Construcción de un índice de genes con anotaciones funcionales consistentes para la interpretación de experimentos de expresión génica en girasol

[1]P. Fernández, [2]M. Soria, [3]D. Príncipi, [3]S. González, [3]S. Lew, [1]R. Heinz y[1]N. Paniego

[1]Instituto de Biotecnología, INTA Castelar, Hurlingham. Pcia. de Bs. As. ARGENTINA
[2]Facultad de Agronomía, Universidad de Buenos Aires, ARGENTINA
[3]Instituto de Ingeniería Biomédica, FI, UBA, Bs. As., ARGENTINA

Resumen

Los proyectos genómicos desarrollados durante la última década han incrementado exponencialmente el número de secuencias disponibles en bases de datos públicas tanto de genomas completos como de ESTs, incluido el girasol. Para esta especie se dispone de más de 130.000 secuencias en GenBank, y existen pocos estudios desarrollados y disponibles para determinar la identidad funcional de dichas secuencias. Dado que esta información es clave para interpretar los resultados de los análisis conducidos a nivel transcripcional, el objetivo de este trabajo es el análisis bioinformático de secuencias parciales expresadas (ESTs) que serán objeto de posteriores estudios de expresión génica para caracteres de importancia agronómica a partir del diseño de un microarreglo de oligonucleótidos. Para ello, inicialmente se realizó un proceso de depuración de EST, ensamblado, anotación y re-evaluación de contigs para filtrar aquellos que presentaban redundancia y podrían conducir a la observación de patrones de expresión difíciles de interpretar durante el análisis del microarreglo. De manera exploratoria, se analizó un subconjunto de unigenes a través de un análisis de expresión digital diferencial a partir de la abundancia de transcriptos correspondientes presentes en diferentes clonotecas, para evaluar de manera preliminar la información contenida en el chip de girasol. Los resultados de estos análisis tendrán impacto en el conocimiento de expresión diferencial de posibles genes que serán genuinamente validados a posteriori de manera experimental bajo la utilización de la micromatriz de Helianthus annuus L. sintetizada y en proceso de validación.

Abstract

Plant genome projects developed during the last decade allowed the exponentially increased of public sequences in web databases not only of whole-sequenced genomes but also ESTs sequence for different species, including sunflower. Although more than 130,000 public ESTs sequences are available in GenBank, there are few preliminary studies aimed to elucidate the functional identity of these anonymous sequences.

Considering that this information is important to carry out robust interpretation of transcriptional studies, we propose to obtain a reliable set of sunflower unigenes by applying different bioinformatic cleaning procedures to public ESTs databases. This unigene data base represents the start point in the design of an oligonucleotide microarray for transcriptomic studies. As a preliminary exploratory assay, a subset of these unigenes derived from different cDNA libraries was analyzed using automatic annotation methodologies and a digital expression analysis was conducted. The outcome results will help to have additional information about the data contained in the unigene database for further application to Helianthus annuus L transcriptomic analysis using an oligonucleotide microarray which is already printed and synthesized and in the process of validation.

Palabras Clave: girasol, análisis digital, EST

1. Introducción

1.1. Necesidad del desarrollo bioinformático en la genómica funcional

La Bioinformática es una disciplina científica que surge para cubrir los requerimientos de manejo de información e interpretación de los resultados de los proyectos genómicos y de disciplinas relacionadas tales como la transcriptómica, la proteómica, la metabolómica, el mapeo genético, la genética reversa, etc. Los datos provenientes de las áreas mencionadas constituyen conjuntos de información de gran dimensión, resultando fundamental el proceso de organización, análisis e integración de los mismos para el aprovechamiento eficiente que permitirá generar nuevas hipótesis de trabajo y diseñar adecuadamente experimentos a campo. En este contexto la gestión de la información en el área de las disciplinas post-genómicas se ha convertido en uno de los mayores desafíos que enfrenta la tecnología de la información desde los comienzos de este siglo.

En la Argentina, diversos grupos de investigación realizan esfuerzos para impulsar el área de genómica funcional en diferentes instituciones. El éxito de esos esfuerzos requiere llevar a cabo proyectos interdisciplinarios que involucren el trabajo de especialistas en estadística y tecnologías de información para el desarrollo de metodologías y herramientas para el análisis apropiado al gran volumen de datos que estas tecnologías generan [1]. Si bien la tecnología de los microarreglos de ADN y/u oligonucleótidos ha sido un pilar para el descubrimiento de la funcionalidad de un gran número de genes y/o sus mecanismos de regulación, aún permanecen muchas incertidumbres a la hora de analizar la información obtenida a partir de un valor de fluorescencia, y su correspondiente significancia biológica. La sinergia multidisciplinaria entre estadística, biología molecular e informática emergen como necesarias para dar sustento al uso e interpretación de los resultados obtenidos utilizando micromatrices en el área de genómica funcional.

Los proyectos genómicos desarrollados durante la última década han originado un incremento exponencial en el número de secuencias disponibles en bases de datos públicas tanto a nivel de genomas completos, en el caso de genomas vegetales para (*Arabidopsis thaliana*, *Oryza sativa* y *O. japonica*) o secuencias correspondientes a regiones expresadas o ESTs, en particular para especies de genomas grandes, como el girasol. Para esta última especie se dispone de más de 130.000 ESTs depositados en GenBank [2] para girasol cultivado, no se dispone de la secuencia del genoma y toda la fuente de información genómica asociada proviene de estas secuencias públicas anónimas respecto de su funcionalidad.

1.2. Principales contribuciones al tema por parte del grupo del proyecto.

El grupo de trabajo de Genómica Aplicada del Instituto de Biotecnología ha desarrollado en el marco de distintos proyectos conducidos bajo el programa “Análisis Genómico de Girasol” herramientas moleculares que incluyen marcadores neutros SSR altamente polimórficos [3], un banco de EST órgano-específicos aislados de colecciones de ADN copia usando la tecnología de hibridación sustractiva de secuencias (SSH-ESTs) [4], desarrollo de un mapa saturado de referencia para girasol cultivado [5], estudios de diversidad basados en identificación de SNP/InDels [6, 7] y mapeo de QTLs e

identificación de genes candidatos para resistencia al patógeno *Sclerotinia sclerotiorum* [8, 9]. Recientemente se exploró una estrategia de análisis concertado de expresión génica a través de la tecnología de microarreglos de ADNc, utilizando una colección de ESTs locales que representan secuencias únicas del banco de ESTs desarrollado anteriormente [4], con el propósito de identificar nuevos genes candidatos relacionados con la resistencia a estreses abióticos. Este trabajo consistió un diseño de tres réplicas biológicas para la evaluación de las respuestas iniciales de la planta de girasol a la exposición a bajas temperaturas y salinidad. De estos candidatos iniciales, que incluyen genes potencialmente involucrados en procesos de transcripción, traducción, degradación/plegado o interacción de proteínas o asociados a mecanismos de generación y procesamiento de especies reactivas de oxigeno, 10 fueron validados mediante la técnica de PCR cuantitativa (qPCR) y/o northern blot [10, 11].
A partir de un proyecto financiado por INTA dentro del marco del Plan Estratégico Institucional 2005-2015, fue posible consolidar la creación de una unidad de bioinformática, que cuenta con equipamiento y herramientas de bioinformática para asistir los procesos de anotación de genes, el análisis de experimentos de expresión concertada de genes y la identificación de genes candidatos, entre otros.

1.3. Aporte bioinformático a la expresión diferencial: el análisis digital

El acceso al conocimiento de estas secuencias posibilita el desarrollo de estudios de interpretación de la información codificada en tales genomas y predecir su función. Este último aspecto, abordado por la genómica funcional, combina el análisis de la transcriptómica con el de otras disciplinas postgenómicas como la proteómica y la metabolómica en condiciones biológicas definidas, pretendiendo definir la identidad, la función y la regulación de genes en determinados procesos biológicos. Las tecnologías utilizadas por la genómica funcional requieren cuantiosos trabajos de laboratorio a una escala que ha sido posible alcanzar mediante la automatización de los procesos involucrados. Es por ello que los estudios de expresión digital ofrecen información preliminar sobre el nivel de expresión de diferentes grupos o “clusters” de secuencias inicialmente anónimas respecto de su función, generando datos aproximados de sobre o sub expresión de ESTs aún entre clonotecas de diferentes órganos o tratamientos [12]. Este análisis permite la detección y/o cuantificación de transcriptos aislados de diferentes órganos, tejidos o células incluyendo diferentes condiciones y/o tratamientos a través de una estimación estadística de grupos o subgrupos de genes agrupados en un set de secuencias, generalmente de tipo EST [13].
Este tipo de análisis ha sido ampliamente llevado a cabo como medida de acercamiento a expresión diferencial de ESTs anónimos en plantas de especies no-modelo [14, 15] de manera de obtener información funcional sobre secuencias públicas carentes de anotación.

Elementos del Trabajo y Metodología

Curado, ensamble y anotación de secuencias

Las sondas que van a formar parte de un microarreglo deben reunir tres características esenciales: deben ser sensibles, isotérmicas y específicas [16]. Existen algunas bases de datos de unigenes desarrolladas a partir de ESTs de girasol. Para este trabajo se propone utilizar; la información generada por el proyecto de índice de genes de DFCI [17], considerando los ESTs disponibles. Para el género y especie *Helianthus annuus* existen 133.682 secuencias correspondientes a ESTs en NCBI. A partir de estas secuencias está contemplado desarrollar un chip en el formato 4 x 44k con sondas de 60meros para la obtención de perfiles de ARN en gran escala.

En la primera etapa del proceso de diseño fue importante filtrar aquellos contigs que presentaban redundancia, porque podrían producir un aumento innecesario del número de genes representados en la micromatriz pudiendo conducir a la observación de patrones de expresión difíciles de interpretar. Para estimar el grado de redundancia presente en la biblioteca seleccionada se realizó un análisis general de anotación y un análisis de comparación de secuencias detallado sobre una muestra representativa de unigenes.. Al mismo tiempo, se evaluó el porcentaje de unigenes de tamaño menor a 100-150 nucleótidos que dificultaban la identificación confiable de sondas inequívocamente representativas del mismo de una longitud máxima de 60 nucleótidos, de acuerdo a la tecnología Agilent.. El número de secuencias final ensambladas para incluir en la micromatriz fue de 40.000 unigenes, que fueron previamente anotados utilizando Blast2Go [18] y posteriormente mapeados en vías metabólicas a través de KEGG.

Ensayo de expresión digital

El análisis digital se realizó incluyendo un subgrupo de las 27 clonotecas analizadas de GenBank con ESTs representativos y públicos originados por diferentes grupos de investigación que proveen de secuencias a este repositorio (Tabla 1). Un total de 133.682 fueron utilizados en el análisis [19].

Identificación de clonoteca	Estadio/Caracterización
HaSSH	Molecular characterization of phosphorus-responsive genes in sunflower
CCF (STU)	EST sequences from several different strains/cultivars
QH-RHA 280/QH_ABCDI sunflower RHA801	shoots/hulls/flowers environmental stress/chemical induction
CHA(XYZ) common wild sunflower	girasol silvestre (wild sunflower)
HaHeaS	heart-shaped embryo vs cotyledonary embryo
HaHeaR	heart-shaped embryo
HaCotR	cotyledonary embryo
HaGlbR	globular embryo
HaDevS1	4 days after self-pollination embryo
HaDevS2	7 days after self-pollination embryo
HaDevR1	leaves
HaDevR2	terminal bud
HaDevR3	stem
HaDevR6	embryo
HaDevR5	4 days after self-pollination embryo
HaDevR8	15 days after self-pollination embryo
HaDis	unknown/cotyledons/ (Genoplante)
HaSemS4	hypocotyl
HaDpsR1	hypocotyl
HaDplR2	hypocotyl 1-5 days
HaDplR	protoplast
HaERF	embryo
HaERS	embryo
HaR	INTA: organ-specific cDNA libraries (root)
HaT	INTA: organ-specific cDNA libraries (stem)
HaEF	INTA: organ-specific cDNA libraries (early flower)
HaF	INTA: organ-specific cDNA libraries (flower)
HaH	INTA: organ-specific cDNA libraries (leaf)

Tabla 1 – Caracterización de las clonotecas originarias de los ESTs públicos de girasol cultivado disponibles en GenBank

Partiendo de 27 clonotecas de ESTs ("n clonotecas"), se procedió al ensamblado de contigs. Para ello se utilizó CAP3 [20] de manera que los contigs resultantes fueran formados por combinación entre secuencias de diferentes orígenes. Posteriormente, se identificó a que clonoteca pertenecía cada una de las secuencias que se utilizaron para ensamblar cada uno de los contigs. Se continuó entonces con las comparaciones, teniendo en cuenta que estas se realizarían de a pares. En nuestro caso, con 27 clonotecas, se realizaron 351 comparaciones (n!/2!(n-2)!) [15]. Dadas dos clonotecas cualesquiera, se procedió a contar cuantos ESTs se usaron para ensamblar cada uno de los contigs. Si la cantidad de ESTs utilizados para cada contig era mayor o igual a 5 se continuaba con el método estadístico, si el número de ESTs era menor o igual a 5, no se consideraba ese contig al comparar las dos clonotecas analizadas. El método estadístico consistió en aplicar un test de Poisson modificado para ser simétrico (p valor < 0,05), de manera de poder establecer si para un contig dado la expresión en una condición A es estadísticamente diferente a la expresión en la condición B [12]. En caso de un resultado positivo para dicho test se procede a la anotación del contig (que en nuestro caso ya habia sido realizada) tal cual fue llevado a cabo utilizando el vocabulario GO [21], de manera de conocer a que posible gen está asociado ese contig, pudiendo finalmente estimar que un cierto gen se encuentra más expresado en una condición que en otra. El proceso fue repetido para todos los pares de clonotecas y para todos los contigs formados.

3. Resultados

3.1. Limpieza, depuración y ensamble de secuencias

La remoción de contaminaciones con secuencias provenientes de vectores y/o adaptadores de uso común en *kits* comerciales de aplicación en biología molecular, es una de las operaciones más significativas en el procesamiento de secuencias para su posterior uso en análisis funcional "*in silico*". Los problemas subyacentes a la calidad de los datos depositados en bases de datos públicas internacionales fueron reportados previamente en los inicios de la era de los proyectos genómicos siendo los principales aquellos asociados con redundancias (múltiples representaciones de un mismo gen), contaminación por vector y/o adaptadores [22] y anotación errónea de sitios intrón/exón o *splicing* [23]. Las secuencias denominadas ESTs (del inglés: *expressed sequence tags*) [24] son secuencias originadas a partir de lecturas de una sola pasada a partir de un ADN copia (ADNc). Los ESTs han sido y son en la actualidad ampliamente utilizados en la predicción de genes, *splicing* alternativo y estudios de expresión diferencial en diferentes órganos y/o estadios de un organismo. Determinar la calidad, pureza y limpieza de estas secuencias (las cuales conforman aproximadamente la mitad de la información disponible en GenBank), constituye un hito fundamental para estudios *in silico* de genómica funcional que se aborden a partir de estos datos como fuente. Si bien existen procedimientos informáticos acordados para hacer este proceso más eficiente, como es el caso de phred [25], Cross_match [26] y LUCY [27] del J. Craig Venter Institute [28], la automatización de los mismos hace que porcentajes relativamente altos de secuencias (2.203 sobre un total de 48.212 ESTs al azar analizados en dbEST [release: 18/04/2007] presentaron contaminaciones según base de datos de Univec [29] sigan arrastrando al final de proceso contaminaciones como secuencias de adaptador intercaladas dentro de la secuencia de interés, o aún en los extremos [30]. La herramienta VecScreen de BLAST [31] es muy

eficiente en la identificación de contaminaciones por vector y/o oligonucleótidos usando como referencia la base de datos Univec que contiene la mayoría de los vectores y oligonucleótidos comerciales usados mayoritariamente.

3.2. Anotación funcional de secuencias públicas

Una vez obtenidas las secuencias públicas limpias, ensambladas y depuradas lo más conveniente para su significancia biológica es anotarlas acorde a un vocabulario controlado que pueda describir de manera universal su función, proceso y/o localización celular. El ejemplo más exitoso de sistematización por nomenclador universal es el proyecto Gene Ontology (GO) [21]. Existen muchas aplicaciones en la internet para poder anotar secuencias públicas acordes al vocabulario de Gene Ontology, siendo las más utilizadas GoMiner [32], FatiGO [33], Gotcha [34], GoFigure [35] y Blast2Go [18], entre otros. Blast2GO (B2G) [18] es una herramienta bioinformática desarrollada para la anotación funcional y el análisis de genes o proteínas sin caracterización previa. Básicamente B2G utiliza el algoritmo BLAST de manera local o remota y efectúa búsquedas contra bases de datos para encontrar secuencias similares a la secuencia de interés (secuencia blanco). El programa permite mapear los resultados obtenidos con Blast contra las bases de datos de GO, y extrae los términos GO asociados a cada uno de los hits obtenidos, devolviendo una anotación controlada para la secuencia blanco. Una rutina básica de uso de B2G consiste de 5 pasos: ejecución de BLAST, mapeo GO, anotación funcional, visualización gráfica de las estadísticas asociadas a los resultados obtenidos.

3.3. Análisis diferencial de ESTs a través de la expresión digital

El análisis de expresión digital permitió observar tres grupos (“clusters”) de genes con una notable diferencia de expresión entre clonotecas. Se observó una diferencia significativa entre ESTs provenientes de clonotecas de embrión, hipocótilo y protoplasto vs. ESTs provenientes de clonotecas de órganos vegetativas en estadios avanzados de desarrollo, y/o órganos verdes maduros y reproductivos (tallo, hoja y flor). Los ESTs aislados de la yema terminal mostraron 13 grupos de genes sobreexpresados contra 7 de los ESTs de embrión, mientras que los de hoja contra los de yema terminal mostraron 2 grupos sobreexpresados frente a 21 (Tabla 2).

x \ y	HaHeaS	HaHeaR	HaCotR	HaGlbR	HaDevS1	HaDevS2	HaDevR1	HaDevR2	HaDevR3	HaDevR6	HaDevR5	HaDevR8	HaDis	HaSemS4	HaDpsR1	HaDplR2	HaDplR	HaERF	HaERS	HaR	HaT	HaEF	HaF	HaH
HaHeaS						0\1		0\1		2\1														
HaHeaR			1\2	1\1			1\1	2\2		4\2			2\1		3\2	1\1								
HaCotR				2\1	1\2	3\1	0\1	1\1		8\2	2\0		2\1	1\1	6\1	0\1								
HaGlbR							1\1	1\2		1\1			1\2	0\1	1\0	1\1								
HaDevS1						0\1	0\1	3\1		7\4	1\0		1\2		1\1	0\1								
HaDevS2								1\1		1\4	1\2				0\1									
HaDevR1								2\21	1\0	3\2	2\0		12\19		1\1		0\1				1\0			
HaDevR2									0\1	13\7	2\5	2\1	9\6	0\1	5\2	0\1								
HaDevR3										1\0		1\0	1\0											
HaDevR6											5\12		6\5	1\1	8\4									
HaDevR5													2\3		0\1	0\1								
HaDevR8																								
HaDis														1\1	4\3								0\1	
HaSemS4															1\0	0\1								
HaDpsR1																0\1	0\1							
HaDplR2																								
HaDplR																								
HaERF																								
HaERS																								
HaR																								
HaT																								
HaEF																								
HaF																								
HaH																								

Tabla 2 – Grupos de genes (“clusters”) diferencialmente expresados entre ESTs de clonotecas órgano/estadio específicos públicas de girasol

Discusión

El girasol pertenece a la familia de las *Compositae*, que es una de las más numerosas y diversas dentro del grupo de las Angiospermas, incluye varias especies que a pesar de ser importantes económicamente no han sido demasiado estudiadas. Los representantes más caracterizados desde el punto de vista genómico y funcional son la lechuga y el girasol. Para estas especies existe disponibilidad de recursos genéticos caracterizados, una colección importante de ESTs, genotecas de insertos grandes en cromosomas artificiales de bacterias (BACs), marcadores moleculares, mapas genéticos y QTLs caracterizados. Existen distintas iniciativas de secuenciación de ESTs que comprometen aproximadamente a 20 especies dentro de la familia *Compositae*, lo cual significa que existe información representativa de la diversidad presente en la familia sustentando la posibilidad de estudios comparativos entre los miembros de la misma.
En la actualidad se está estructurando un consorcio de genómica de la familia *Compositae* (http://compgenomics.ucdavis.edu/cwp), como los desarrollados para otras familias dentro del reino vegetal, que albergan entre sus miembros a especies de interés económico. En este contexto se está desarrollando una micromatriz o chip para el análisis genético y de expresión de transcriptos compatible para realizar este tipo de estudios en lechuga y girasol (http://chiplett.ucdavis.edu/). No obstante la disponibilidad de estos recursos y la existencia de iniciativas genómicas, aun existen limitaciones importantes para las Compuestas en relación a otras familias. Para el género y especie *Helianthus annuus* existen 133.682 secuencias correspondientes a ESTs en NCBI. Existe una base de datos específica para la información genómica referida a las Compuestas en el sitio “Compositae Genome Project Database” (CGPDB) (http://cgpdb.ucdavis.edu), y se dispone también de otros sitios como el “Plant Genome Database” (http://www.plantgdb.org) con secuencias muy bien identificadas por especie y ensambladas en contigs y singletons.. Este trabajo tiene como objetivo central el diseño de una micromatriz de alta densidad para el girasol cultivado que represente todas las secuencias disponibles en GenBank/NCBI incluidas en las bases de datos de ESTs, con el propósito de ser utilizada en estudios de expresión génica para caracteres de importancia agronómica. Este trabajo contempló la limpieza de secuencias contaminantes presentes en los ESTs y secuencias depositadas en bases de datos públicas, el ensamblado de ellos en genes únicos (unigenes) genuinos y representativos (eliminación de redundancias y armado de contigs) y la anotación funcional de los mismos según un vocabulario controlado que pueda asociar la secuencia funcional del EST a una función hipotética determinada. El número de secuencias final ensambladas para incluir en la micromatriz será de 22.000 unigenes, que fueron previamente anotados utilizando Blast2Go [18] y posteriormente mapeados en vías metabólicas a través de KEGG.

Con respecto al análisis de expresión digital llevado a cabo también en este trabajo, los ESTs aislados de la clonoteca de embrión fueron los más diferencialmente expresados, principalmente con la clonoteca generada con embriones posteriores al cuarto día de auto-polinización y la clonoteca de cotiledón e hipocótilo. No se observaron perfiles de expresión diferencial para los ESTs provenientes de las clonotecas de embriones de cuatro días, protoplasto, hipocótilo de 1 a 5 días, junto con las aisladas en nuestro instituto [36] no se agruparon dentro de grupos con posible expresión diferencial.
Este análisis permite analizar de manera preliminar genes o grupos de genes de funcionalidad anónima. Si bien existe en progreso un chip de ADN de girasol que será

validado con diferentes experimentos de genómica funcional de manera de identificar genes candidatos asociados a diferentes eventos moleculares de interés agronómico, el análisis de expresión digital es una herramienta que da indicios de los posibles grupos asociados a expresión diferencial.

Datos de Contacto: *Dra. Paula Fernández, Instituto de Biotecnología, INTA Castelar, N. Repetto y Los Reseros S/N, (1686) Hurlingham, Pcia. de Buenos Aires, ARGENTINA*
E-mail: pfernandez@cnia.inta.gov.ar

Referencias

1. Buckingham S: **Programmed for success.** *Nature* 2003, **425**.
2. **National Center for Biotechnology Information (NCBI) (http://blast.ncbi.nlm.nih.gov/)**
3. Paniego N, Echaide M, Munoz M, Fernandez L, Torales S, Faccio P, Fuxan I, Carrera M, Zandomeni R, Suarez EY, Hopp HE: **Microsatellite isolation and characterization in sunflower (*Helianthus annuus L.*).** *Genome* 2002, **45:**34-43.
4. Fernandez P, Paniego N, Lew S, Hopp HE, Heinz RA: **Differential representation of sunflower ESTs in enriched organ-specific cDNA libraries in a small scale sequencing project.** *BMC Genomics* 2003, **4:**40.
5. Kiani P, Grieu P, Maury P, Hewezi T, Gentzbittel L, Sarrafi A: **Genetic variability for physiological traits under drought conditions and differential expression of water stress-associated genes in sunflower (*Helianthus annuus L.*).** *Theor Appl Genet* 2007, **114:**193-207.
6. Fusari C, Lia V, Hopp HE, Heinz RA, Paniego N: **Identification of Single Nucleotide Polymorphisms and analysis of Linkage Disequilibrium in sunflower elite inbred lines using the candidate gene approach.** *BMC Plant Biology* 2008, **8:7**.
7. Fusari C, Lia V, Nishinakamasu V, Zubrzycki JE, Puebla AF, Maligne AE, Hopp HE, Heinz RA, Paniego N: **Single Nucleotide Polymorphism Genotyping by Heteroduplex Analysis in Sunflower (*Helianthus annuus L.*).** *Molecular Breeding* 2010, **To be published.**
8. Maringolo C: **Mapeo de QTL asociados a resistencia a podredumbre húmeda del capítulo de girasol (*Sclerotinia sclerotiorum (Lib.) De Bary*). .** FCA Universidad Nacional de Mar del Plata, FCA Balcarce - Unidad Integrada INTA Balcarce; 2007.
9. Peluffo L, Lia V, Maringolo C, Troglia C, Hopp HE, Paniego N, Escande A, Paniego N, Fernie A, Heinz RA, Carrari F: **Metabolic shifts associated to *Sclerotinia sclerotiorum* infection in resistant and susceptible sunflower genotypes.** *Phytochemistry* 2010, **71:**70-80.
10. Fernandez P: **Análisis genómico de girasol: Desarrollo de colecciones de ESTs y de una plataforma bioinformática para estudios de expresión de genes candidatos en respuestas a estreses abióticos.** UBA, FCEyN; 2007.
11. Fernandez P, Di Rienzo J, Fernandez L, Hopp H, Paniego N, R.A. H: **Transcriptomic identification of candidate genes involved in sunflower responses to chilling and salt stresses based on cDNA microarray analysis.** *BMC Plant Biology* 2008, **8**.
12. Audic S, Claverie JM: **The Significance of Digital Gene Expression Profile.** *Genome Research* 1997, **7**.
13. Wang J, Liang P: **DigiNorthern, digital expression analysis of query genes based on ESTs.** *Bioinformatics* 2003, **19:**653-654.
14. Keilin T, Pang X, Venkateswari J, Halaly T, Crane O, Keren A, Ogrodovitch A, Ophir R, Volpin H, Galbraith D, Or E: **Digital expression profiling of a grape-bud EST collection leads to new insight into molecular events during grape-bud dormancy release.** *Plant Science* 2007, **173:**446-457.
15. Vizoso P, Meisel LA, Tittarelli A, Latorre M, Saba J, Caroca R, Maldonado J, Cambiazo V, Campos-Vargas R, Gonzalez M, et al: **Comparative EST transcript profiling of peach fruits under different post-harvest conditions reveals candidate genes associated with peach fruit quality.** *BMC Genomics* 2009, **10**.
16. Stekel D: **Image Processing.** In *Microarray Bioinformatics.* Edited by University PC. NY, USA; 2003

17. **The Dana Farber Cancer Institute Sunflower Gene Index Project (http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/gimain.pl?gudb=sunflower).**
18. Conesa A, Götz S, García-Gomez JM, Terol J, Talón M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.** *Bioinformatics* 2005, **21:**3674-3676.
19. Fernandez P, Blesa D, Príncipi D, Fusari C, Soria M, Reynares C, Angelone L, Delfino S, Conesa A, Tapia E, et al: **"Sunflower Functional Genome Database, a curated unigene database to support functional diversity studies en sunflower ".** In *LA-ISCB 2010*; *13-16 de marzo; Montevideo, Uruguay*. 2010
20. Huang X, Madan A: **CAP3: A DNA sequence assembly program.** *Genome Research* 1999, **9:**868-877.
21. Ashburner M, Ball CA, Blake JA, Botstein D, Buttler H, Cherry JM, Davis AP, Dolinsky K, Dwight SS, Eppig JT: **Gene Ontology tool for the unification of biology.** *Nature Genetics* 2000, **25:**25-29.
22. Seluja A, Farmer A, McLeod M, Harger G, Schad P: **Establishing a method of vector contamination identification in database sequences.** *Bioinformatics* 1999, **15**.
23. Korning PG, Hebsgaard SM, Rouze P, Brunak S: **Cleaning the GenBank *Arabidopsis thaliana* data set.** *Nucleic Acids Research* 1996, **24:**316-320.
24. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Ciao H, Merril CR, Wu A, Olde B, Moreno RF: **Complementary DNA sequencing: expressed sequence tags and human genome project.** *Science* 1991, **252:**1651-1656.
25. Band MR, Olmstead C, Everts RE, Liu ZL, Lewin HA: **A 3,800 gene microarray for cattle functional genomics: comparison of gene expression in spleen, placenta, and brain.** *Animal Biotechnology* 2002, **13**.
26. **Cross_match (http://www.phrap.org)**
27. Chou HH, Holmes MH: **DNA sequence quality trimming and vector removal.** *Bioinformatics* 2001, **17:**1093-1104.
28. **J. Craig Venter Institute (http://www.jcvi.org/).**
29. **Univec: (http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html)**
30. Chen YA, Lin C, Wan C, Wu HB, Hwang PI: **An optimized procedure greatly improves EST vector contamination removal.** *BMC Genomics* 2007, **8:416**.
31. Altschul S, Gish W, Miller W, Myers E, Lipman D: **Basic local alignment search tool.** *Journal of Molecular Biology* 1990, **215:**403-410.
32. Zeeberg BR, Feng W, Wang G, Wang MD, Fojo TA, Sunshine M, Narasinham S, Kane DW, Reinhold WC, Lababidi S, et al: **GoMiner: A Resource for Biological Interpretation of Genomic and Proteomic Data.** *Genome Biology* 2003, **4:**(28).
33. Al-Shahrour F, Díaz-Uriarte R, Dopazo J: **FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes.** *Bioinformatics* 2004, **20:**578-580.
34. Martin DM, Berriman M, Barton GJ: **GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes.** *BMC Bioinformatics* 2004, **18:**178.
35. Khan S, Situ G, Decker K, Schmidt CJ: **GoFigure: automated Gene Ontology annotation.** *Bioinformatics* 2003, **19:**2484-2485.
36. Fernandez P, Paniego N, Lew S, Hopp HE, Heinz RA: **Differential representation of sunflower ESTs in enriched organ-specific cDNA libraries in a small scale sequencing project.** *BMC Genomics* 2003, **4:**40.