EST
asmbl_66440
asmbl_66441
PASA Aligned EST/cDNA
gi|5761350|gb|AI966709.1|AI966709
gi|163921515|gb|
gi|151394676|gb
gi|151394742|gb
gi|163926995|gb
gi|163927146|gb
gi|192294947|gb
gi|163921374|gb
Plant TA/EST/cDNA
BLASTX Plant Pept
Gaps
DNA sequence/GC content (zoom in to 100bp)
6-frame translation
consensus 3.0 map
Transposable element models
name: Satt534
Note: Map3.0 SSR B2/Gm14 cM 75.7


Glycine max gene Glyma14g36440 :
About this gene
Sequences
Peptide Homologs
Gene Ancestry
Info:
Locus name Glyma14g36440
Transcript name Glyma14g36440.1
Description
Links to external DBs
Functional annotations:
Pfam:03106 WRKY DNA -binding domain
GO:0003700 transcription factor activity
GO:0043565 sequence-specific DNA binding
GO:0045449 regulation of transcription
Protein domain view:
1
180
Genomic view:
Location: Gm14: 45729563 - 45731082 :: View in Gbrowse environment
Gm14
45730k
45731k
Transcript
Glyma14g36440.1
Alt_transcript

satt70


Buscar
Punto o región de referencia:
Gm14:34218977..34238976
Buscar
Fuente de datos
Glycine max
Reportes & análisis:
Anotar VISTA Plot A. thaliana (TAIR9)
Configurar...
Ir
Avanzar/Acercar:
Mostrar 20 kbp
Dar la vuelta
Resumen
Gm14
0M
10M
20M
30M
40M
consensus 3.0 map
Detalles
34220k
34230k
RepeatMasker
Transcript
Glyma14g27900.1
UserBlast
your.query BLAST Gm14 feature1 [e-Value = 9.2e-6]
Alt_transcript
PASA Assembled EST
asmbl_66035
asmbl_66036
PASA Aligned EST/cDNA
gi|193540042|gb|FK495582.1|FK495582
gi|15204620|gb|E
Plant TA/EST/cDNA
BLASTX Plant Pepti
Gaps
name: Satt070
Note: Map3.0 SSR B2/Gm14 cM 63.3
DNA sequence/GC content (zoom in to 100bp)
6-frame translation
consensus 3.0 map
Transposable element models



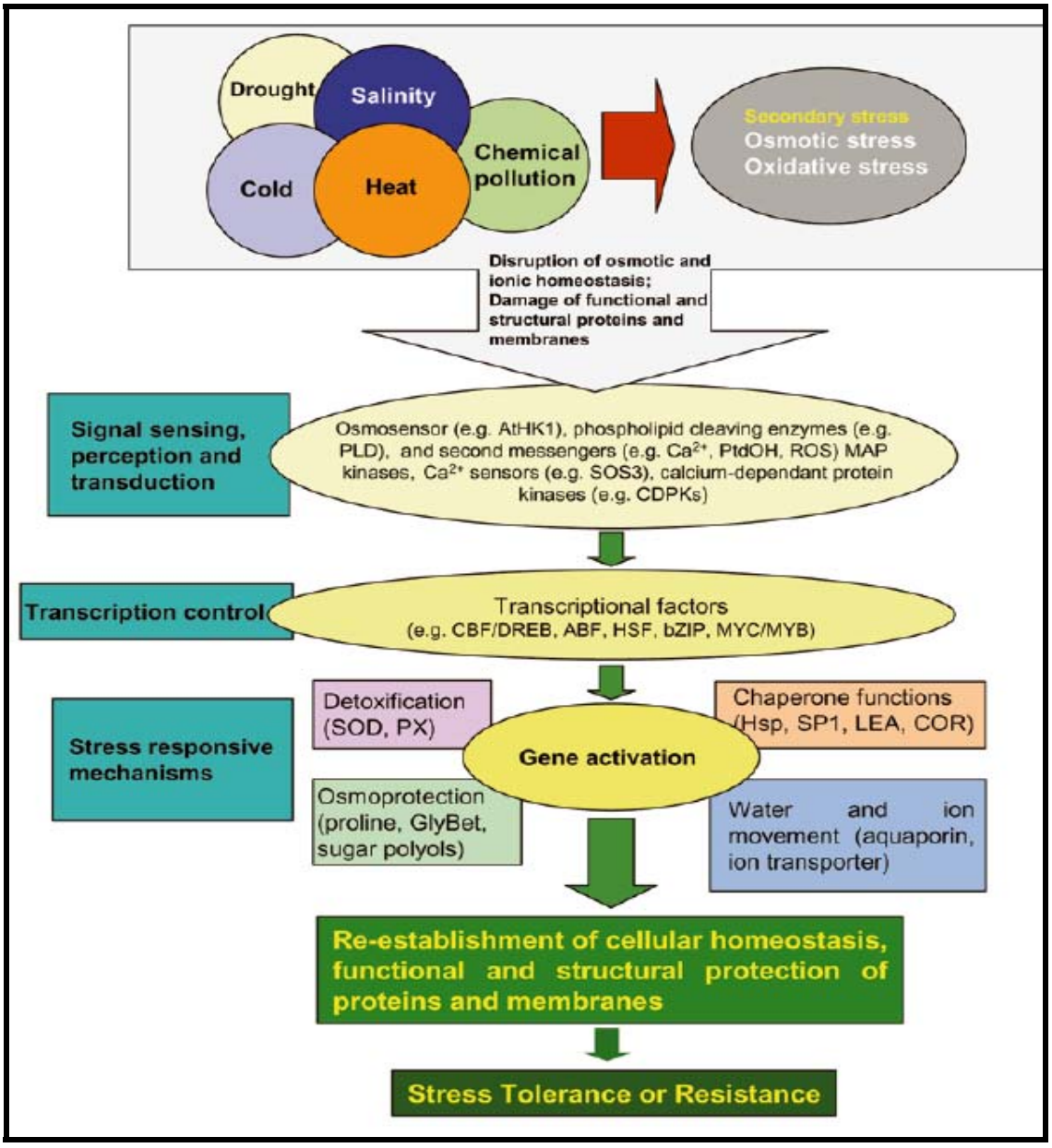
9.6 ANEXO 6: Planilla con los 44 marcadores funcionales anotados manualmente

Accesión	Sequence ID	Repetido	GO process (GOntology)	e-value
NP566897.1	>Contig60	pentanucleotide	Abscisic acid mediated signaling (ARAC8/ ATROP10 / ROP10)	2E-59
ABQ81887.1	>Contig75	trinucleotide	Response to abscisic acid stimulus (KS- type dehydrin SLTI629) G. max	0,028
	>Contig75	trinucleotide		
	>Contig75	dinucleotide		
NP181507.1	>Contig99	tetranucleotide	Response to heat, to ozone, to salt stress, to jasmonic acid stimulus (CYT1-cytokinesis defective 1)	3E-103
NP197570.1	>Contig177	trinucleotide	Biological process (senescence associated protein- related)	1E-17
NP193898.3	>Contig360	pentanucleotide	Response to abiotic stimulus (CPL1)	2E-8
NP567042.1	>Contig430	trinucleotide	Abscisic acid mediated signaling (UBA2A; RNA binding)	1E-42
	>Contig430	trinucleotide		
NP188929.1	>Contig520	trinucleotide	Response to salt stress (APS1) ATP sulfurylase 3	2E-120
NP565107.1	>Contig671	pentanucleotide	Response to oxidative stress (Isoflavone reductase)	1E-46
NP001078346.1	>Contig787	pentanucleotide	Response to oxidative stress, to cold, etc. (SAG 21-senescence associated gene 21)	7E-8

NP196972.1	>Contig807	pentanucleotide	Response to stress (USP)		7E-74
AAA75414.1	>Contig894	tetranucleotide	Response to cold (transcription factor bZip 52) G.max		2E-105
NP188945.1	>Contig951	dinucleotide	Water deprivation (IAA7-auxin resistent 2-transcription factor)		3E-82
NP567042.1	>Contig1333	trinucleotide	abscisic acid mediated signaling	(UBA2A ,RNA binding)	2E-4
NP200911.1	>Contig1366	trinucleotide	response to cold	(GR-RBP3- glicine rich RNA binding protein)	0,27
NP189499.1	>Contig1502	trinucleotide	response to abscisic acid stimulus	(AAA type ATP ase family protein)	4E-64
NP196746.2	>Contig1648	pentanucleotide	response to water deprivation	(protein kinase family protein)	1E-34
NP181401.1	>Contig1990	pentanucleotide	response to cold,related to freezing tolerance	(PECT1-fosforiletanolamine cytidyltransferase 1)	5E-40
NP200414.1	>Contig2193	tetranucleotide	response to heat, to water deprivation	(HSP81-2)	9E-108
NP171654.1	>Contig2262	dinucleotide	response to desication	(LEA14-late embryogenesis abundant 14)	6E-28
NP566050.1	>Contig2389	trinucleotide	response to salt stress	response to ABA stimulus	5E-46
NP568755.1	>Contig2595	dinucleotide	response to cold	(AP2 domain-containing transcription factor)	1E-16
	>Contig2595	trinucleotide			
NP193326.1	>Contig2713	pentanucleotide	response to water deprivation, to cold ,	to oxidative stress, to reactive oxygen species	5E-5
NP195236.1	>Contig2757	trinucleotide	response to salt stress	(PIP3-plasma membrane intrinsic protein 3)	2E-99
NP177745.1	>Contig2764	trinucleotide	early response to dehydration 14	(ERD14)	1E-8
NP850017.1	>Contig2783	trinucleotide	response to cold	(glycine rich RNA binding protein)	7E-33
	>Contig2783	trinucleotide			
NP192343.1	>Contig2830	trinucleotide	early responsive to dehydration protein related	(ERD protein related)	2E-29
NP192056.1	>Contig2844	tetranucleotide	response to salt stress, response to ionic osmotic stress	GAMMA-TIP3/TIP1	1E-104
NP176602.1	>Contig2860	trinucleotide	response to cold	(VHA-E3)	8E-71
NP174810.1	>Contig2912	dinucleotide	response to osmotic stress	(ANNAT1- annexin arabidopsis 1)	7E-74
NP181221.1	>Contig2990	dinucleotide	response to salt stress	GAMMA TIP ,water channel)	9E-62
NP189283.1	>Contig3215	tetranucleotide	response to salt stress,aquoporin expression to environmental stress	GAMMA TIP2	4E-95
ABH02827.1	>Contig3253	trinucleotide	response to salt stress	USP	1E-67
	>Contig3253	trinucleotide			
	>Contig3253	trinucleotide			
NP182011.1	>Contig3282	dinucleotide	ethylene mediated signaling pathway	(ATERF13-EREBP)	9E-37

NP196972.1	>Contig3353	pentanucleotide	USP	universal stress protein	7E-29
NP195197.1	>Contig3442	trinucleotide	response to osmotic,oxidative,salt stress	ADC2-arginine decarboxylase 2)	7E-46
NP192763.1	>Contig3501	trinucleotide	response to heat shock	(AtHSP 22.0)	2E-26
NP177563.1	>Contig3528	trinucleotide	freezing tolerance	(GR-RBP5- glycine rich RNA binding protein 5)	3E-23
	>Contig3528	trinucleotide			
NP174468	>Contig3583	trinucleotide	dehydration responsive protein	ERD-3	7E-68
NP566108.1	>Contig3603	trinucleotide	USP-universal stress protein	(USP family protein)	6E-47
AF169205.1	>Contig3609	trinucleotide	response to cold	(GR- RNA binding protein)	7E-19
	>Contig3609	trinucleotide			
	>Contig3609	trinucleotide			
AAV51938.1	>Contig3753	trinucleotide	salt stress	AP2/EREBP transcription factor ERF-1)	8E-23
	>Contig3753	trinucleotide			
AAS67006.1	>Contig3797	dinucleotide	salt stress (ionic osmotic stress)	(phosphoenolpyruvate carboxylase)	2E-33
NP192839.1	>Contig3829	trinucleotide	oxidative stress	(NDPK3-nucleoside diphosphate kinase 3)	7E-87
NP200137.1	>Contig3835	trinucleotide	drought stress	(AGP22/ATAGP22-arabinogalactan proteins 22)	3E-5

9.7 ANEXO 7: Proceso de estrés abiótico en plantas



Tomado de **Wang et al. 2003**-Plant responses to drought, salinity and extreme temperatures: towards genetic engineering for stress tolerance. Planta 218:1-14. (Page 2)

9.8 ANEXO 8: Resultados del análisis discriminante mediante uso de SAS

1a)

DISCRIMINANTE CON SELECCIÓN DE VARIABLES

OBS	CULTIVAR	SATT173_213	SATT175_205	SATT231_235	SATT231_259	SATT534_246	SATT534_270
1	DM4200	1	0	0	0	0	1
2	NM55R	1	0	0	1	0	1
3	NM70R	1	0	0	0	0	0
4	RA514	1	1	0	0	0	0
5	RA518	1	0	1	0	0	0
6	TJS2055	1	0	0	0	1	0
7	TJS2178R	0	0	0	0	0	0
8	DM4200	1	0	0	0	0	1
9	NM55R	1	0	0	1	0	1
10	NM70R	1	0	0	0	0	0
11	RA514	1	1	0	0	0	0
12	RA518	1	0	1	0	0	0
13	TJS2055	1	0	0	0	1	0
14	TJS2178R	0	0	0	0	0	0
15	DM4200	1	0	0	0	0	1
16	NM55R	1	0	0	1	0	1
17	NM70R	1	0	0	0	0	0
18	RA514	1	1	0	0	0	0
19	RA518	1	0	1	0	0	0
20	TJS2055	1	0	0	0	1	0
21	TJS2178R	0	0	0	0	0	0

1b)

SISTEMA SAS

PROCEDIMIENTO DISCRIM

OBSERVACIONE	21	TOTAL DF	20
VARIABLES	6	CLASES WITHIN DF	14
CLASES	7	CLASES BETWEEN DF	6

INFORMACIÓN DEL NIVEL DE LA CLASE

CULTIVAR	NOMBRE DE VARIABLE	FRECUENCIA	PESO	PROPORCIÓN	PROBABILIDAD ANTERIOR
DM4200	DM4200	3	3.0000	0.142857	0.142857
NM55R	NM55R	3	3.0000	0.142857	0.142857
NM70R	NM70R	3	3.0000	0.142857	0.142857
RA514	RA514	3	3.0000	0.142857	0.142857
RA518	RA518	3	3.0000	0.142857	0.142857
TJS2055	TJS2055	3	3.0000	0.142857	0.142857
TJS2178R	TJS2178R	3	3.0000	0.142857	0.142857

2a)

SISTEMA SAS

PROCEDIMIENTO DISCRIM
RESUMEN DE CLASIFICACIÓN PARA LOS DATOS CALIBRADOS: WORK.SOJA_ALELOS
RESUMEN DE RESUSTITUCIÓN USANDO VECINO MÁS CERCANO

FUNCIÓN DE LA DISTANCIA CUADRADA

$$D^2(X,Y) = (X-Y)' \, COV^{-1} \, (X-Y)$$

PROBABILIDAD POSTERIOR DE MIEMBRO EN CADA CULTIVAR

$M_K(X)$ = PROPORCIÓN DE OBS EN GRUPO K EN EL VECINO MÁS CERCANO DE X

$$PR(J|X) = M_J(X) \, PRIOR_J \; / \; SUM_K (M_K(X) \, PRIOR_K)$$

NÚMERO DE OBSERVACIONES Y PORCENTAJE CLASIFICADO EN CULTIVAR

DE CULTIVAR	DM4200	NM55R	NM70R	RA514	RA518	TJS2055	TJS2178R	TOTAL
DM4200	3	0	0	0	0	0	0	3
	100.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00
NM55R	0	3	0	0	0	0	0	3
	0.00	100.00	0.00	0.00	0.00	0.00	0.00	100.00
NM70R	0	0	3	0	0	0	0	3
	0.00	0.00	100.00	0.00	0.00	0.00	0.00	100.00
RA514	0	0	0	3	0	0	0	3
	0.00	0.00	0.00	100.00	0.00	0.00	0.00	100.00
RA518	0	0	0	0	3	0	0	3
	0.00	0.00	0.00	0.00	100.00	0.00	0.00	100.00
TJS2055	0	0	0	0	0	3	0	3
	0.00	0.00	0.00	0.00	0.00	100.00	0.00	100.00
TJS2178R	0	0	0	0	0	0	3	3
	0.00	0.00	0.00	0.00	0.00	0.00	100.00	100.00
TOTAL	3	3	3	3	3	3	3	21
	14.29	14.29	14.29	14.29	14.29	14.29	14.29	100.00
ANTERIORES	0.14286	0.14286	0.14286	0.14286	0.14286	0.14286	0.14286	

2b)

SISTEMA SAS
PROCEDIMIENTO DISCRIM
RESUMEN DE CLASIFICACIÓN PARA LOS DATOS CALIBRADOS: WORK.SOJA_ALELOS
RESUMEN DE RESUSTITUCIÓN USANDO VECINO MÁS CERCANO

ESTIMACIONES DE CUENTA DE ERROR PARA CULTIVAR

	DM4200	NM55R	NM70R	RA514	RA518	TJS2055	TJS2178R	TOTAL
TASA	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ANTERIORES	0.1429	0.1429	0.1429	0.1429	0.1429	0.1429	0.1429	

3a)

SISTEMA SAS

OBS	CULTIVAR	LAU1_139	LAU2_141	P127_325	Q43453_289
1	DM4200	0	0	0	1
2	NM55R	0	0	1	1
3	NM70R	0	0	1	1
4	RA514	1	0	1	1
5	RA518	1	0	1	1
6	TJS2055	1	0	1	0
7	TJS2178R	1	1	1	1
8	DM4200	0	0	0	1
9	NM55R	0	0	1	1
10	NM70R	0	0	1	1
11	RA514	1	0	1	1
12	RA518	1	0	1	1
13	TJS2055	1	0	1	0
14	TJS2178R	1	1	1	1
15	DM4200	0	0	0	1
16	NM55R	0	0	1	1
17	NM70R	0	0	1	1
18	RA514	1	0	1	1
19	RA518	1	0	1	1
20	TJS2055	1	0	1	0
21	TJS2178R	1	1	1	1

3b)

SISTEMA SAS
PROCEDIMIENTO DISCRIM

OBSERVACIONE	21	TOTAL DF	20
VARIABLES	4	CLASES WITHIN DF	14
CLASES	7	CLASES BETWEEN DF	6

INFORMACIÓN DEL NIVEL DE LA CLASE

NOMBRE DE CULTIVAR	VARIABLE	FRECUENCIA	PESO	PROBABILIDAD PROPORCIÓN	PROBABILIDAD ANTERIOR
DM4200	DM4200	3	3.0000	0.142857	0.142857
NM55R	NM55R	3	3.0000	0.142857	0.142857
NM70R	NM70R	3	3.0000	0.142857	0.142857
RA514	RA514	3	3.0000	0.142857	0.142857
RA518	RA518	3	3.0000	0.142857	0.142857
TJS2055	TJS2055	3	3.0000	0.142857	0.142857
TJS2178R	TJS2178R	3	3.0000	0.142857	0.142857

4a)

SISTEMA SAS
PROCEDIMIENTO DISCRIM
RESUMEN DE CLASIFICACIÓN PARA LOS DATOS CALIBRADOS: WORK.SOJA_ALELOS
RESUMEN DE RESUSTITUCIÓN USANDO VECINO MÁS CERCANO

FUNCIÓN DE LA DISTANCIA CUADRADA

$$D^2(X,Y) = (X-Y)' COV^{-1} (X-Y)$$

PROBABILIDAD POSTERIOR DE MIEMBRO EN CADA CULTIVAR

$M_K(X)$ = PROPORCIÓN DE OBS EN GRUPO K EN EL VECINO MÁS CERCANO DE X

$$PR(J|X) = M_J(X) PRIOR_J / SUM_K (M_K(X) PRIOR_K)$$

NÚMERO DE OBSERVACIONES Y PORCENTAJE CLASIFICADO EN CULTIVAR

DE CULTIVAR	DM4200	NM55R	NM70R	RA514	RA518	TJS2055	TJS2178R	OTRO	TOTAL
DM4200	3	0	0	0	0	0	0	0	3
	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00
NM55R	0	0	0	0	0	0	0	3	3
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	100.00
NM70R	0	0	0	0	0	0	0	3	3
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	100.00
RA514	0	0	0	0	0	0	0	3	3
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	100.00
RA518	0	0	0	0	0	0	0	3	3
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	100.00
TJS2055	0	0	0	0	0	3	0	0	3
	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	100.00
TJS2178R	0	0	0	0	0	0	3	0	3
	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	100.00
TOTAL	3	0	0	0	0	3	3	12	21
	14.29	0.00	0.00	0.00	0.00	14.29	14.29	**57.14**	100.00
ANTERIORES	0.14286	0.14286	0.14286	0.14286	0.14286	0.14286	0.14286		

4b)

SISTEMA SAS
PROCEDIMIENTO DISCRIM
RESUMEN DE CLASIFICACIÓN PARA LOS DATOS CALIBRADOS: WORK.SOJA_ALELOS
RESUMEN DE RESUSTITUCIÓN USANDO VECINO MÁS CERCANO

ESTIMACIONES DE CUENTA DE ERROR PARA CULTIVAR

	DM4200	NM55R	NM70R	RA514	RA518	TJS2055	TJS2178R	TOTAL
TASA	0.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	**0.5714**
ANTERIORES	0.1429	0.1429	0.1429	0.1429	0.1429	0.1429	0.1429	

5a)

SISTEMA SAS
PROCEDIMIENTO DISCRIM
RESUMEN DE CLASIFICACIÓN PARA LOS DATOS CALIBRADOS: WORK.SOJA_ALELOS
RESUMEN DE VALIDACIÓN CRUZADA USANDO VECINO MÁS CERCANO

FUNCIÓN DE LA DISTANCIA CUADRADA

$$D^2(X,Y) = (X-Y)' \, COV^{-1} \, (X-Y)$$

PROBABILIDAD POSTERIOR DE MIEMBRO EN CADA CULTIVAR

$M_K(X)$ = PROPORCIÓN DE OBS EN GRUPO K EN EL VECINO MÁS CERCANO DE X

$$PR(J|X) = M_J(X) \, PRIOR_J \, / \, SUM_K (M_K(X) \, PRIOR_K)$$

NÚMERO DE OBSERVACIONES Y PORCENTAJE CLASIFICADO EN CULTIVAR

DE CULTIVAR	DM4200	NM55R	NM70R	RA514	RA518	TJS2055	TJS2178R	OTRO	TOTAL
DM4200	3	0	0	0	0	0	0	0	3
	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00
NM55R	0	0	0	0	0	0	0	3	3
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	100.00
NM70R	0	0	0	0	0	0	0	3	3
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	100.00
RA514	0	0	0	0	0	0	0	3	3
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	100.00
RA518	0	0	0	0	0	0	0	3	3
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	100.00
TJS2055	0	0	0	0	0	3	0	0	3

	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00
100.00								
TJS2178R	0	0	0	0	0	0	3	0
3								
	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00
100.00								
TOTAL	3	0	0	0	0	3	3	12
21								
	14.29	0.00	0.00	0.00	0.00	14.29	14.29	57.14
100.00								
ANTERIORES	0.14286	0.14286	0.14286	0.14286	0.14286	0.14286	0.14286	

5b)

SISTEMA SAS
PROCEDIMIENTO DISCRIM
RESUMEN DE CLASIFICACIÓN PARA LOS DATOS CALIBRADOS: WORK.SOJA_ALELOS
RESUMEN DE VALIDACIÓN CRUZADA USANDO VECINO MÁS CERCANO

ESTIMACIONES DE CUENTA DE ERROR PARA CULTIVAR

	DM4200	NM55R	NM70R	RA514	RA518	TJS2055	TJS2178R	TOTAL
TASA	0.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.5714
ANTERIORES	0.1429	0.1429	0.1429	0.1429	0.1429	0.1429	0.1429	