FACULTAD DE CIENCIAS
UNIVERSIDAD DE CHILE

"DESARROLLO, IMPLEMENTACION Y VALIDACION DE NUEVA METODOLOGIA PARA EL ANALISIS PRECISO DE LA EXPRESION GENICA A GRAN ESCALA MEDIANTE SAGE"

RODRIGO FERNANDO MALIG FUENTES

Ingeniería en Biotecnología Molecular

Santiago de Chile, Septiembre de 2007

"DESARROLLO, IMPLEMENTACION Y VALIDACION DE NUEVA METODOLOGIA PARA EL ANALISIS PRECISO DE LA EXPRESION GENICA A GRAN ESCALA MEDIANTE SAGE"

Seminario de Título entregado a la Universidad de Chile en cumplimiento parcial de los requisitos para optar al Título de Ingeniero en Biotecnología Molecular.

RODRIGO FERNANDO MALIG FUENTES

DR. FRANCISCO MELO LEDERMANN
Director Seminario de Título

Comisión de Evaluación Seminario de Título

DR. JOSÉ ALBIRDUA SEPÚLVEDA
Presidente Comisión

DR. RICARDO CABRERA PAUCAR
Corrector

Santiago de Chile, Septiembre de 2007

FACULTAD DE CIENCIAS
UNIVERSIDAD DE CHILE

DESARROLLO, IMPLEMENTACION Y VALIDACION DE NUEVA METODOLOGIA PARA EL ANALISIS PRECISO DE LA EXPRESION GENICA A GRAN ESCALA MEDIANTE SAGE

Memoria de Título

Entregada a la

Universidad de Chile

en cumplimiento parcial de los requisitos

para optar al Titulo de

Ingeniero en Biotecnología Molecular

Por:
Rodrigo Fernando Malig Fuentes

Septiembre 2007
Santiago - Chile

Director de Memoria de Título: Dr. Francisco Melo Ledermann
Co-director de Memoria de Título: Eduardo Agosin Trumper

FACULTAD DE CIENCIAS
UNIVERSIDAD DE CHILE

INFORME DE APROBACION
MEMORIA DE TITULO

Se informa a la Escuela de Ciencias de la Facultad de Ciencias que la Memoria de Título presentada por el candidato.

Rodrigo Fernando Malig Fuentes

Ha sido aprobada por la Comisión de Evaluación de la Memoria de Título como requisito para optar al Título de Ingeniero en Biotecnología Molecular, en el examen de Defensa de la Memoria rendido el día , del mes de año

Director de Memoria de Título:

Dr..... .....

Comisión de Evaluación de la Memoria de Titulo

Dr..... .....

Dr..... .....

Dr..... .....

Dr..... .....

AGRADECIMIENTOS

Quiero agradecer a toda mi familia, en especial a mis padres, Fernando y Patricia, por la preocupación, dedicación y confianza depositada en mí. También quiero agradecer de forma muy especial a mi tutor Dr. Francisco Melo, gracias por la confianza, el apoyo y todas las enseñanzas, las cuales van más allá sólo de la parte académica. Gracias también a las personas que ayudaron a realizar este trabajo: Dr. Eduardo Agosin, Tomás Norambuena, Fernando Faunes, Francisco Pizarro, Hannetz Roschzttardtz, Dr. Juan Larrain y Gaëlle Lehouque.

INDICE

INDICE DE TABLAS

INDICE DE FIGURAS

Abreviaturas standard y convenciones

ADN	Acido desoxirribonucléico
cADN	Acido desoxirribonucléico complementario
ARN	Acido ribonucléico
UTR	Región no traducida
Nts	Nucleótidos

Técnicas

PCR	Reacción en cadena de la polimerasa
RT-PCR	Transcripción reversa de reacción en cadena de la polimerasa
SAGE	Análisis seriado de expresión génica

RESUMEN

Las actuales metodologías que analizan los datos experimentales obtenidos por análisis seriado de expresión génica (SAGE) producen un gran número de tags no identificados. Además, una fracción significativa de los tags de SAGE no es específica y tiene múltiples anotaciones, lo cual produce errores en la determinación de los niveles de expresión génica. Finalmente, la consideración de la información genómica en el proceso de mapeo de tags experimentales producidos por SAGE aumenta la posibilidad de descubrir nuevos genes y/o elementos regulatorios, los cuales pueden estar jugando un rol importante dentro de la célula bajo ciertas condiciones ambientales.

En esta tesis, se desarrolló un nuevo método para el proceso de mapeo de tags de SAGE llamado asignación génica jerárquica. Este método entrega una anotación completa de los potenciales tags de SAGE dentro de un genoma, junto con estimar una probabilidad de observación experimental para cada uno de ellos. Se aplicó esta nueva metodología al genoma de Saccharomyces cerevisiae, produciendo la anotación más completa y precisa de tags virtuales actualmente disponible para este organismo. Luego, se validó y se demostró la utilidad de este método con datos de SAGE experimentales. Se contribuye con nuevas pistas y nuevos resultados obtenidos desde datos ya publicados en experimentos de SAGE en levadura. Además, se presentan los beneficios de usar este nuevo método sobre genomas más grandes y complejos, donde esta metodología será de mayor utilidad.

En esta tesis también se implementó un servidor web para el mapeo de tags experimentales de SAGE. El núcleo de este servidor se basa en la nueva metodología descrita. Este servidor reduce en gran medida las asignaciones

ambiguas de los tags que tienen múltiples anotaciones en el genoma, además de facilitar el análisis de los datos producidos por SAGE.

Finalmente, usando datos de SAGE ya publicados y el nuevo servidor, se descubrieron nuevos genes dentro del genoma de Saccharomyces cerevisiae y se comenzó el proceso de validación experimental para cada uno de ellos.

ABSTRACT

The current methodologies to analyze the experimental data obtained with the technique of serial analysis of gene expression (SAGE) produce a large number of non-identified tags. In addition to this, the specificity or differential quality of a significant fraction of the experimentally observed SAGE tags cannot be assessed properly, which can lead to errors in the determination of gene expression levels. Finally, the consideration of genome information in the tag mapping process of SAGE experiments increases the possibility of discovering new genes and/or regulatory elements in the gene-to-protein production process that may be playing an important role in a living cell under certain environmental conditions.

In this thesis, a novel method for the tag-to-gene assignment process in SAGE, called hierarchical gene assignment or HGA, has been developed. The method provides a full annotation of the potential virtual SAGE tags within a genome, along with an estimation of their confidence for experimental observation. This method was applied to the Saccharomyces cerevisiae genome, producing the most thorough and accurate annotation of virtual SAGE tags that is available today for this organism. This method was then validated, and its usefulness demonstrated, based on known experimental data. New clues and results obtained from the analysis of existing SAGE experiments in yeast, which have not been dispatched yet, are also reported. The benefits of using the hierarchical gene assignment method on more large and complex genomes, where it will be most useful, are also highlighted.

In this thesis, a web server for the accurate mapping of experimental tags in SAGE was also implemented. The core of the server relies on a database of genomic virtual tags built by the new method developed here. The server attempts to reduce

the amount of ambiguous assignments for those tags that are not unique in the genome, along with facilitating the analysis of data from SAGE experiments.

Finally, using published experimental SAGE data and the server implemented here, new genes are discovered in the Saccharomyces cerevisiae genome and the experimental validation process for each of them was started.

1. INTRODUCCION

1.1.- La fermentación alcohólica

Durante el proceso de vinificación las levaduras están sujetas a múltiples y cambiantes condiciones de estrés, las cuales incluyen: shock hiperosmótico, limitación de nutrientes, variaciones de temperatura y toxicidad por altas concentraciones de etanol.

Para sobreponerse a estas condiciones extremas, en las levaduras productoras de vino han evolucionado respuestas metabólicas que capacitan a la célula a adaptarse a sus nuevas condiciones ambientales. Estas respuestas les permiten censar y adaptarse adecuada y oportunamente a los cambios en su ambiente, permitiéndoles mantener su actividad metabólica e integridad celular (Bauer & Pretorius, 2000). Una adaptación frente a este entorno desfavorable implica una reorganización metabólica para mantener la actividad celular. A veces, esta reorganización conduce a una reducción de la actividad celular provocando una disminución de la tasa fermentativa generando fermentaciones deficientes, lo que finalmente resulta en fermentaciones "estancadas" (Walter, 1998). Estas adaptaciones también involucran cambios en los niveles de expresión génica, donde un gran número de genes varían su expresión significativamente durante todo el proceso de fermentación alcohólica (Rossignol y col, 2003). De esta forma, obtener los perfiles detallados de expresión génica de la levadura productora de vino durante la fermentación alcohólica permitiría el entendimiento, a nivel molecular, de los procesos biológicos de la fermentación del vino y de la regulación de la expresión génica en respuesta a los cambios ambientales. Entendimiento que finalmente podría servir para optimizar el proceso de vinificación.

1.2.- Métodos para medir el nivel de expresión génica en transcriptomas completos

Actualmente los métodos más comúnmente utilizados para medir la expresión de miles de genes a nivel celular son: 1) Microarreglos de ADN (Schena y col, 1995) y 2) Análisis seriado de expresión génica (SAGE) (Velculescu y col, 1995). Ambos métodos tienen ventajas y desventajas. La mayor ventaja de los microarreglos de ADN es el tiempo y costo requerido para obtener datos experimentales, cuando el chip de ADN se encuentra disponible. Entre las mayores desventajas de este método están: 1) Su baja reproducibilidad entre experimentos independientes, 2) Genes para los cuales la sonda no está disponible, no pueden ser medidos, 3) Los datos de expresión son obtenidos de tal forma que este método mide la expresión de genes de forma cualitativa, y no cuantitativa, 4) Baja sensibilidad de detección (transcritos poco abundantes son difíciles de cuantificar) y 5) Baja especificidad como consecuencia de hibridaciones cruzadas, las cuales pueden provocar la obtención de un perfil de expresión génico incorrecto. Por otro lado, los beneficios de SAGE es que es una técnica cuantitativa, genes no caracterizados pueden ser medidos, facilitando el descubrimiento de nuevos genes y los transcritos que son expresados a un bajo nivel pueden ser detectados. La mayor desventaja de la técnica de SAGE es el tiempo y costos requeridos para su realización.

1.3.- Análisis seriado de expresión génica (SAGE) en levadura durante la fermentación alcohólica

La tecnología de SAGE ha sido pionera en el uso de secuencias nucleotídicas cortas, denominadas tags, las cuales son extraídas desde ARN poliadenilado. Las

librerías de tags permiten inferir los niveles de expresión génica a gran escala (Velculescu y col, 1995), ya que la frecuencia de los tags obtenidos por SAGE es directamente proporcional a la abundancia del transcrito desde donde estos provienen.

SAGE ha sido descrita como una poderosa técnica para analizar transcriptomas completos y puede ser muy eficiente para el descubrimiento de nuevos genes y la anotación de genomas (Caron y col, 2001; Boheler & Stern 2003; Sun y col, 2004; Tuteja & Tuteja 2004a; Tuteja & Tuteja 2004b; Harbers & Carninci 2005). Su éxito se basa en la habilidad para medir valores absolutos de expresión génica sin la necesidad del uso de sondas, previamente conocidas.

Sin considerar los costos de la recolección de datos, los antecedentes previamente descritos posicionan a SAGE como la mejor técnica, actualmente disponible, para medir la expresión génica a gran escala, razón por la cual se ocupó esta técnica para cuantificar los perfiles de expresión génica bajo condiciones de fermentación vínica en la levadura de vino *Saccharomyces cerevisiae* EC1118 (Varela y col, 2005). Este estudio permitió analizar la expresión diferencial de genes relacionados con respuesta a estrés, entre otros, y permitió el descubrimiento de nuevos genes posiblemente involucrados en los cambios metabólicos que ocurren durante la fermentación alcohólica en la levadura. Esto debido a que se observó un gran número de tags que provienen de regiones intergénicas del genoma de la levadura secuenciada (tags llamados NORFs, non-open reading frame), e incluso, 324 tags (20% de todos los tags producidos) ni siquiera mapearon dentro del genoma de la cepa de *Saccharomyces cerevisiae* secuenciada. Estos últimos se denominan tags NIDs (non-identified tags).

1.4.- Metodología y desventajas de las actuales asignaciones de tags de SAGE

1.4.1.- Protocolo de la extracción de tags en SAGE

La metodología de SAGE comienza con la digestión del cADN, generado a partir de una muestra biológica, con una enzima llamada "enzima ancla" (comúnmente NlaIII cuya secuencia de reconocimiento es CATG) para posteriormente ligar un adaptador en el extremo cohesivo que se generó como consecuencia de la digestión. Este adaptador contiene un sitio de reconocimiento para una endonucleasa de clase II, enzima llamada "enzima de etiquetado" y una secuencia a la que posteriormente se unirá un partidor durante los pasos de PCR. Luego una endonucleasa de clase II, por ejemplo BsmFI, corta a 14 pares de bases de distancia río abajo de su sitio de reconocimiento, produciendo fragmentos de ADN con extremos cohesivos, los cuales permiten ligar los productos de la digestión, produciendo ditags. Estos ditags se amplifican por PCR, luego se digieren con la enzima ancla – para liberar los adaptadores – se concatenan, se clonan y se secuencian. Desde aquí en adelante, mediante herramientas bioinformáticas, se extraen secuencias de 14 nucleótidos (secuencias denominadas tags), se determina la frecuencia de los tags y se determinan los transcritos desde donde estos provienen (Figura 1).

1.4.2.- Mapeo de tags producidos por SAGE y desventajas de los métodos actuales que realizan esta tarea

Un paso crítico dentro de la metodología de SAGE es el proceso de asignar cada tag al transcrito desde donde fue generado, proceso llamado mapeo de tags.

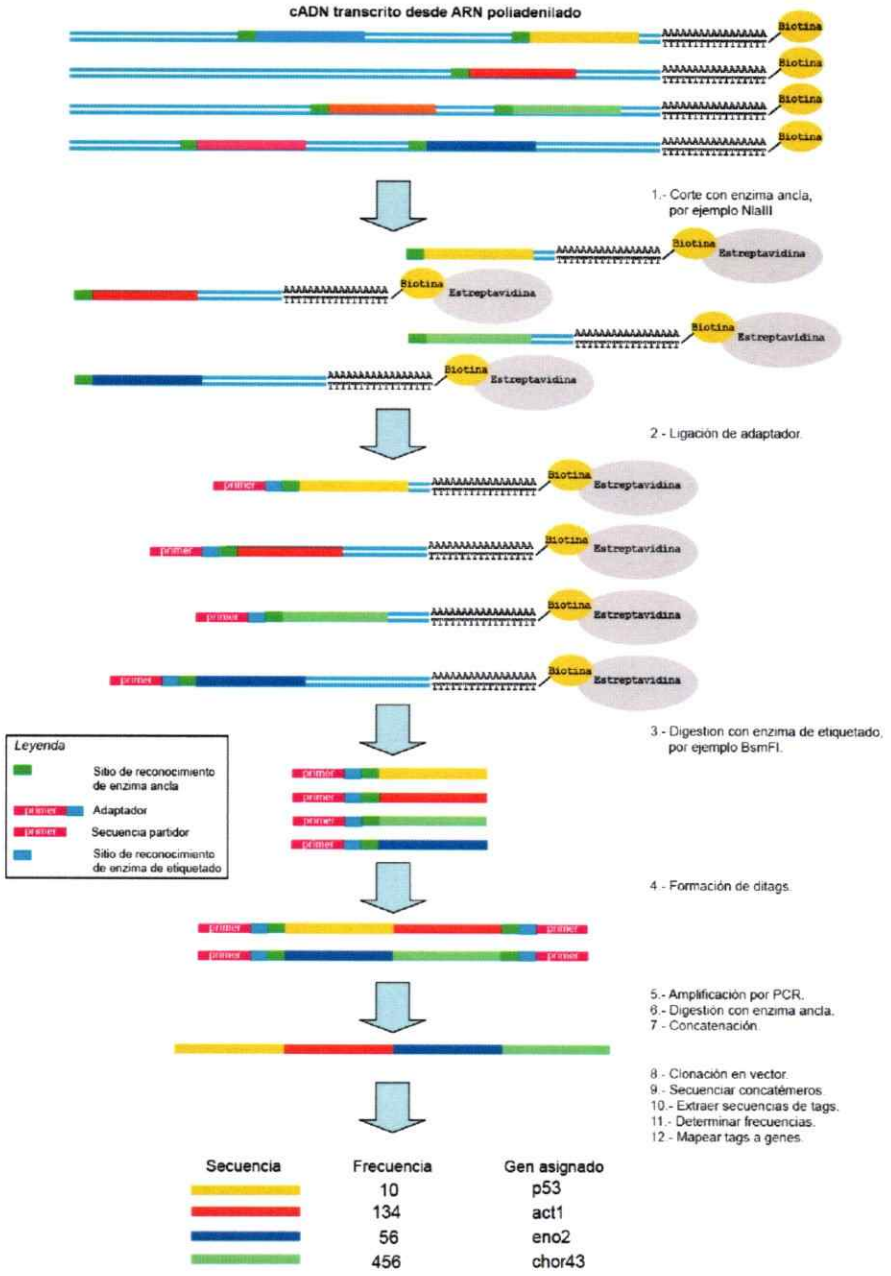


Figura 1. Metodología típica para la preparación de una librería de tags en SAGE. Una población de transcritos poliadenilados son extraídos desde un tejido o tipo celular. Este RNA es transcrito reversamente para producir cADN de doble hebra. El cADN es digerido con una enzima ancla, generalmente NlaIII, y los fragmentos más 3' son capturados mediante el uso de biotina ligada a el partidor oligo dT utilizado durante la síntesis de cADN (1). A los extremos cohesivos creados por la digestión con la enzima ancla es unido un adaptador, el cual introduce un sitio de unión de un partidor y un sitio de reconocimiento de una enzima etiquetadora, adyacente al fragmento de cADN (2). Fragmentos que contienen la secuencia del adaptador y una corta secuencia propia del cADN (secuencia llamada tag) son liberados luego de la digestión con la enzima etiquetadora (3), los cuales son ligados entre ellos para formar ditags (4). La secuencia del adaptador a ambos lados de los ditags es utilizada para amplificar por PCR (5). Debido a una re-digestión de los productos de PCR con la enzima ancla, todos los adaptadores son removidos de los ditags (6) antes de que sean ligados en grandes concatémeros (7), los cuales son clonados en un vector (8). Los vectores obtenidos son secuenciados (9), luego mediante herramientas bioinformáticas los tags de 14 nucleótidos son extraídos (10), sus frecuencias determinadas (11) y finalmente se establecen los genes desde donde provienen los tags (12).

Actualmente, el mapeo de tags involucra la búsqueda de los tags observados experimentalmente dentro del transcriptoma del organismo desde donde se extrajo la librería de tags. Esta estrategia permite solamente asignar parcialmente los tags, debido a que las fuentes actuales con datos de transcriptomas están incompletas para la mayoría de las especies y organismos. Por lo tanto, una fracción significativa de los tags obtenidos experimentalmente permanece no asignada. Estos tags son llamados o categorizados como NIDs (non-identified). Las bases de datos comúnmente utilizadas para mapear tags de SAGE (van Kampen y col, 2000; Lash y col, 2000; Divina & Jiri, 2004) utilizan clusters UniGene (Schuler, 1997) de ESTs (Expressed Sequence Tags) para asignar los tags experimentales de SAGE al tag potencial más 3' de los transcritos conocidos. Cada cluster UniGene contiene una colección de secuencias expresadas (mRNA bien caracterizados, secuencias de cDNA y ESTs), las cuales pueden estar representando a un único transcrito. Sin embargo, esta estrategia tiene varias desventajas cuando se emplea para mapear tags de SAGE. Primero, un mismo gen puede estar representado en múltiples clusters resultando en una asignación ambigua de los tags. Segundo, los ESTs, el mayor componente de los clusters UniGene (Lash y col, 2000), tienen una tasa de error estimada aproximadamente de un 1% (1 en 100 nts), resultando en una tasa errónea de asignación de tags cercana al 10% (Lash y col, 2000). Tercero, los clusters UniGene no contienen la totalidad de los transcritos de un organismo, por lo que existirán tags que no serán asignados (como en el caso de los genes hipotéticos). Por ejemplo, estudios de SAGE en humanos mostraron que el 60% de los tags experimentales no fueron asignados a ningún cluster UniGene (Chen y col, 2002). La correspondencia entre los tags no asignados y los transcritos desde donde provenían fue demostrada por RT-PCR, donde sobre el 90% de todos los tags no asignados provenían de un transcrito verdadero (Chen y col, 2002). Cuarto, en los clusters UniGene están representados principalmente los genes más

abundantes, por lo que detectar genes poco abundantes (por ejemplo, factores de transcripción) mediante SAGE usando estas bases de datos es poco probable. Quinto, mapear tags de SAGE en estas bases de datos no permite el descubrimiento de genes nuevos, la cual es una de las características más importantes que posee SAGE.

1.4.3.- Ventajas de la utilización de la información genómica para la asignación de tags de SAGE

Como se mencionó en el punto 1.3, SAGE puede ser muy eficiente para el descubrimiento de nuevos genes y anotación de genomas (Boheler & Stern 2003; Harbers & Carninci 2005; Sun y col, 2004). Por esta razón, la utilización de información genómica en lugar de información del transcriptoma, debe ser usada en el proceso de mapeo de tags experimentales de SAGE. Esta estrategia soluciona el problema de estar limitado sólo a los genes que poseen un EST descrito. Junto con lo anterior, las secuencias genómicas tienen una baja tasa de error en sus secuencias, menos de 0.0001% (Adams y col, 2000) y la cantidad de genes anotados es significativamente mayor que el conjunto de secuencias expresadas para un organismo. De este modo, la información genómica constituye una mejor fuente de información para el mapeo de tags de SAGE, permitiendo además el descubrimiento de nuevos genes de una manera más eficiente. Sin embargo, el uso de genomas completos para mapear tags representa un desafío bioinformático debido a que la complejidad de los grandes genomas hace más improbable la existencia de tags únicos, lo que provoca una gran ambigüedad en el mapeo de cada una de estas secuencias (Wahl y col, 2004).

1.5.- Hipótesis

Considerando todos los antecedentes mencionados, las hipótesis de este trabajo son las siguientes:

1) La utilización de información genómica resultará en una asignación de tags de SAGE más completa y confiable en comparación con los métodos existentes.

2) El mapeo de tags de SAGE contra el genoma permitirá el descubrimiento de nuevos genes en levadura.

1.6.- Objetivo general

Desarrollar un nuevo método bioinformático que permita utilizar los beneficios y solucionar las dificultades del uso de información genómica para el proceso de mapeo de tags de SAGE.

1.7.- Objetivos específicos

1.- Aumentar la precisión de la asignación de tags experimentales de SAGE.

2.- Disminuir la cantidad de tags experimentales de SAGE no asignados.

3.- Disminuir la ambigüedad en las asignaciones de tags experimentales de SAGE.

4.- Desarrollar un servidor que utilice las nuevas herramientas desarrolladas en este trabajo.

5.- Descubrir nuevos genes en levadura que pudieran ser importantes para el proceso de fermentación vínica.

6.- Aumentar el entendimiento de los perfiles de expresión génica durante fermentaciones alcohólicas en levadura.

2. RESULTADOS

2.1.- Mapeo de tags por Asignación Génica Jerárquica (HGA)

En este trabajo se creó un nuevo método para llevar a cabo el proceso de mapeo de tags producidos por SAGE. Este método combina información genómica y su actual anotación junto con datos experimentales de SAGE anteriores, de modo de aumentar la precisión y reducir la ambigüedad de asignación de tags. Este nuevo método es llamado mapeo de tags por asignación génica Jerárquica (Hierarchical Gene Assignment, HGA) y consiste de 4 etapas principales, las cuales son descritas en detalle a continuación. Un diagrama de flujo de cada uno de los pasos involucrados en el HGA es ilustrado en las Figuras 2.1-2.2.

2.1.1.- Paso 1: Extracción y anotación de tags potenciales de SAGE en el genoma

En el genoma completo de un organismo se buscan los sitios de reconocimiento de la enzima ancla utilizada en SAGE. Luego todos los tags de SAGE potenciales son extraídos en combinación con la enzima etiquetadora de elección. Los tags extraídos de este modo son comparados en pares todos contra todos y la frecuencia de ocurrencia en el genoma para cada uno de ellos es determinada (Figura 2.1A, derecha). Por otro lado, las tablas de transcritos, las cuales contienen las anotaciones de las posiciones de los genes de cada organismo son ocupadas por el método HGA para mapear todos los genes en las diferentes posiciones del genoma bajo estudio (Figura 2.1A, izquierda).

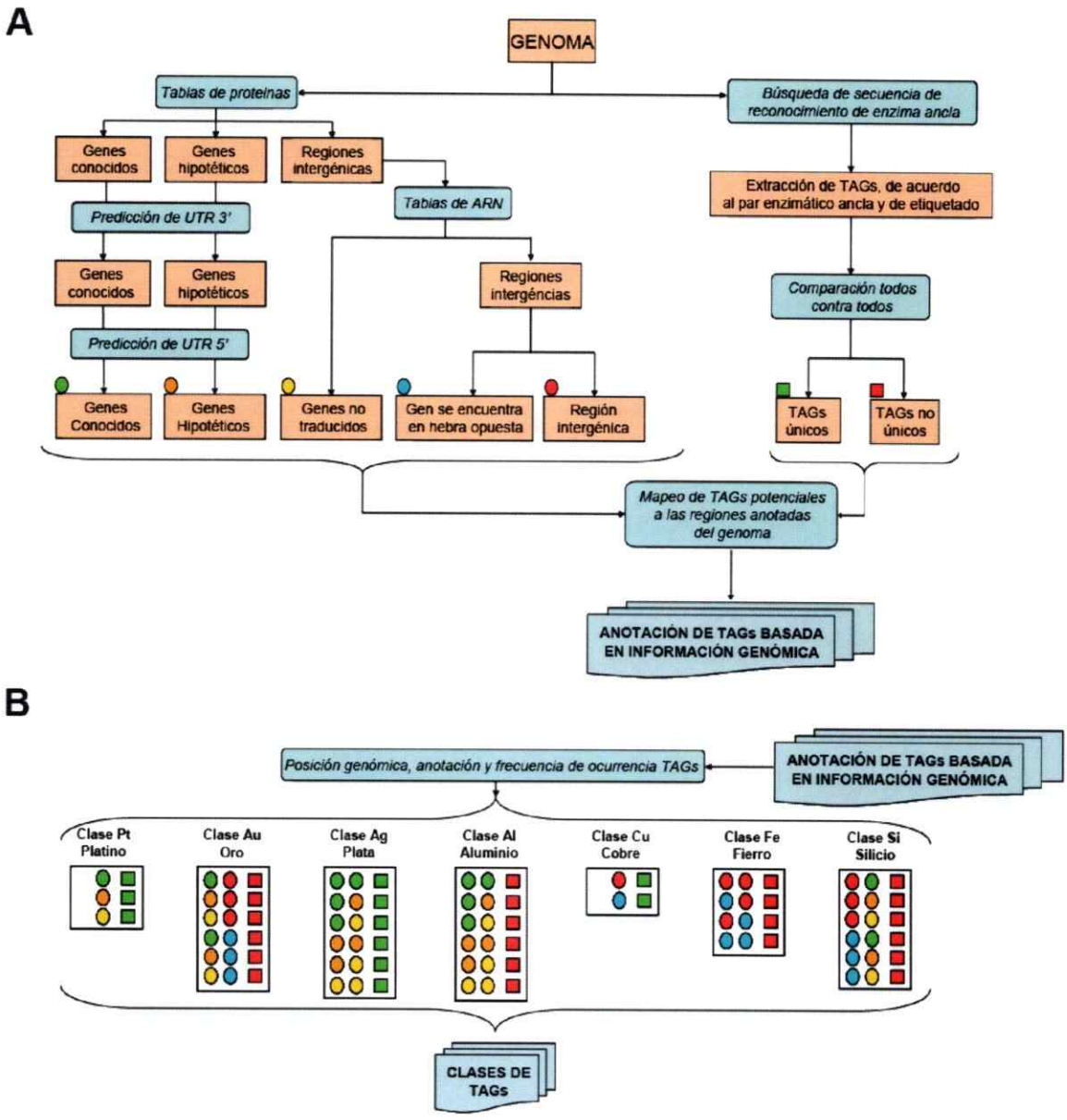


Figura 2.1 Primera parte del diagrama de flujo del método HGA. El método consiste principalmente de 4 pasos secuenciales. **A**.- Paso 1) Primero, todos los tags potenciales del genoma son extraídos y comparados, mientras que la frecuencia de cada tag es registrada, junto con su posición dentro del genoma. Luego utilizando las tablas de proteínas y ARNs para algún genoma en particular, junto con la asignación de los UTRs para los genes, todos los transcritos potenciales y las regiones intergénicas son extraídas y sus posiciones son almacenadas. Toda la información obtenida en este paso es cruzada produciendo una detallada anotación de los tags de SAGE virtuales basada en el genoma. **B**.- Paso 2) Basándose en la posición genómica, anotación y frecuencia de los tags potenciales dentro del genoma, cada tag virtual es asignado a una de siete clases posibles. Este nuevo esquema de clasificación ayuda a asignar una confiabilidad a los tags en las etapas posteriores. Una explicación detallada de cada una de las clases de tag está disponible en la Tabla 1.

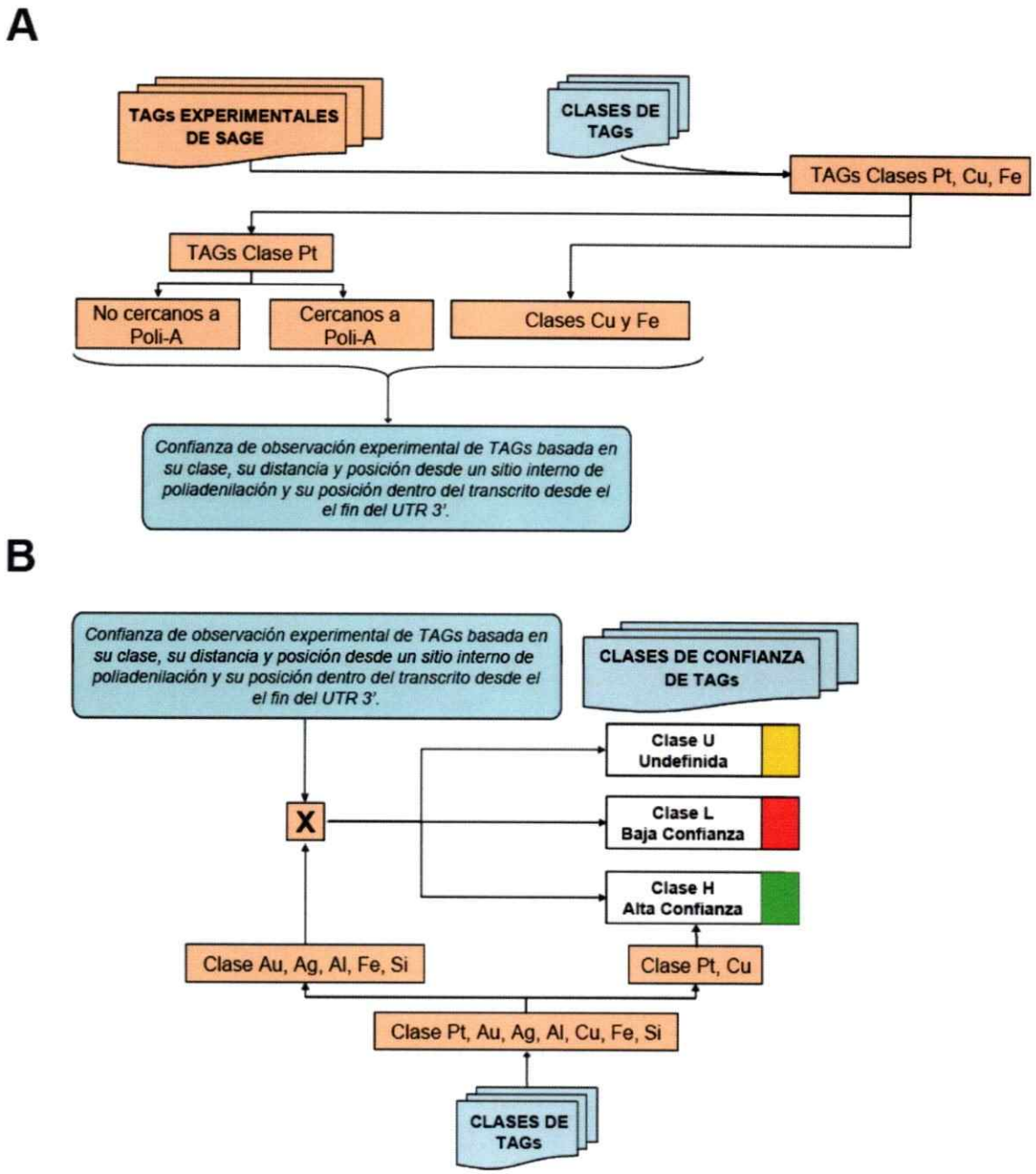


Figura 2.2 Segunda parte del diagrama de flujo del método HGA. A.- Paso 3) Todos los tags de SAGE experimentales (Tabla 2) son cruzados contra la clasificación de tags virtuales previamente descrita, y sólo los tags experimentales pertenecientes a las clases Platino, Cobre y Hierro son seleccionados. El conjunto de tags perteneciente a la clase Platino es subdividido en 2 grupos: i) Aquellos tags que mapean en un transcrito y no están localizados río arriba de una región de poliadeninas internas y ii) Aquellos tags que mapean en un transcrito y que están cercanos a un región de poliadeninas internas. La anotación genómica y clasificación de cada tag es utilizada para determinar su probabilidad de ser observada experimentalmente. Una detallada descripción de las funciones de probabilidad que son derivadas de los datos experimentales es mostrada en la figura 3. **B**.- Paso 4) La clasificación generada en el paso 2 es cruzada contra las probabilidades obtenidas en el paso 3 para producir una asignación de confianza (alta, baja o indefinida) para cada una de las tags potenciales dentro del genoma. Esta información puede ser utilizada para mapear sin ambigüedad tags de SAGE experimentales a los transcritos o regiones genómicas desde donde provienen, junto con una estimación de la confianza de la asignación del tag. Una explicación detallada de los

diferentes pasos utilizados durante el proceso del HGA es proporcionada en las secciones resultados y métodos.

En el caso que las tablas de proteínas de un determinado organismo sólo especifiquen las regiones codificantes de cada gen y no tengan asignadas las regiones no traducidas (UTRs) en los extremos 5' y 3' de cada transcrito, es necesario asignar estas regiones. Una asignación precisa de estas regiones es particularmente relevante para el caso de los UTRs 3', debido a que se espera que una fracción significativa de los tags producidos por SAGE se produzcan de estas regiones. El grado de precisión con que se pueden asignar los UTRs, depende del conocimiento del transcriptoma del organismo bajo estudio. Entonces, luego de asignar los UTRs de todos los genes, las regiones que aún permanecen como intergénicas son categorizadas en dos clases, dependiendo si en la hebra opuesta a esta región existe o no un transcrito anotado (Figura 2.1A, medio). Una vez que el HGA ha completado los procesos anteriores, la información genómica anotada es clasificada dentro de alguna de las siguientes categorías: 1) Genes conocidos, 2) genes hipotéticos, 3) Genes no traducidos, 4) Región intergénica donde un gen se encuentra en la hebra opuesta y 5) Región intergénica en ambas hebras.

2.1.2.- Paso 2: Definición de clases de tags y sus características

La información estructurada del genoma generada en el paso 1 es cruzada contra todos los tags virtuales, generando una anotación de todos los tags potenciales en el genoma. Los tags resultantes son categorizados en alguna de 7 clases, las cuales dependen de la posición genómica, anotación y frecuencia de ocurrencia de cada tag potencial en el genoma (Figura 2.1B). La definición detallada de cada clase de tag es proporcionada en la Tabla 1. Esta nueva propuesta de categorización facilita la deducción no sólo del grado de confianza de la asignación,

sino que también permite un manejo más ordenado y fácil de la información que produce SAGE.

Tabla 1: Definición de clases de tags virtuales genómicos producidos por SAGE.

CLASE DE TAG		
ID	**NOMBRE**	**DESCRIPCION**
Pt	**Platino**	Tag es único en el genoma y mapea un transcrito
Au	**Oro**	Tag no es único en el genoma, mapea un gen mientras todas sus anotaciones restantes son intergénicas.
Ag	**Plata**	Tag es único en el genoma pero mapea dos o más genes sobrelapados localizados en la misma región genómica.
Al	**Aluminio**	Tag no es único en el genoma, mapea un transcrito pero sus otras ocurrencias también mapean otro transcrito localizado en una región genómica distinta.
Cu	**Cobre**	Tag es único en el genoma y mapea con una región intergénica.
Fe	**Fierro**	Tag no es único en el genoma, pero todas sus ocurrencias mapean un región anotada como intergénica.
Si	**Silicio**	Tag no es único en el genoma, mapea un región intergénica, pero otras ocurrencias de esa secuencia mapean un transcrito.

Como un importante complemento de este nuevo esquema de clasificación, el método HGA también incorpora dos características adicionales de los tags, las cuales tienen el objetivo de reducir alguna de las distorsiones posibles que pueden afectar la interpretación de los resultados de SAGE. Primero, todas las subsecuencias de 8 o más adeninas consecutivas dentro de un gen anotado son identificadas de modo de considerar el posible apareamiento del oligo dT con regiones de poliadeninas internas de moléculas de ARNs, apareamiento que puede ocurrir durante el primer paso del protocolo de SAGE, la transcripción reversa (Figura 1). Ha sido demostrado que este proceso ocurre con una alta frecuencia, causando que cerca del 12% de los ESTs estén truncados debido al apareamiento del oligo dT en regiones de poli(A) internas (Nam y col, 2002). Así, aquellos tags que mapeen dentro de un gen y estén situados cerca y río arriba de un sitio de poliadenilación interno son clasificados por el método HGA como tags 'cercanos a poli(A)' (Para detalles ver materiales y métodos). Los tags que no pertenecen a la categoría anterior serán clasificados como tags 'no cercanos a poli(A)'. Segundo, el efecto de splicing y su potencial impacto sobre la generación o modificación de secuencias de tags es también considerada por el método HGA. Los tags que mapean en un gen en el límite exón-intrón son etiquetados como 'tags potenciales de splicing'. En este caso, un splicing virtual es generado en el computador y la nueva secuencia es registrada como 'tag splicing'. Cada 'tag splicing' es clasificada nuevamente siguiendo los mismos criterios mencionados anteriormente. En los casos donde se produce un nuevo sitio de reconocimiento de la enzima ancla como consecuencia del splicing, la nueva secuencia del tag es generada y registrada. Estos tags son clasificados como 'nuevos tags potenciales' y su correspondiente clase es calculada *de novo*. Los tags restantes son clasificados como 'tags no splicing'.

2.1.3.- Paso 3: Determinación de los valores de probabilidad para la observación de tags a partir de datos experimentales

La clasificación de tags resultante, junto con las características adicionales mencionadas anteriormente, son utilizadas para seleccionar tags particulares desde el genoma. Luego, estos tags seleccionados son examinados por su ocurrencia en datos de SAGE experimentales y descritos en la literatura para algún organismo de interés. Estos tags permitirán obtener probabilidades de observación experimental, probabilidades que posteriormente serán heredadas por todos los tags virtuales (Figura 2.2A). Los tags seleccionados deben pertenecer a tres clases diferentes: platino (Pt), cobre (Cu) y fierro (Fe) (ver Tabla 1 para detalles). Sólo los tags pertenecientes a estas clases son escogidos debido a que estos pueden ser asignados sin ambigüedad a un único transcrito o alguna región intergénica en el genoma. De este modo, la probabilidad de que un tag potencial con características especiales sea observado experimentalmente puede ser obtenida. Para aquellos tags que mapean en un transcrito, son únicos en el genoma, tienen una única anotación (tags clase platino) y además son clasificados como 'no cercanos a un poli(A)', la probabilidad de observarlos experimentalmente en función de su posición de mapeo dentro del transcrito (desde el fin del UTR3') puede ser calculada. La posición donde mapea el tag dentro del transcrito se refiere a la posición que posee el tag potencial que es observado experimentalmente, en donde cada tag potencial tiene una posición dentro de un transcrito, y siendo el tag potencial más 3' la posición número 1 (Figura 3A). En este caso, se obtuvieron diferentes valores de probabilidad dependiendo de la posición dentro del transcrito donde el tag mapeó. De este modo el HGA incorpora el efecto de digestiones incompletas por la enzima ancla en experimentos de SAGE (la Figura 3B muestra la función derivada para levadura). Por otro lado, para los tags Pt clasificados como 'cercanos a poli(A)', un

sólo valor de probabilidad es derivado desde los datos experimentales. Este valor representa la probabilidad de obtener un tag experimental como consecuencia del apareamiento del oligo dT con regiones de poli(A) internas durante la síntesis del cADN (La Figura 3C exhibe el valor calculado para levadura). Finalmente, los tags pertenecientes a las clases cobre (Cu) y fierro (Fe) corresponden a los tags que mapean a una región intergénica del genoma, en donde la clase Cu representa a los tags con secuencias únicas en el genoma, mientras que la clase Fe agrupa a los tags con múltiples instancias en el genoma. En ambos casos, la frecuencia de ocurrencia de estas clases de tags en datos experimentales, representan la probabilidad de que un tag provenga de una región intergénica, de acuerdo a la actual anotación de un determinado organismo (La Figura 3C presenta el valor calculado para levadura). Entones para resumir, a todos los tags potenciales se les asigna un valor de probabilidad de acuerdo a las regiones genómicas donde mapean y a las características específicas que ellos tienen. Por ejemplo, los tags que son asignados a transcritos tendrán distintos valores de probabilidad de acuerdo a la posición donde mapean dentro de estos y a la proximidad de poli(A) internos, mientras que los tags que son asignados a regiones intergénicas tendrán un sólo valor de probabilidad.

2.1.4.- Paso 4: Razones de probabilidad para asignar un grado de confianza a los tags virtuales de SAGE

Las funciones de probabilidad descritas en el paso anterior son luego cruzadas contra todos los tags potenciales, lo que finalmente permitirá obtener un grado de confianza de observar experimentalmente cada tag potencial del genoma (Figura 2.2B).

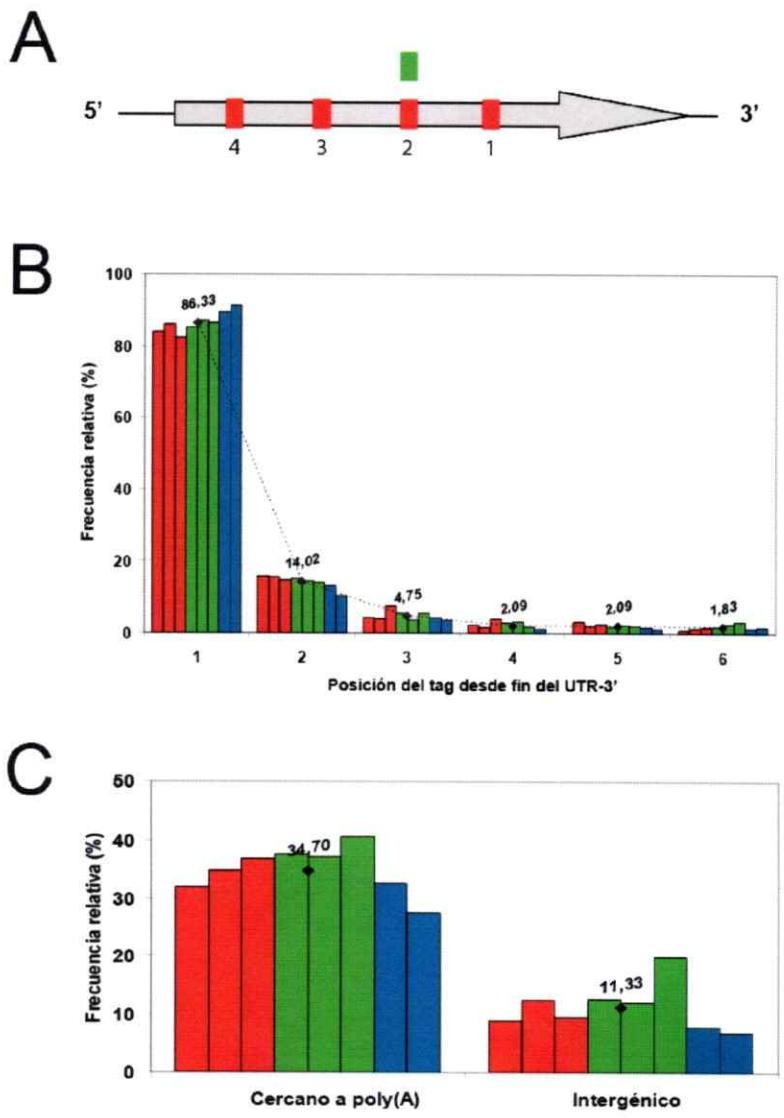


Figura 3. Frecuencia experimental de observación de tags producidos por SAGE en levadura. Las probabilidades de observar un tag en particular en un experimento de SAGE fueron derivadas desde datos de SAGE experimentales de 8 condiciones de cultivo distintas (Tabla 2). Las barras rojas corresponden a los datos Var-1, Var-2 y Var-3 obtenidos por Varela y colaboradores (Varela y col, 2005); las barras verdes corresponden a los datos Vel-1, Vel-2 y Vel-3 obtenidos por Velculescu y colaboradores (Velculescu y col, 1997) y las barras azules corresponden a los datos Kal-1 y Kal-2 obtenidos por Kal y colaboradores (Kal y col, 1999). Los valores promedio de los datos de los ocho puntos están mostrados sobre un punto negro. **A.-** Ejemplo de posición donde mapea un tag dentro de un transcrito. Se representan los tags potenciales en color rojo, un tag experimental en color verde y las posiciones de los tags potenciales con su número correspondiente. En este caso el tag experimental mapea en la posición dos del transcrito debido a que esa posición tiene asignado el tag potencial que tiene la misma secuencia del tag experimental. **B.-** Este panel muestra los valores de las probabilidades de observar un tag platino – 'no cercano a poli(A)' – de acuerdo a la posición dentro del transcrito desde el fin del UTR-3'. Los tags experimentales seleccionados fueron aquellos que fueron únicos en el genoma, tuvieron una única anotación, mapearon a transcritos con regiones no traducidas (UTRs) 3' predichas y categorizados como 'no cercanos a poli(A)'. Tags platino que mapearon en transcritos sin un UTR-3' predicho no fueron incluidos en este estudio para evitar errores en la estimación de la posición de los tags. **C.-** Este panel muestra el valor de la probabilidad de observar un tag platino clasificado como 'cercano a poli(A)'. Los tags experimentales seleccionados para calcular este valor fueron tags platino categorizados como tag 'cercano a poli(A)'. La probabilidad de observar un tag que mapee en una región intergénica en el genoma de levadura fue derivada desde tags experimentales pertenecientes a las clases cobre y fierro (Tabla 1).

Estos valores de confianza expresan la probabilidad de asignar apropiadamente un tag experimental a un tag virtual genómico y es representado por el HGA en tres posibles categorías: 1) Confianza alta, 2) Confianza baja o 3) Confianza Indefinida. La clase confianza alta significa que un tag tiene una alta probabilidad de ser correctamente asignado. Todos los tags que son únicos en el genoma y que tienen una única anotación son clasificados dentro de esta categoría. En aquellos casos donde la secuencia de un tag ocurre dos o más veces en el genoma, las razones de las probabilidades entre todas las instancias de ese tag son calculadas. Una confianza alta ha sido arbitrariamente definida como por lo menos 5 veces más probable que cada una de las otras asignaciones alternativas. El umbral de razón de probabilidades igual a 5 es un parámetro del programa, por lo que puede ser fácilmente modificado. La clase confianza baja es la opuesta a la confianza alta, lo que significa que hay otra alternativa para la misma secuencia del tag que es por lo menos 5 veces más probable. La clase confianza indefinida es asignada a aquellos casos donde no hay una sola anotación genómica de la misma secuencia del tag que pueda ser asignada con confianza alta (por ejemplo, entre todas las ocurrencias de un tag que es observado múltiples veces en el genoma, no hay ni un caso en que una probabilidad de ocurrencia sea por lo menos 5 veces superior cuando se compara con todas las otras instancias). En estos casos, las distintas instancias de un tag aún pueden ser ordenadas de acuerdo a los valores de las razones de las probabilidades que ellas exhiben, opción que posee el método HGA. Algunos ejemplos ilustrando como el método HGA asigna la clase de confianza a los tags, son mostrados en la Figura 4.

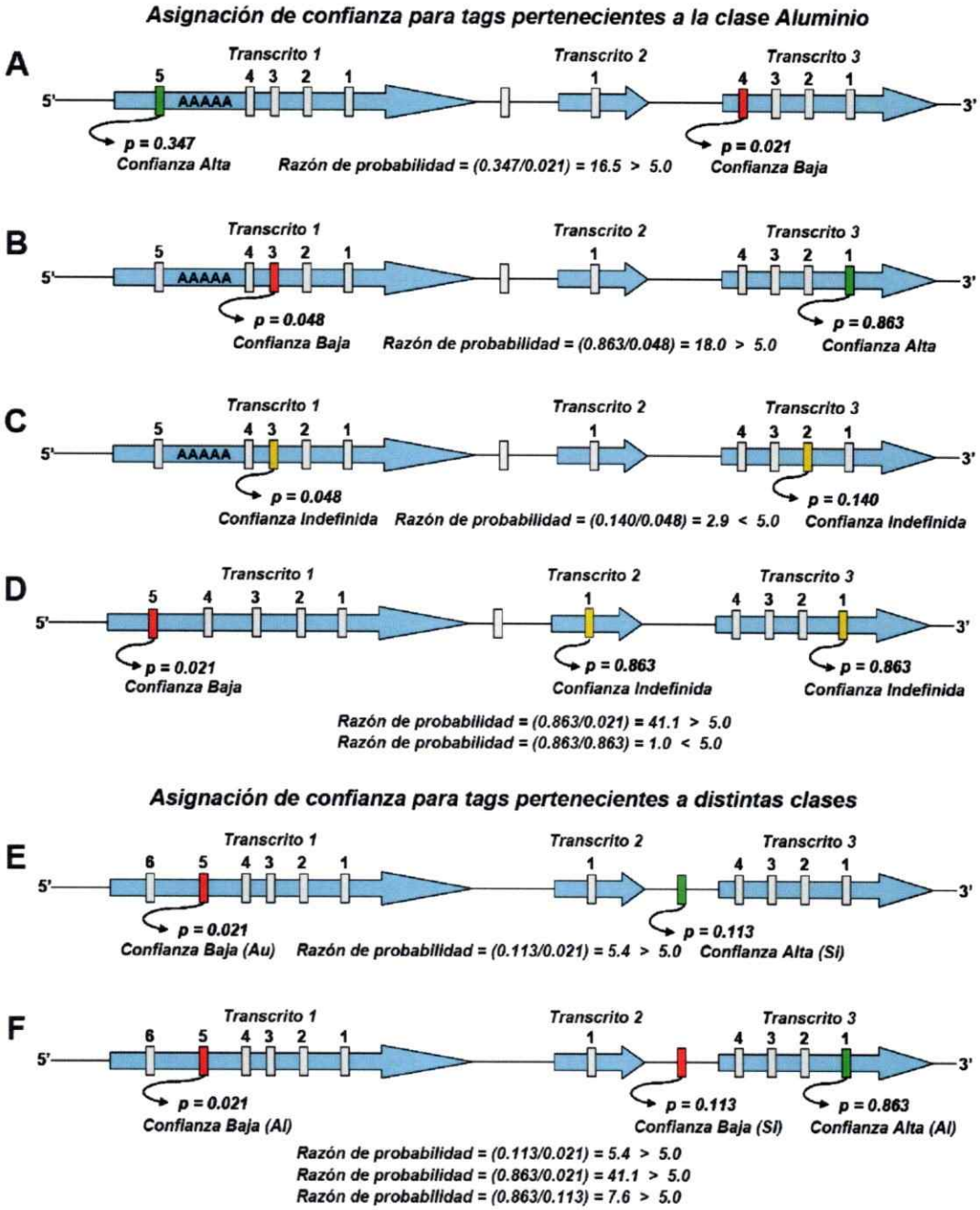


Figura 4: Ejemplos de asignación de confianza a tags de SAGE por el método HGA. Se muestran distintos escenarios de asignación de clases de confianza a tags virtuales genómicos. Desde los paneles A hasta D, se ejemplifican asignaciones de confianza para tags clase aluminio. En los paneles E y F se muestra como se asignan confianzas a diferentes clases de tags. En cada panel los tags genómicos que comparten la misma secuencia se muestran en colores distintos del plomo. También se muestra la posición dentro del transcrito en los casos donde un tag mapea dentro de este. Tags pertenecientes a la clase 'cercanos a poli(A)' son identificados por la presencia de una región de poliadeninas río abajo de su secuencia. Antes de asignar clases de confianza a los tags, la probabilidad de observación experimental de cada instancia de un tag en particular es obtenida de la Tabla 3. Luego, todas las razones de probabilidad son calculadas utilizando los mayores valores en el numerador. En aquellos casos donde una instancia de un tag en particular exhibe una razón de probabilidad mayor a 5 cuando es comparada contra todas las otras instancias, una confianza alta se le asigna a esa instancia del tag. Las instancias restantes son clasificadas como de confianza baja. La confianza indefinida es asignada cuando la razón de probabilidades entre dos instancias es menor que el umbral 5.

2.2.- Aplicación del método HGA al genoma de levadura

En el caso de Saccharomyces cerevisiae, sus tablas de proteínas sólo especifican las regiones codificantes de cada gen (genes verificados, dudosos y hipotéticos), mientras que las regiones no traducidas (UTRs) en los extremos 5' y 3' de los genes no están anotadas, razón por la cual es necesario asignar estas regiones. Para levadura, aproximadamente el 50% de sus genes conocidos tiene un UTR 3' predicho con alta exactitud. Esta predicción se basa en la identificación de señales de poliadenilación río abajo del codón de término (Graber y col, 2002). Para aquellos casos donde la asignación del UTR 3' no estaba disponible, un largo fijo fue asignado. Todos los autores que han reportado experimentos de SAGE en levadura han asignado arbitrariamente un largo fijo de 500 nts río abajo del codón de término, en la ausencia de un gen anotado a una distancia más corta. En este trabajo se sugiere que para estimar el largo de los UTR 3' de genes que no poseen información de esta región, se debería ocupar información experimental o predicha de los UTRs 3' para estimar un largo fijo. Se decidió asignar un largo fijo de 370 nucleótidos río abajo del codón de término, a los genes que no tuvieran un UTR 3' predicho, debido a que sobre el 95% de los UTRs 3' predichos de levadura son de una longitud menor que este largo asignado (figura 5).

Una vez que todos los UTRs 3' fueron asignados a sus respectivos genes, el método HGA completa la anotación de los transcritos con la asignación de los UTRs 5'. En el caso de levadura, poco se conoce acerca de los UTRs 5', pero en este trabajo se asignó a esta región un largo fijo de 100 nts. Esto, debido a que más del 95% de los tags experimentales que mapean dentro del UTR 5' son observados a una distancia menor de 100 nts desde el codón de inicio del marco abierto de lectura (Zhang & Dietrich, 2005).

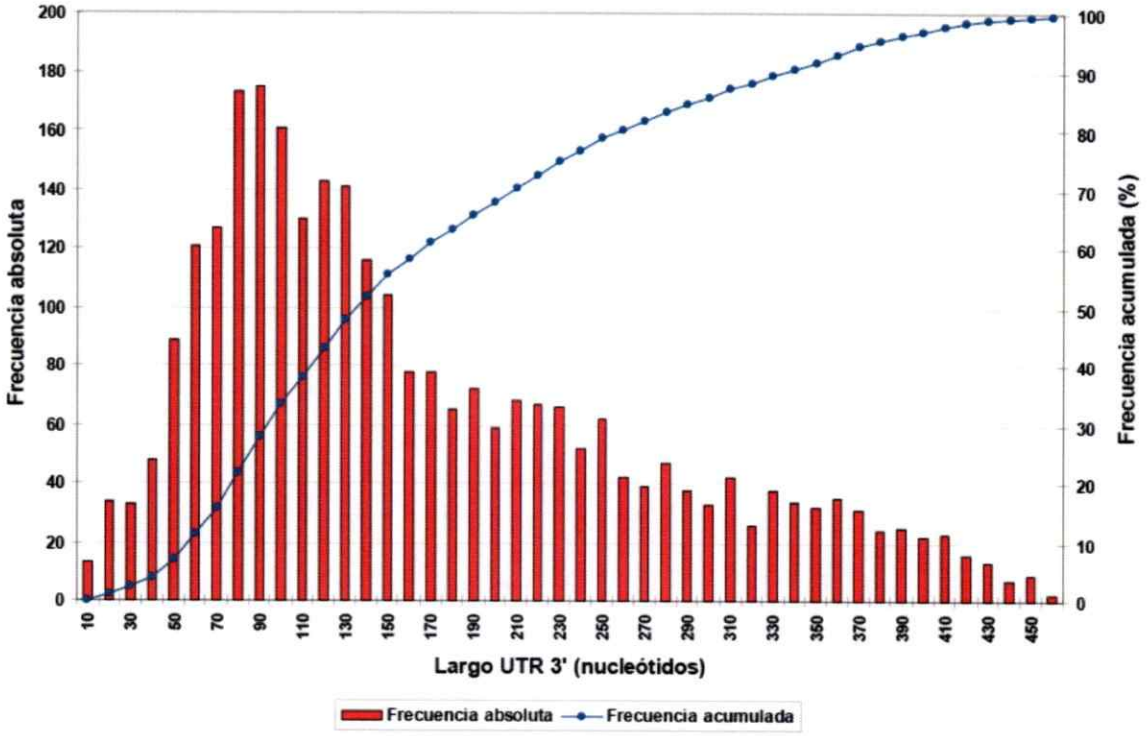


Figura 5: Histograma de largos de UTRs 3' de levadura. Gráfico generado a partir de los datos de Graber y col, 2002.

Luego de este punto, los genes conocidos e hipotéticos han sido anotados lo más precisa y completamente posible (Figura 2.1A). Luego, tablas de ARN son utilizadas para mapear y asignar los transcritos no traducidos al genoma. Esta característica del HGA es nueva, debido a que trabajos anteriores en SAGE no han utilizado explícitamente información proveniente de transcritos no traducidos para mapear tags experimentales. De hecho, la mayoría de los transcritos no traducidos no contiene colas poli(A), y por lo tanto, no deberían ser observados en experimentos de SAGE. Sin embargo, un estudio reciente demostró que algunos ARNs ribosomales de levadura son poliadenilados, incluso en la ausencia de las señales de poliadenilación canónicas (Kuai y col, 2004).

Como se señala en la sección 2.1.2, apareamiento del oligo dT a secuencias poliadeniladas internas durante la transcripción reversa ocurre con una alta frecuencia (Nam y col, 2002), por lo que todos los tags cercanos y río arriba de una región de poli(A) interna fueron categorizados como 'cercanos a poli(A)' (Para detalles ver materiales y métodos). Por lo tanto, se incluyeron transcritos no traducidos e información de regiones de poli(A) internas a la anotación del genoma de levadura para la construcción de la librería virtual de tags de SAGE en levadura.

Como se mencionó en el punto 2.1.3, ciertos tags virtuales genómicos son buscados por su ocurrencia en librerías de tags de SAGE experimentales, procedimiento que permite extraer la probabilidad de observar experimentalmente cada tag potencial del genoma. Para el caso de levadura, todos los experimentos de SAGE en levadura descritos hasta hoy se encuentran listados en la Tabla 2. La probabilidad de obtener un tag de SAGE en levadura de acuerdo a la posición dentro de los transcritos se mostró en la Figura 3B. Los tags experimentales seleccionados para determinar estos valores de probabilidad son los tags de clase Pt y etiquetados como 'no cercanos a poli(A)'.

Tabla 2: Librerías publicadas de tags de SAGE experimentales de Levadura.

ID	TAGs UNICOS	DESCRIPCION	REFERENCIA
Var-1	908	Fase exponencial durante el proceso de fermentación vínica	(Varela y col, 2005)
Var-2	725	Fase estacionaria temprana durante el proceso de fermentación vínica	(Varela y col, 2005)
Var-3	641	Fase estacionaria tardía durante el proceso de fermentación vínica	(Varela y col, 2005)
Vel-1	2.226	Crecimiento logarítmico	(Velculescu y col, 1997)
Vel-2	2.341	Detenido en fase S	(Velculescu y col, 1997)
Vel-3	2.154	Detenido en límite de G2/M	(Velculescu y col, 1997)
Kal-1	1.268	Levadura nativa crecida en oleato	(Kal y col, 1999)
Kal-2	649	Levadura mutante Pip2/oaf1 crecida en oleato	(Kal y col, 1999)

La probabilidad de observar un tag generado de una región clasificada como intergénica (probabilidad establecida considerando sólo los tags pertenecientes a las clases Cu y Fe) en el genoma de *Saccharomyces cerevisiae* es ilustrada en la figura 3C, y corresponde a un 11.3%. En esta figura también es ilustrada la probabilidad de obtener un tag experimental clase Pt y categorizado como 'cercano a poli(A)', probabilidad que corresponde a un 34.7%. Las razones de las probabilidades de todas las posibles combinaciones de las diferentes categorías generadas en este trabajo son mostradas en la Tabla 3.

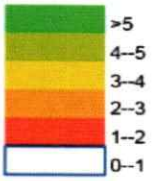
2.3.- Anotación de los tags virtuales genómicos de SAGE en levadura por el método HGA

Se aplicó el método HGA al genoma entero de *Saccharomyces cerevisiae* (Tabla 4). Se encontró que el 80% de los 76.826 tags potenciales del genoma son únicos (tags clases platino y cobre). 54% de los tags potenciales mapearon a regiones intergénicas y el 46% restante a las regiones anotadas como transcritas. Cerca del 60% de estos tags intergénicos tenían un gen anotado en la hebra opuesta, aunque esto se esperaba debido a la alta densidad de regiones codificantes presentes en el genoma de levadura.

Cuando el HGA fue aplicado al genoma de levadura, 82% de todos los tags virtuales fueron clasificados como de alta confianza, reduciendo la ambigüedad en un 2% dentro de los tags que no son únicos en el genoma, tags que abarcan el 20% de los tags únicos. Por lo tanto, se alcanzó un beneficio cercano al 10% cuando se utiliza el HGA para reducir la ambigüedad de los tags virtuales no únicos en este genoma.

Tabla 3: Razones de probabilidad para las asignaciones jerárquicas de tags en el genoma de levadura.

		TAG							
		Transcritos sin sitios de poliadenilación interno						Cercano a poli(A)	Intergénico
		Posición de tag en el transcrito							
		1	2	3	4	5	>=6	X	N.A.
	Probabilidad	**0,863**	**0,140**	**0,048**	**0,021**	**0,021**	**0,018**	**0,347**	**0,113**
1	**0,863**	1,00	6,16	18,17	41,31	41,31	47,16	2,49	7,64
2	**0,140**	0,16	1,00	2,95	6,71	6,71	7,66	0,40	1,24
3	**0,048**	0,06	0,34	1,00	2,27	2,27	2,60	0,14	0,42
4	**0,021**	0,02	0,15	0,44	1,00	1,00	1,14	0,06	0,18
5	**0,021**	0,02	0,15	0,44	1,00	1,00	1,14	0,06	0,18
>=6	**0,018**	0,02	0,13	0,39	0,88	0,88	1,00	0,05	0,16
Cercano a poli(A)	**0,347**	0,40	2,48	7,30	16,61	16,61	18,96	1,00	3,07
Intergénico	**0,113**	0,13	0,81	2,38	5,41	5,41	6,17	0,33	1,00



Las razones de probabilidades entre todos los tipos de tags son calculadas en base a las ocho probabilidades calculadas para la observación de tags experimentales mostrada en las figuras 3B-C. Estas probabilidades están incluidas en la quinta fila y en la segunda columna de la tabla, mientras que la descripción del tipo de tag es mostrada en la segunda, tercera fila y en la primera columna de la tabla. Las razones de probabilidades son calculadas al considerar en el numerador la probabilidad de un tipo de tag mostrado en cada fila; la correspondiente probabilidad del tipo de tag mostrada en cada columna es utilizada en el denominador.

Como se presume, la mayoría de los tags virtuales que mapearon en un transcrito anotado se encontraban en la región codificante (80%), mientras que un 15% mapeó en los UTRs 3' y sólo un 5% fue encontrado en los UTRs 5'. Estos resultados se correlacionan con los largos observados de estos elementos. Sólo una pequeña fracción de los tags virtuales se encuentra en intrones (1%), mismo porcentaje de tags virtuales que se encuentran en genes no traducidos. Muy pocos tags potenciales se encuentran en el límite exón-intrón (0,02%), totalizando 13 nuevas secuencias de tags generadas debido a splicing *in silico*. El número total de tags virtuales versus posición del tag dentro del transcrito muestra una relación linear inversa, como es de esperarse, debido a que las posiciones de los tags dentro de los transcritos se correlaciona con el largo de los genes, por ejemplo gran cantidad de genes tendrán tags en la posición 1 en comparación con los genes que tendrán tags en la posición 10.

Cuando los tags virtuales únicos del genoma fueron considerados, tags que representan los tags potenciales que pueden ser observados por un experimento de SAGE, los resultados anteriores permanecen sin variaciones para la mayoría de las categorías de tags (Tabla 4, columnas derechas). Las únicas clases de tags que mostraron diferencias significativas en comparación con los resultados mencionados anteriormente involucran clases y confianzas de tags. En el primer caso, la fracción de tags pertenecientes a las clases platino y cobre, clases que representan tags únicos en el genoma, aumentan. Las otras clases de tags disminuyen por lo menos dos veces en proporción a la anotación del genoma completo, debido a que la mayoría de los tags no únicos están repetidos dos veces en el genoma.

Tabla 4: Anotación de tags de SAGE virtuales de levadura según el método HGA.

	Todos los tags		Tags únicos	
	Frecuencia	**Porcentaje**	**Frecuencia**	**Porcentaje**
Clases				
Platino (Pt)	28.948	37,68	28.948	43,40
Oro (Au)	2.075	2,70	815	1,22
Plata (Ag)	363	0,47	143	0,21
Aluminio (Al)	4.139	5,39	1.580	2,37
Cobre (Cu)	31.844	41,45	31.844	47,74
Fierro (Fe)	6.417	8,35	2.037	3,05
Silicio (Si)	3.040	3,96	1.336	2,00
Confianza				
Alta	62.780	81,72	62.780	94,12
Indefinida	11.082	14,42	3.923	5,88
Baja	2.964	3,86	0	0,00
Contexto transcrito				
UTR5	1.669	4,70	1.481	4,70
ORF (CDS)	28.645	80,63	25.049	79,56
UTR3	5.211	14,67	4.956	15,74
Intergénico y transcritos				
Intergénico	41.301	53,76	35.217	52,80
Opuesto a transcrito	30.208	39,32	27.407	41,09
Completo en transcrito	35.172	45,78	31.143	46,69
Parcial en transcrito	353	0,46	343	0,51
Total transcritos	35.525	46,24	31.486	47,20
Intrones				
Intrón	380	1,07	212	0,67
No intrón	35.145	98,93	31.274	99,33
Cercano a poli(A)				
Cercanos a poli(A)	1.606	4,52	1.541	4,89
No cercanos a poli(A)	33.919	95,48	29.945	95,11
ARNs no traducidos				
ARNs no mensajeros	367	1,03	203	0,64
Splicing				
Tags modificados por splicing	13	0,02	12	0,02
Posición del tag en transcrito				
Posición 1	6.415	18,06	6.156	17,48
Posición 2	5.733	16,14	5.423	15,40
Posición 3	4.943	13,91	4.306	12,23
Posición 4	4.054	11,41	3.542	10,06
Posición 5	3.217	9,06	2.787	7,91
Posición 6	2.537	7,14	2.186	6,21
Posición 7	1.965	5,53	1.672	4,75
Posición 8	1.516	4,27	1.267	3,60
Posición 9	1.163	3,27	972	2,76
Posición 10	894	2,52	754	2,14
Total				
Total	76.826	100,00	66.703	100,00

En el caso de las confianzas de los tags, el número total de tags con confianza alta permanece igual, pero la fracción de esta clase de tags aumenta un 12% debido a que el número total de tags virtuales es menor, ya que la mayoría de los tags no únicos son descartados. La proporción de tags con confianza indefinida disminuye en aproximadamente un 9% porque muchas instancias de los tags repetidas fueron eliminadas. En los casos donde un tag tiene múltiples instancias en el genoma y una de ellas tiene confianza alta o indefinida, las instancias con confianza baja no son consideradas debido a que sólo sus contrapartes de confianza alta o indefinida son asignadas.

2.4.- Mapeo de tags de SAGE experimentales de levadura contra la anotación generada por el método HGA

Se colectaron todos los datos publicados de experimentos de SAGE en levadura (Tabla 2). Luego se usó la anotación de los tags virtuales generada por el método HGA, para este organismo, para asignar todos los tags experimentales (Tabla 5). Algunos de estos resultados contribuyeron a validar el método HGA. Primero, entre el 82 y 90% de los tags experimentales mapearon a transcritos y no a regiones intergénicas, como es de esperarse para un organismo con tan completa anotación. La mayoría de estos tags pertenecieron a las clases platino, aluminio y oro. Segundo, una gran fracción de estos tags fueron asignados a regiones codificantes y UTR-3' de transcritos y sólo unos pocos mapearon en UTRs 5'. Para todos los casos, más del 91% de los tags experimentales, que mapearon en el genoma, fueron clasificados con confianza alta de acuerdo a la anotación del HGA. Estos hechos sugieren que el método HGA es confiable. Por otro lado, hay algunos

hechos interesantes, nunca antes observados, vinculados a datos de SAGE experimentales que se presentan en este trabajo.

Tabla 5: Asignación de tags de SAGE experimentales en levadura utilizando el método HGA.

	Varela		Velculescu		Kal	
	Frecuencia	Porcentaje	Frecuencia	Porcentaje	Frecuencia	Porcentaje
Clases						
Platino (Pt)	884	69,44	2.187	72,88	993	75,17
Oro (Au)	54	4,24	134	4,47	68	5,15
Plata (Ag)	5	0,39	12	0,40	5	0,38
Aluminio (Al)	112	8,80	255	8,50	139	10,52
Cobre (Cu)	183	14,38	304	10,13	82	6,21
Fierro (Fe)	26	2,04	75	2,50	23	1,74
Silicio (Si)	9	0,71	34	1,13	11	0,83
Confianza						
Alta	1.177	92,46	2.756	91,84	1.204	91,14
Indefinida	96	7,54	245	8,16	117	8,86
Posición dentro de transcrito						
Primera posición	665	67,36	1586	65,16	799	71,66
No primera posición	322	32,64	848	34,84	316	28,34
Contexto transcrito						
UTR5	15	1,52	29	1,19	14	1,26
ORF (CDS)	574	58,16	1.699	69,80	693	62,15
UTR3	398	40,32	706	29,01	408	36,59
Intergénicos y transcritos						
Intergénico	190	16,14	322	11,68	89	7,39
Opuesto a transcrito	140	11,89	255	9,25	60	4,98
Completo en transcrito	973	82,67	2.407	87,34	1.098	91,20
Parcial en transcrito	14	1,19	27	0,98	17	1,41
Total transcritos	987	83,86	2.434	88,32	1.115	92,61
Intrones						
Intrón	6	0,61	5	0,21	1	0,09
Cercano a poli(A)						
Cercanos a poli(A)	70	7,09	189	7,76	73	6,55
ARNs no traducidos						
ARNs no mensajeros	12	1,22	7	0,29	5	0,45
Splicing						
Tags modificados por splicing	1	0,08	1	0,03	4	0,30
Total						
Total	1.273	100,00	3.001	100,00	1.321	100,00

Primero, aunque hay pocas instancias, se encontraron tags que mapearon a intrones en todos los experimentos de SAGE reportados. Segundo, en todos los experimentos de SAGE una fracción significante de tags experimentales está localizada cerca de un poli(A) interno. Tercero, en todos los experimentos reportados, tags de SAGE mapearon a genes no traducidos. Casi todos estos últimos casos corresponden a tags clasificados como 'no cercanos a poli(A)' y mapean en la primera posición dentro de los transcritos, sugiriendo que poliadenilación en el extremo 3' ocurre frecuentemente en ARNs ribosomales. Cuarto, de manera análoga a lo observado en intrones, tags generados como consecuencia de splicing fueron observados en todos los experimentos de SAGE. Esta es la primera vez en que tags de SAGE experimentales son mapeados a tags potenciales provenientes de splicing desde un genoma. Quinto, una fracción significante de tags de SAGE experimentales mapean a regiones del genoma que actualmente están anotadas como intergénicas. Aunque este hecho ya ha sido previamente observado, debe mencionarse que es la primera vez que este análisis es llevado a cabo considerando grados de confianza de los tags, así otorgando una visión más precisa de estos casos. Estos tags intergénicos podrían representar nuevos genes no descritos aún en levadura. Por lo que utilizando la anotación basada en el HGA, ahora ellos pueden ser fácilmente ordenados de acuerdo a su grado de confianza, lo cual facilita y permite optimizar futuros experimentos que permitan el descubrimiento de estos posibles nuevos genes.

Finalmente, una gran fracción de los tags experimentales que mapearon en una región intergénica tienen un gen anotado en su hebra opuesta. Estos tags podrían corresponder a nuevos genes o a nuevos elementos regulatorios, como ARNs antisentido (Quere y col, 2004; Shena y col, 1995). El mapeo genómico detallado basado en el método HGA de todos los tags experimentales disponibles de levadura está disponible como material suplementario.

2.5.- Comparación entre las asignaciones experimentales previas de tags de SAGE y la anotación realizada por el método HGA

Se compararon las asignaciones de tags de SAGE experimentales realizadas por los autores de los diferentes experimentos de SAGE en levadura (Tabla 2) con las efectuadas por el método HGA (Tabla 6). Cuando se consideran todos los tags de SAGE experimentales por autor, entre 8 y 10% de las asignaciones ambiguas pudieron ser clasificadas como no ambiguas por el método HGA. En estos casos, los autores de los experimentos de SAGE asignaron un tag a dos o más genes, mientras que el HGA asignó estas mismas secuencias de tags a un sólo gen con una alta confianza. Cuando se calcula la ganancia de asignaciones no ambiguas realizada por el HGA, considerando sólo los casos con ambigüedad (por ejemplo, aquellos tags con múltiples asignaciones realizadas por los autores) independientemente de la clase del tag, los porcentajes obtenidos son altamente significativos (Figura 6). Para los tags de las clases oro, aluminio y silicio, se logró desde un 57 a un 70% de ganancia de asignaciones no ambiguas. Por lo que, la reducción de la ambigüedad en la asignación de tags, cuando se ocupa HGA es relevante. La ganancia de alrededor de 16% obtenida para los tags clase plata es baja, debido a que en la mayoría de estos casos todas las instancias de una secuencia de un tag plata mapean en las primeras posiciones de cada transcrito. Sin embargo, esta clase de tag es la menos abundante, por lo que tiene un pequeño impacto en el valor absoluto de la ganancia que se logra. Por otro lado, también se encontraron algunas asignaciones conflictivas, entre la asignación original y la desarrollada en este trabajo (Tabla 6). En el caso de los tags pertenecientes a la clase platino, los autores asignaron un gen distinto del asignado por HGA.

Tabla 6: Comparación de las asignaciones de los tags de SAGE experimentales de levadura.

	Varela		Velculescu		Kal	
	Frecuencia	**Porcentaje**	**Frecuencia**	**Porcentaje**	**Frecuencia**	**Porcentaje**
Asignaciones realizadas por autores						
Anotación única	1.067	83,82	2.491	83,01	1.074	81,30
Anotaciones múltiples	206	16,18	510	16,99	247	18,70
Asignaciones realizadas por el método HGA						
Anotación única	1.177	92,46	2.756	91,84	1.204	91,14
Anotaciones múltiples	96	7,54	245	8,16	117	8,86
Ganancia de asignaciones no ambiguas realizadas por el método HGA						
Ganancia	110	8,64	265	8,83	130	9,84
Asignaciones conflictivas (entre las asignaciones realizadas por los autores y el método HGA)						
Confianza alta	69	5,42	154	5,13	115	8,71
Clase platino	41	3,22	90	3,00	36	2,73
Clase cobre	28	2,20	64	2,13	79	5,98
Número total de asignaciones						
Total	1.273	100,00	3.001	100,00	1.321	100,00

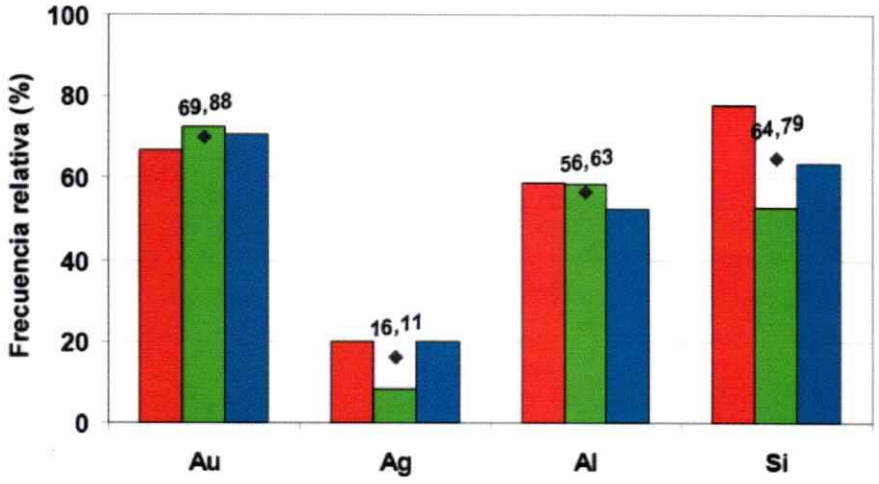


Figura 6: Ganancia de asignaciones no ambiguas realizadas por el método HGA. El porcentaje de casos donde los otros autores asignaron dos o más genes a un tag, mientras que el método HGA asignó la misma secuencia del tag experimental a una sola región con alta confianza es graficado independientemente para cada clase y para cada una de las publicaciones de los experimentos de SAGE en levadura. Barras rojas corresponden a los datos acumulados de los experimentos de Var-1, Var-2 y Var-3 (Varela y col, 2005); las barras verdes corresponden a los datos acumulados de los experimentos de Vel-1, Vel-2 y Vel-3 (Velculescu y col, 1997); y las barras azules corresponden a los datos acumulados de los experimentos de Kal-1 y Kal-2 (Kal y col, 1999). En los casos de tags pertenecientes a la clase silicio, el método HGA los asignó a una sola región intergénica, mientras que los otros autores asignaron estos tags a las instancias de los genes anotados.

La pequeña cantidad de asignaciones conflictivas de esta clase de tags contribuye a validar el método HGA. Para los tags pertenecientes a la clase cobre, en la literatura se asignaron algunos tags de esta clase a genes, mientras que según el HGA estos corresponden a regiones intergénicas. La mayoría de estas asignaciones conflictivas son debido al distinto largo asignado al UTR-3' entre los distintos autores (largo fijo de 500 nucleótidos) y el HGA (largo variable, con un valor máximo de 370 nts para los casos donde en gen no tiene un UTR-3' predicho).

2.6.- Caracterización de las proteínas hipotéticas observadas por SAGE durante la fermentación alcohólica

El mapeo de tags realizado por el método HGA depende en sus inicios de la disponibilidad y de la calidad de las tablas de proteínas anotadas para cada organismo (Figura 2.1A). En el caso de levadura, sus tablas de proteínas son bastante completas y aún así algunas descripciones de genes hipotéticos y dudosos no están disponibles. Esto es aún más evidente en las tablas de proteínas de organismos cuyos genomas están parcialmente anotados, por ejemplo *Xenopus tropicalis*, y en tablas de proteínas de ciertas bases de datos donde los genes hipotéticos no tienen ninguna descripción, por ejemplo las tablas de proteínas de NCBI. En vista de lo anterior y para asignar lo más completamente posible los tags de SAGE experimentales, se desarrolló un programa computacional que permite superar estas limitaciones. De modo de aumentar el conocimiento con respecto a los perfiles de expresión génica durante la fermentación alcohólica, se ejecutó este programa utilizando como archivo de entrada aquel que contiene las secuencias en formato FASTA de todos los genes sin descripción, generalmente genes hipotéticos y dudosos, que se expresan durante la fermentación alcohólica (Varela y col, 2005).

Los genes hipotéticos que originaron tags clase platino, que mostraban los mayores porcentajes de identidad y que sobre el 70% de su secuencia se alineaba con alguna proteína conocida se ilustran en la Tabla 7.

2.7.- SAGExplore, una aplicación web basada en el método HGA

Para hacer el método HGA disponible para toda la comunidad científica, se creó el servidor web llamado SAGExplore. Este servidor (Figura 5) permite el mapeo de tags de SAGE contra la anotación de todas las tags de levadura llevada a cabo previamente por el método HGA. También permite estimar *a priori* los resultados de SAGE para algún organismo en particular, lo que finalmente permite el diseño adecuado de un experimento de SAGE. El servidor está construido sobre una base de datos producida por el método HGA y tiene características que permiten el fácil y rápido manejo de datos experimentales de SAGE. También ofrece múltiples representaciones gráficas que facilitan en gran medida el análisis de los datos experimentales. Junto con lo anterior, posee herramientas que permiten el análisis de secuencias, como por ejemplo, la base de datos está ligada directamente a la aplicación BLAST de NCBI, permitiendo así el alineamiento de cualquier región de interés del genoma contra la base de datos de Genbank, entre otras muchas características.

2.7.1.- Genome explore, una herramienta de SAGExplore que permite explorar genomas en términos de los tags potenciales que se observarían por SAGE

Uno de los enlaces que posee SAGExplore es 'Genome explore' (Figura 5a), página que permite examinar el genoma de un determinado individuo basándose en los tags de SAGE potenciales clasificados por el método HGA.

Tabla 7: Caracterización de genes hipotéticos que se expresan durante la fermentación alcohólica.

TAG	Locus	Exp[1]	Early[2]	Late[3]	ID*	e-value	Descripción
CATGTCAAACATCC	YLR327C	6	4	41	100	1,0E-180	ATP synthase regulating factor
CATGTAGCTTTTTC	YJR096W	6	21	3	94,8	3,0E-25	Related to 2, 5-diketo-D-gluconic acid reductase [Neurospora crassa]
CATGTATGTCGGTG	YFR006W	3	17	0	75,6	6,0E-70	Prolidase [Aspergillus nidulans]
CATGTTATCAGAGG	YNR036C	3	8	13	67,6	7,0E-06	Ribosomal protein S12 [Solanum nigrum]
CATGGGTGGCGAGG	YGR210C	0	0	10	65,1	4,0E-25	GTP-binding protein [Schizosaccharomyces pombe]
CATGAGACAAACTT	YJL200C	0	0	6	60,9	0,0E+00	Aconitate hydratase [Schizosaccharomyces pombe]
CATGACACCTGCTG	YPL088W	0	0	6	58,7	1,0E-132	Putative oxidoreductase [Schizosaccharomyces pombe]
CATGTATAGGTCAA	YHR112C	6	0	6	53,1	6,0E-96	Putative cystathionine gamma-synthase [Schizosaccharomyces pombe]
CATGCAAGAAGCGG	YJR096W	0	4	3	51,2	2,0E-74	Related to 2, 5-diketo-D-gluconic acid reductase [Neurospora crassa]
CATGTAAATATGTT	YGL068W	6	0	3	50,5	7,0E-86	60S ribosomal protein L7
CATGAAACCGTCCC	YPL183W-A	3	4	3	46,6	1,0E-154	Mitochondrial ribosomal protein L36; putative BRCA1-interacting protein [Homo sapiens]
CATGAAACCGTCCC	YPL183W-A	3	4	3	45,6	2,0E-32	Ribosomal protein L36 [Deinococcus radiodurans]
CATGCGACCATCGT	YKL107W	6	8	0	45,5	4,0E-32	Probable oxidoreductase protein [Ralstonia solanacearum]
CATGCGATTGAATA	YIL103W	0	8	0	45,4	2,0E-79	Putative diphthamide synthesis protein [Caenorhabditis elegans]
CATGCATACTCTAT	YKL091C	6	4	0	45,2	2,0E-09	SEC14 cytosolic factor (Phosphatidylinositol)
CATGATTCTCTTTT	YNL200C	3	4	0	44,3	3,0E-77	apoA-I binding protein [Homo sapiens]
CATGCTGATTACGG	YGL101W	3	4	0	43,7	1,0E-60	Similar to CGI-130 protein [Homo sapiens]
CATGGTTGTGAATA	YNL274C	0	8	0	43,7	3,0E-77	Putative 2-hydroxyacid dehydrogenase [Schizosaccharomyces pombe]
CATGTTAATGAAAA	YBL036C	6	0	0	43,7	1,0E-60	Proline synthetase co-transcribed [Mus musculus]
CATGGCTGTGACTT	YMR251W	6	0	0	43,3	3,0E-57	Protein with Glutathione S transferase domain [Schizosaccharomyces pombe]
CATGCGTTCATCCG	YDL086W	0	8	0	42,4	1,0E-125	Dienelactone hydrolase family protein [Shewanella oneidensis MR-1]
CATGCAGCAGGTTT	YER156C	3	4	0	42,4	2,0E-44	Yeast hypothetical protein YEY6 like [Caenorhabditis elegans]
CATGGTATTCCTGA	YIR035C	3	8	0	42,2	2,0E-64	Short chain dehydrogenase [Schizosaccharomyces pombe]
CATGCTGTTTTGGG	YKL033W-A	0	8	0	42,1	2,0E-29	Haloacid dehalogenase-like hydrolase [Schizosaccharomyces pombe]
CATGATAAGGATGG	YMR278W	3	4	0	41,4	5,0E-95	Similarity to phosphomannomutases [Schizosaccharomyces pombe]
CATGCAATCGAGGC	YNL045W	3	4	0	40,5	7,0E-66	Probable leukotriene a-4 hydrolase [Schizosaccharomyces pombe]
CATGTAACATTGTG	YOR006C	6	0	0	40,4	1,0E-27	Similar to RIKEN cDNA 0610007P22 gene [Mus musculus]
CATGCAGCAATTTA	YPR004C	6	0	0	40,2	1,0E-40	Electron transfer flavoprotein alpha-subunit [Brucella melitensis]

[1] Frecuencia del tag durante la fase exponencial de la fermentación alcohólica.
[2] Frecuencia del tag durante la fase estacionaria temprana de la fermentación alcohólica.
[3] Frecuencia del tag durante la fase estacionaria tardía de la fermentación alcohólica.
* Porcentaje de identidad calculado por Blast.

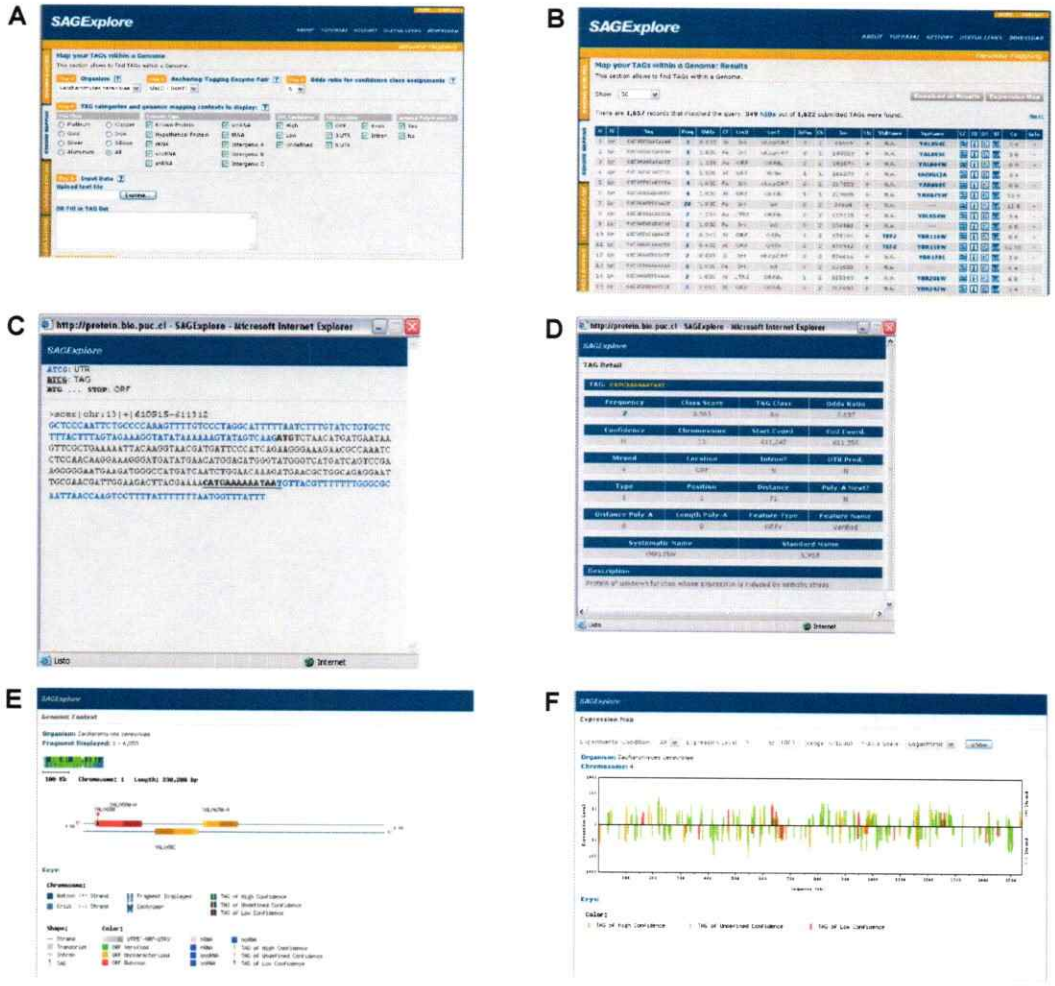


Figura 5: Algunas páginas de SAGExplore. Todas estas páginas están basadas en la anotación producida por el HGA y en información experimental provista por el usuario. **A**.- Página de SAGExplore, llamada 'Genome mapping', que permite mapear los tags experimentales producidos por SAGE mediante la anotación generada por el HGA. Esta página tiene los mismos 4 primeros pasos que la página 'Genome explore'. Junto con mapear los tags experimentales, este formulario también permite seleccionar varias características propias de la clasificación del HGA y el contexto genómico de los tags. Por ejemplo, permite seleccionar tags de alguna clase particular o seleccionar tags que mapeen en regiones intergénicas, UTRs, cerca de poli(A) internos, etc. **B**.- Tabla que se genera luego que SAGExplore asigna todos los tags experimentales a sus respectivas regiones genómicas. Esta tabla tiene los mismos campos que la tabla que genera la página 'Genome explore' luego de haber seleccionado la región genómica que se quiere explorar, excepto por los dos últimos campo (frecuencia del tag e información experimental). Los paneles C, D y E se generan desde esta tabla. **C**.- Ventana que muestra la secuencia nucleotídica del tag seleccionado y sus regiones flanqueantes. En el caso de que el tag seleccionado se encuentre dentro de un transcrito, esta opción mostrará la secuencia de los UTRs (color azul) y del ORF que posee el tag seleccionado, la secuencia del tag estará con letra negrita y subrayada. Cuando el tag seleccionado se encuentra en una región anotada como intergénica, esta ventana presentará la secuencia del tag junto a los 500 nts río arriba y abajo de su secuencia. Esta característica de SAGExplore es principalmente útil para diseñar partidores que amplifiquen la región que contiene el tag (por ejemplo, para comprobar que una región anotada como intergénica se expresa) y porque estas secuencias se pueden alinear automáticamente contra la base de datos de proteínas no redundantes de NCBI. **D**.- Ventana que muestra todas las características que el método HGA asigna a los tags. **E**.- Opción que permite ver el contexto genómico de un determinado tag. Esta vista gráfica proporciona una visión nunca antes apreciada, que permitirá aumentar la comprensión acerca de la producción de tags experimentales de SAGE. **F**.- Mapa de expresión cromosómico que muestra las frecuencias de los tags experimentales, suministrados por el usuario, versus su posición en el cromosoma.

Esta página permite extraer los tags potenciales mediante una combinación de los siguientes pasos: 1.- Elección del organismo de estudio, 2.- Elección del par enzimático, enzima ancla y la enzima etiquetadora (este punto es importante debido a que la enzima etiquetadora determina el largo del tag). Por ejemplo, existe una variación del SAGE tradicional llamada LongSAGE (Saha y col, 2002) donde utilizando la enzima etiquetadora Mmel se producen tags de 21 nucleótidos, 3.- Elección del umbral de la razón de las probabilidades que se considerará para estimar las confianzas de asignación de cada tag, 4.- Selección de las categorías y el contexto genómico de los tags que se quieren analizar, 5.- Selección de la región genómica desde donde se extraerán los tags y 6.- Ingreso de datos de acuerdo a lo escogido en el punto 5. Una vez completados todos los pasos anteriores, el servidor entrega una tabla similar a la de la Figura 5b, donde están todos los tags que cumplen con todos los requisitos anteriores. La página que contiene esta tabla posee una serie de enlaces interesantes, como por ejemplo, un campo de esta tabla está ligado a la Saccharomyces Genome Database (SGD), base de datos que permite examinar el detalle del gen seleccionado. También permite explorar gráficamente los tags virtuales seleccionados, de modo de tener una visión más clara de la región bajo estudio (Figura 5c), entre otras muchas características. En base a lo anterior, el usuario puede planear un experimento de SAGE virtual y estimar *a priori* los resultados que podría obtener ocupando un par enzimático en particular. Por ejemplo, mediante esta estrategia puede ver cuantos tags de alta, baja o indefinida confianza obtendría asumiendo o no un patrón de expresión definido.

También puede evaluar la confianza de los tags que debería obtener de los genes de su interés, de este modo sabiendo anticipadamente si puede evaluar efectivamente estos genes.

2.7.2.- Mapeo de tags experimentales de SAGE mediante SAGExplore

Entre las características más importantes que tiene SAGExplore se encuentra aquella que permite mapear los tags experimentales producidos por SAGE utilizando el método HGA y toda la clasificación desarrollada en este trabajo (Figura 5a). Para hacer uso de esta herramienta, el usuario debe ingresar su librería de tags experimentales con sus respectivas frecuencias y seguir pasos muy similares a los del punto 2.8.1. Cabe destacar que en el paso 4, o sea la elección de las categorías y el contexto genómico de los tags que se desean obtener, es donde el usuario puede enfocar su análisis detallado de los tags. Por ejemplo se puede enfocar sólo en los tags de alta confianza, en los tags clase platino o si la idea es descubrir genes nuevos debe enfocar su análisis principalmente sobre los tags clase cobre. Una vez que el servidor analiza los datos ingresados, genera la página de la Figura 5b. Esta página contiene una tabla que posee los tags mapeados por el método HGA según las opciones elegidas por el usuario. También tiene algunos enlaces a representaciones gráficas útiles para un mejor entendimiento del contexto genómico de los tags y de los perfiles de expresión génica que se producen bajo las condiciones experimentales escogidas para realizar el experimento de SAGE (Figura 5d-e-f). También posee la opción de extraer y visualizar la secuencia que flanquea a un determinado tag, lo que permite hacer un BLAST-X automático de esta secuencia contra la base de datos de proteínas no redundantes del NCBI (Figura 5c). Esta característica es especialmente importante en el caso de genes hipotéticos sin descripción disponible o en el caso de querer descubrir nuevos genes, ya que permite alinear estas regiones contra la base de datos de proteínas y determinar si esta secuencia tiene alta identidad con alguna proteína conocida.

2.7.2.1.- Mapas de expresión cromosómicos

Representaciones gráficas de la expresión génica en cromosomas humanos, revelan un gran orden de organización a nivel genómico, ya que los genes que se sobre-expresan bajo ciertas condiciones están agrupados dentro de una región cromosómica definida (Caron y col, 2001). A raíz de lo anterior, SAGExplore posee entre algunas de sus características únicas, la opción de generar gráficos dinámicos de expresión génica basados en la frecuencia de los tags experimentales obtenidos por SAGE (Figura 5f). Esta herramienta provee información valiosa de los dominios de expresión cromosómicos, pudiendo evaluar por ejemplo cambios de expresión a nivel cromosómico entre los distintos puntos experimentales de SAGE o el descubrimiento de cluster génicos de expresión no identificados anteriormente.

2.7.2.2.- Mapeo de tags contra librerías ya existentes de SAGE

Otra de las características de SAGExplore, es que permite mapear tags experimentales de SAGE contra librerías de SAGE ya existentes. Esta herramienta permite principalmente comparar los niveles de expresión de un tag entre las distintas librerías, o sea, entre distintas condiciones, lo que finalmente permite comparar los niveles de expresión de uno o más genes entre los distintos experimentos de SAGE.

2.7.3- Próximas actualizaciones de SAGExplore

En este momento SAGExplore sólo permite explorar y mapear tags en el genoma de levadura (*Saccharomyces cerevisiae*), pero pronto se incluirán los genomas de

ratón común (*Mus musculus*), rana africana (*Xenopus tropicalis*) y humano (*Homo sapiens*).

2.8.- Descubrimiento de nuevos genes en la levadura EC1118

Utilizando el método HGA para asignar todos los tags experimentales de SAGE obtenidos durante una fermentación alcohólica (Varela y col, 2005), se obtienen 183 tags clase cobre (tags únicos en el genoma que mapean en una región intergénica) y 26 tags clase fierro (tags no únicos en el genoma, pero todas sus instancias mapean en alguna región intergénica). Para confirmar que las regiones genómicas desde donde provienen estos tags se expresan bajo condiciones vínicas, se realizaron experimentos de RT-PCR. Se eligieron los 5 tags clase cobre que mostraron los mayores aumentos en su frecuencia (aumentos estadísticamente significativos) entre la fase exponencial o la fase estacionaria temprana y la fase estacionaria tardía de la cinética de crecimiento de levadura durante el proceso de fermentación alcohólica (Tabla 8). Se obtuvo productos de PCR del tamaño esperado desde 3 de las 5 regiones anotadas actualmente como intergénicas (Figura 6).

Tabla 8: Tags escogidos para el descubrimiento de nuevos genes en la levadura EC1118.

Tags Intergénicos sin ORF en hebra opuesta									
ID[1]	Secuencia de tag	Clase	Exp[2]	Tem[3]	Tar[4]	Cromosoma	Inicio	Término	Hebra
I1	CATGAGGCTACCTA	Cu	0	0	24	XV	908512	908525	-
I2	CATGTAGTTGCTCC	Cu	6	0	20	VII	317113	317126	+
I3	CATGAGAGGTGATC	Cu	0	0	17	VII	811261	811274	+
I4	CATGGAATTTATAG	Cu	0	0	17	XIII	259930	259943	+
I5	CATGGCGACTTGAT	Cu	0	0	17	II	720810	720823	-

[1] Nombre de identificación del tag.
[2] Ocurrencias del tag durante la fase exponencial de la fermentación alcohólica.
[3] Ocurrencias del tag durante la fase estacionaria temprana de la fermentación alcohólica.
[4] Ocurrencias del tag durante la fase estacionaria tardía de la fermentación alcohólica.

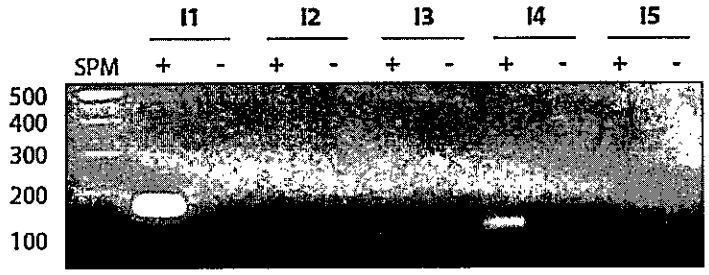


Figura 6: Evaluación de la transcripción de regiones de ADN anotadas actualmente como intergénicas en S. cerevisiae. Resultado del RT-PCR basado en las posiciones de los tags intergénicos descritos en la Tabla 8. Se amplificaron las regiones genómicas que contienen los tags I1 a I5 con partidores en las secuencias flanqueantes (para detalles ver materiales y métodos). Como control negativo (-) para cada muestra se realizó una transcripción reversa en ausencia de la enzima RT. SPM: Standard de peso molecular.

2.8.1.- Comprobación de la transcripción de regiones intergénicas

Algunas de las regiones anotadas como intergénicas, que se demostraron que se expresan bajo condiciones de fermentación alcohólica, se encuentran cerca de extremos de genes conocidos. Debido a esto, se evaluó si las regiones intergénicas estudiadas en este trabajo correspondían a marcos abiertos de lectura (ORF) conocidos con UTRs excepcionalmente largos u ORFs conocidos con UTRs mal anotados o predichos. Para corroborar lo anterior se diseñaron partidores dentro de la secuencia de los ORFs cercanos a la región anotada actualmente como intergénica, de modo que si luego de un RT-PCR se observa alguna banda, esto indica que la región anotada como intergénica corresponde a parte del ORF desde donde se diseñó uno de los partidores. El resultado del procedimiento anterior demostró que ninguna de las regiones que anteriormente se confirmó que se transcriben pertenece a genes cercanos (resultados no mostrados).

2.8.2.- Asignación de función a los genes descubiertos

De modo de poder asignar una posible función a los transcritos que provienen de regiones anotadas actualmente como intergénicas, se procedió a hacer un BLAST-X de las secuencias flanqueantes de los tags asignados a estas regiones. El largo de las regiones extraídas fue de 500 nts a cada lado del tag, por lo que todas las secuencias alineadas tenían un largo total de 1014 nts. Los resultados de este estudio arrojaron que ninguna de las regiones evaluadas contenía sobre 30% de identidad con otra proteína conocida.

2.9.- Descubrimiento de posibles ARNs antisentidos en la levadura EC1118

Como se mencionó en el punto 2.9, gran cantidad de tags de SAGE producidos durante condiciones vínicas mapearon en regiones intergénicas del genoma de levadura. Dentro de esta clase de tags, 218 secuencias fueron asignadas, según el método HGA, a regiones intergénicas donde en la hebra opuesta existe un gen anotado. Para confirmar que algunas de estas regiones intergénicas se transcriben durante una fermentación alcohólica, se realizó una técnica llamada RT-PCR específico de hebra (Kessler y col, 2003), la cual permite demostrar la expresión de una hebra particular del ADN. Luego de efectuar esta técnica, se obtuvo un producto de PCR del tamaño esperado (Figura 7) desde todas las regiones evaluadas (Tabla 9).

Tabla 9: Tags escogidos para el descubrimiento de posibles ARNs antisentidos en la levadura EC1118.

Tags Intergénicos con ORF en hebra opuesta									
ID[1]	Secuencia de tag	Clase	Exp[2]	Tem[3]	Tar[4]	Cromosoma	Inicio	Término	Hebra
O1	CATGGCCAATGATA	Cu	6	8	62	V	468390	468403	-
O2	CATGCCCACGTAAG	Cu	0	4	34	IV	954494	954507	-
O3	CATGGGTAATCGAA	Cu	0	0	27	IV	832428	832441	+
O4	CATGGCCAGGACAA	Cu	3	0	24	III	156712	156725	+
O5	CATGATAATGAGGA	Cu	3	0	24	XIII	14417	14430	+

[1] Nombre de identificación del tag.
[2] Ocurrencias del tag durante la fase exponencial de la fermentación alcohólica.
[3] Ocurrencias del tag durante la fase estacionaria temprana de la fermentación alcohólica.
[4] Ocurrencias del tag durante la fase estacionaria tardía de la fermentación alcohólica.

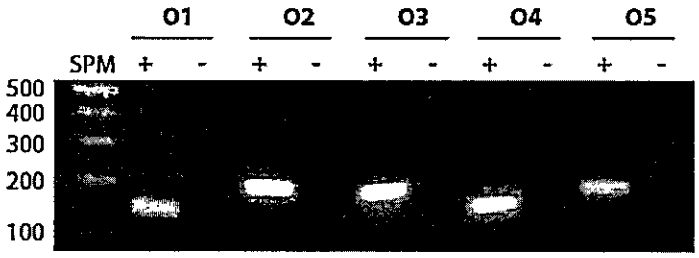


Figura 7: Evaluación de la transcripción de regiones de ADN anotadas actualmente como intergénicas opuestas a un ORF en S. cerevisiae. Resultado del RT-PCR específico de hebra basado en las posiciones de los tags intergénicos de la Tabla 9. Se amplificaron las regiones genómicas que contienen los tags O1 a O5 con partidores en las secuencias flanqueantes (para detalles ver materiales y métodos). Como control negativo (-) para cada muestra se realizó una transcripción reversa en ausencia de la enzima RT. SPM: Standard de peso molecular.

3.- DISCUSION

En este trabajo, se presentó un nuevo método bioinformático llamado Asignación Génica Jerárquica o HGA, procedimiento que permite un preciso y eficiente mapeo de tags de SAGE. El HGA tiene dos grandes ventajas en comparación con los métodos actuales de asignación de tags (Lash y col, 2000; van Kampen y col, 2000; Divina y col, 2004): 1) El nuevo esquema de clasificación de tags. Esquema útil para una identificación preliminar de las características de los tags de SAGE y para inferir sus capacidades en la asignación a transcritos o regiones intergénicas. 2) Asignación de confianzas a todos los tags de SAGE potenciales en el genoma. Estas dos ventajas permiten minimizar el número de asignaciones ambiguas de los tags experimentales de SAGE a regiones genómicas.

3.1.- Mejoras en el proceso de mapeo de tags por el método HGA

El método HGA incorpora varias características nuevas que mejoran la precisión y la integridad del proceso de asignación de tags de SAGE. Estas características se detallan a continuación: primero, en lugar de utilizar sólo las regiones codificantes de genes conocidos e hipotéticos, se asignaron UTRs de la forma más precisa posible, generando así los transcritos putativos más exactos. Transcritos maduros e inmaduros fueron generados al considerar los posibles sitios de splicing, de este modo manteniendo y utilizando toda la información genómica relevante que está disponible. Cuando no hubo información disponible a cerca de los UTRs para algún gen, se utilizó un largo fijo, el cual fue estimado desde datos experimentales (Figura 5). Es importante mencionar que la mayoría de las asignaciones conflictivas entre la asignación del HGA y la de trabajos anteriores (Velculescu y col, 1997; Kal y col, 1999; Varela y col, 2005) (Tabla 6) fue observada debido a los distintos largos de

UTRs asignados entre los distintos métodos. La asignación precisa de los UTRs 3' es fundamental para un mapeo eficiente de tags de SAGE. Este último punto constituye una de las contribuciones importantes del HGA para el análisis de datos producidos por SAGE. Segundo, los ARNs no codificantes, junto con los genes conocidos e hipotéticos, también fueron incluidos en la anotación genómica. Aunque la cantidad de tags que mapean a los ARNs no codificantes anotados en levadura es baja (Tabla 5), es importante incluirlos ya que ellos constituyen una fracción importante en genomas de eucariontes superiores (Stolc y col, 2005; Claverie 2005) y pueden desarrollar funciones esenciales para la célula (Peng y col, 2003). Tercero, tags que mapean a regiones anotadas como intergénicas en el genoma, donde un transcrito se encuentra en la hebra opuesta, también son considerados por el método HGA. Estos supuestos tags intergénicos, en caso de ser observados experimentalmente, podrían corresponder a ARNs antisentido, los cuales tienen un rol importante en la regulación génica de eucariontes (RIKEN y col, 2005). Se demostró en este trabajo que gran cantidad de tags experimentales de levadura fueron asignados a este tipo particular de transcritos (Tabla 5), aún en *S. cerevisiae*, organismo que constituye uno de los genomas mejor anotados en la actualidad. Si solamente información proveniente desde ESTs o transcriptomas se hubiera utilizado para mapear los tags experimentales, no se habrían podido identificar a los tags provenientes de hebras opuestas a transcritos. Por lo tanto, un método que considera directamente estos elementos durante el proceso de anotación, permite acelerar el proceso de descubrimiento de nuevos elementos regulatorios. La identificación de elementos regulatorios de este tipo es importante para una interpretación completa y precisa de los patrones de expresión génicos. Cuarto, al utilizar información genómica en el proceso de mapeo de tags, el método HGA también permite la asignación de tags a regiones donde no existe anotación génica en ninguna de las dos hebras. En este trabajo se demostró que una fracción

significativa de estos tags fue observada en experimentos de SAGE en levadura (Tabla 5). Es importante mencionar nuevamente que aunque la anotación del genoma de levadura es bastante completa en comparación a otros organismos y que una fracción significativa del genoma está actualmente anotada como codificante, la gran cantidad de tags experimentales asignados como intergénicos sugiere que existen muchos transcritos desconocidos aún por descubrir. Esta observación concuerda con los resultados obtenidos de experimentos de microarreglos (Havilio y col, 2005). Quinto, el uso de información genómica junto con la generación de los tags potenciales producidos luego de splicing - tags no disponibles explícitamente desde la secuencia genómica - permite al método HGA estimar la unicidad de los tags de una forma más precisa. Sexto, el registro de las posiciones de regiones de poli(A) internas dentro de los transcritos anotados es también una característica nueva importante del método HGA. Esto se consideró debido a que el análisis reciente de datos de ESTs mostró que una fracción significativa del proceso de transcripción reversa se inicia en regiones de poli(A) internas, regiones de más de 8 adeninas consecutivas (Nam y col, 2002). La asignación de tags experimentales de SAGE realizada en este trabajo (Tabla 5) confirmó esta situación, ya que gran cantidad de tags fueron clasificados como 'tags cercanos a poli(A)'. Es interesante mencionar que cerca del 5% de los tags virtuales que mapean a transcritos fueron clasificados como 'cercanos a poli(A)', dando cuenta de un total de 1.606 ocurrencias en el genoma. Esto sugiere que estas instancias no deberían ser menospreciadas al momento de mapear tags experimentales al genoma. Más aún, en el caso de tags clase platino y 'cercanos a poli(A)', la posición del tag dentro del transcrito no debería ser relevante para estimar la probabilidad de observar experimentalmente esta clase de tags. Esto se verificó al encontrar que un pequeño e insignificante efecto se advierte dependiendo de la posición del tag, tal y como se esperaba (resultados no mostrados). Séptimo,

la nueva definición de clases de tags considerada por el método HGA (Tabla 1) facilita la comprensión del origen genómico del tag, junto con una estimación inicial de la confianza de que éste sea observado en un experimento de SAGE. Octavo, el cálculo de las probabilidades de los tags, a partir de datos experimentales, junto con su nueva clasificación, permiten al método HGA obtener la probabilidad o confianza de que un tag, de alguna región genómica en particular, sea observado experimentalmente. Esto constituye el núcleo del método HGA y una de las contribuciones más importantes de este trabajo para reducir la cantidad de asignaciones ambiguas de tags experimentales en SAGE. Además, también se demostró que cerca del 30% de los tags experimentales que fueron asignados a un transcrito mapearon en alguna posición distinta de la más 3' (Tabla 5). Estos tags correspondían a aproximadamente el 25% de todos los tags experimentales. Por lo tanto, si se consideran sólo los tags más 3' para asignar tags experimentales - metodología que ocupan las bases de datos actuales (van Kampen y col, 2000; Lash y col, 2000; Divina & Jiri, 2004) - una fracción importante de los tags experimentales no serán asignados, provocando una pérdida de información significativa. Finalmente, es importante mencionar que incluso en los casos donde la ambigüedad no puede ser completamente eliminada, el método HGA puede reducir el número de asignaciones posibles, de este modo, disminuyendo la ambigüedad total de un tag cuando su secuencia se encuentra repetida múltiples veces en el genoma.

3.2.- Parámetros del HGA que dependen de la anotación genómica

La probabilidad de obtener tags experimentales desde regiones intergénicas depende en gran medida de la calidad de la anotación del genoma. Para los

genomas que están pobremente anotados, o sea con una pequeña cantidad de genes asignados, esta probabilidad aumentará cuando se aplique el método HGA. Esta es una característica deseable para estimar la probabilidad de esta clase de tags, debido a que el HGA está orientado al descubrimiento de genes nuevos en aquellos organismos donde existe una anotación incompleta. En *Saccharomyces cerevisiae*, el 11,3% de todos los tags experimentales producidos por SAGE (Velculescu y col, 1997; Kal y col, 1999; Varela y col, 2005) mapean en regiones actualmente anotadas como intergénicas. Esto sugiere que aún queda por descubrir una cantidad significativa de transcritos, incluso en genomas tan bien anotados como el de levadura.

3.3.- Significancia de la anotación, basada en el método HGA, para asignar tags en genomas complejos

En este trabajo, el método HGA logró un aumento cercano al 10% en las asignaciones no ambiguas cuando se consideraron todos los tags experimentales de SAGE en levadura (Velculescu y col, 1997; Kal y col, 1999; Varela y col, 2005) (Tabla 6). Estos valores mejoraron hasta un 70% cuando se consideraron sólo aquellos tags con múltiples repeticiones en el genoma (Figura 4). En levadura, organismo que tiene un genoma relativamente pequeño - alrededor de 12 millones de pares de bases - los tags únicos de 14 nts constituyen el 87% de todos los tags virtuales genómicos. Por lo tanto, un aumento del 10% en las asignaciones de tags no ambiguas tiene un interés limitado. Sin embargo, cuando genomas de mayor tamaño son considerados (por ejemplo, genomas de billones de pares de bases), la fracción de tags únicos de 14 nts en el genoma se reduce de manera significativa a aproximadamente un 10% de todos los tags virtuales genómicos (Figura 6). Por lo tanto, actualmente el principal problema de utilizar secuencias de grandes genomas

para el mapeo de tags es que su gran tamaño y complejidad hacen más improbable la unicidad de las secuencias de los tags, lo que hace aún más difícil una asignación de tags no ambigua. Es en estos casos donde el método HGA sería de máxima utilidad, debido a que la reducción de asignaciones ambiguas será mucho más importante que lo observado en levadura.

LongSAGE ha sido propuesto para reducir la ambigüedad del mapeo de tags a grandes genomas (Saha y col, 2002; Wahl y col, 2004). Sin embargo, esta técnica tiene algunas desventajas importantes, tales como: 1) Altos costos en comparación a la técnica SAGE tradicional, 2) baja eficiencia de secuenciación (menos tags por clon) y 3) aumento significativo de las tasas de error de secuenciación en tags de 20-21 nts., errores que se estiman que ocurren en un 20% de los tags experimentales derivados de LongSAGE (Akmaev & Wang, 2004). Si SAGE o LongSAGE debe ser utilizado para analizar transcriptomas completos, es aún tema de debate. La generación de tags de SAGE de 30 nts implica triplicar los costos de secuenciación cuando es comparado con tags de 10 nts. Este incremento sólo conlleva a un aumento de asignaciones no ambiguas cercano al 8% cuando se utilizan las bases de datos UniGene para el mapeo de tags (Lee y col, 2002). Contrario a esto, otros autores han demostrado que un aumento en el largo de la secuencia del tag es crucial para una asignación eficiente de tags experimentales a grandes genomas (Wahl y col, 2005). Para estimar de mejor manera cuál metodología de SAGE se debería utilizar, en este trabajo se desarrollaron algunos análisis bioinformáticos de los tags virtuales generados a partir de la última versión del genoma de *Xenopus tropicalis*. Este genoma tiene 1.5 billones de pares de bases, lo cual es la mitad del tamaño del genoma humano y 125 veces más grande que el genoma de levadura. Los estudios realizados sobre todos los tags potenciales del genoma de *X. tropicalis* mostraron que el 9,1% de los tags potenciales de 14 nts son únicos en el genoma, una fracción muy baja en

comparación al 80,6% de tags potenciales únicos de 21 nts (Figura 7). Cuando se construyó un histograma de frecuencia de los tags virtuales dentro del genoma de *X. tropicalis* (Figura 8), se encontró que el 60% de las secuencias de tags virtuales de 14 nts tenían menos de 9 ocurrencias en el genoma, mientras que el 90% tenían menos de 20 ocurrencias. Esta baja cantidad de ocurrencias de tags de 14 nts en este genoma sugiere que el uso de la metodología HGA debería permitir una apropiada asignación de tags experimentales de 14 nts en genomas complejos. Más importante aún, cuando se construyó una base de datos de los tags virtuales del genoma de *X. tropicalis* utilizando sólo algunos de los parámetros que ocupa el método HGA, se obtuvo que el 40% de los tags virtuales de 14 nts tuvieron una clasificación de confianza alta. Estos tags representan un aumento cercano al 31% en el número de asignaciones no ambiguas. Este porcentaje de ganancia es significativamente mayor que el 10% obtenido para levadura, aunque su estimación se basó en los tags genómicos virtuales en lugar de los tags experimentales de SAGE, estimación que debería aumentar al ocupar información experimental, tal como se observó en levadura (Tabla 4). Junto con lo anterior y aunque en algunos casos la ambigüedad no podrá ser eliminada completamente, por lo menos será significativamente reducida. Toda esta evidencia sugiere que el método HGA será de mayor utilidad para asignación de tags en genomas grandes y complejos, proporcionando una alternativa eficiente y de bajo costo en comparación a LongSAGE. Sin embargo, genomas más grandes y complejos tienen nuevos desafíos que no fueron abordados en este trabajo con el genoma de levadura. Entre estos desafíos se encuentra la gran cantidad de splicing alternativo y polimorfismos de un nucleótido que se observan en algunos genomas complejos. Estas características de ciertos genomas constituyen un problema para la estimación precisa de las funciones de probabilidad que tienen las distintas clases de tags.

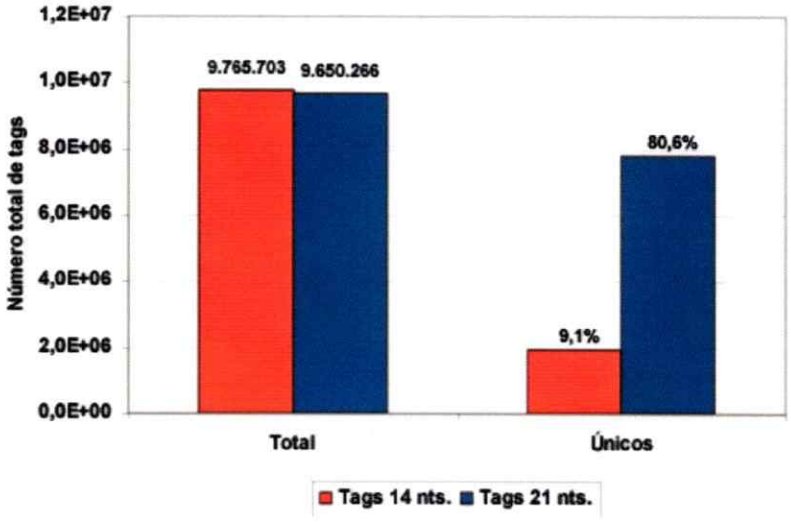


Figura 7: Número de tags virtuales de 14 y 21 nts. en Xenopus tropicalis. Todos los tags potenciales de 14 y 21 nts fueron extraídos desde el genoma de *X. tropicalis* versión 3.0. Una comparación de todos contra todos, dentro del conjunto de tags del mismo tamaño, arrojó que un 9.1% de los tags de 14 nts y un 80.6% de los tags de 21 nts son únicos en el genoma.

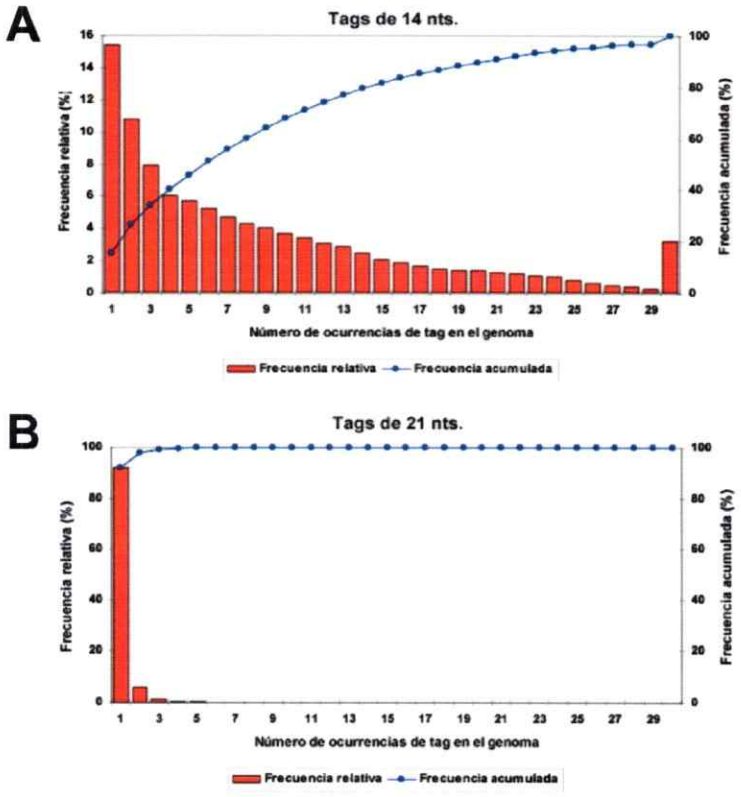


Figura 8: Frecuencia de ocurrencia de secuencias de tags experimentales en X. tropicalis. La frecuencia de ocurrencia de todas las secuencias de tags, únicas o no únicas, fueron calculadas y registradas. El panel A muestra los resultados realizados para tags de 14 nts. Panel B muestra los resultados sobre los tags de 21 nts.

Por lo tanto, será necesario hacer ciertas modificaciones a la metodología HGA propuesta en este trabajo, las cuales dependerán del genoma particular que se esté estudiando.

Actualmente se está adaptando el método HGA para anotar más precisamente todos los tags potenciales de los genomas de rana africana (*Xenopus tropicalis*), ratón común (*Mus musculus*) y humano (*Homo sapiens*).

3.4.- Asignación de probable función a proteínas sin caracterización

La Tabla 7 muestra el resultado del alineamiento de los genes sin descripción, y por lo tanto sin función asignada, observados experimentalmente bajo condiciones de fermentación alcohólica contra la base de datos de proteínas no redundantes del NCBI. Cabe destacar que esta tabla muestra solamente los genes que produjeron tags de la clase platino y que poseen sobre un 70% de su secuencia alineada con alta identidad (sobre 40% de identidad) con alguna proteína de función conocida. Basándose en estos parámetros, se puede asignar con relativamente alta probabilidad una función a las proteínas bajo estudio. Esto debido a que proteínas con secuencias similares adoptan estructuras similares (Zuckerkyl & Pauling, 1965; Doolittle, 1981,1986; Chothia & Lesk, 1986), las que finalmente definen su función. Además, se ha determinado que proteínas con sobre 40% de identidad comparten su estructura terciaria en un alto grado, lo que finalmente se traduce en que comparten la misma función (Rost, 1999). Considerando los antecedentes anteriores, las asignaciones de funciones realizadas en este trabajo (Tabla 7) serían confiables.

3.5.- Transcritos poliadenilados en *Saccharomyces cerevisiae*

Debido a que la técnica de SAGE se basa en la extracción de tags desde transcritos poliadenilados, solamente los transcritos que poseen colas poli(A) se podrán evaluar mediante esta técnica. Actualmente se piensa que en levadura solamente los ARNs mensajeros y algunos ARNs ribosomales se poliadenilan luego de su transcripción (Kuai y col, 2004). En cambio lo que nos revelan las asignaciones de tags experimentales de levadura realizada por el método HGA, es que no solamente los ARNs mensajeros y los ARNs ribosomales se poliadenilan, sino que también otras clases de ARNs no codificantes (ncRNAs). Esta hipótesis se fundamenta en que en todos los experimentos de SAGE descritos en levadura (Velculescu y col, 1997; Kal y col, 1999; Varela y col, 2005) se observaron tags de la clase platino (alta confianza) mapeando en la primera posición de ncRNAs. Cabe mencionar que es altamente improbable que estos tags se hayan generado debido a errores de secuenciación o debido a errores durante la transcripción reversa ya que todos los tags mencionados anteriormente poseen frecuencias mayores a 5 en alguna de las condiciones evaluadas. También es altamente improbable que estos ncRNAs no sean poliadenilados y que igualmente hayan sido reclutados durante la captura de los ARNs con colas poli(A). Esto es debido a que no poseen subsecuencias internas mayores a 5 adeninas consecutivas – por lo que no podrían producir tags a consecuencia de transcripción reversa interna – y tampoco muestran altos valores de identidad con secuencias reversas-complementarias de genes codificantes de proteínas o ARNs ribosomales. Este último punto es importante considerarlo debido a que ARNs sin colas poli(A) podrían, en teoría, ser capturados al hibridarse con transcritos poliadenilados ligados a biotina. Otro punto importante es que, debido a que por lo menos algunos ncRNAs tendrían colas

poli(A), éstos deberían tener alguna señal de poliadenilación, señal que actualmente no se está considerando para la búsqueda *in silico* de estos genes (McCutcheon & Eddy 2003; Schattner y col, 2004). Como consecuencia de lo anterior, algunos de los SNRs estarían anotados con una longitud más corta de lo que realmente codifica el gen. Evidentemente esto es sólo una hipótesis y necesita ser demostrada experimentalmente.

3.6.- Descubrimiento de nuevos genes en *Saccharomyces cerevisiae*

En este trabajo se propone que se transcriben ocho regiones anotadas actualmente como intergénicas en el genoma de *Saccharomyces cerevisiae*. De estas ocho regiones, tres no poseen ningún gen anotado en ambas hebras, mientras que las cinco restantes corresponden a regiones intergénicas que tienen en su hebra opuesta un gen anotado. Las tres regiones que no tienen ningún gen anotado en ninguna de las dos hebras, y que se demostró que se transcriben (sección 2.9), no poseen más de un 30% de identidad con ninguna de las proteínas actualmente conocidas. Por lo tanto, estas regiones podrían corresponder a genes que codifican proteínas nuevas o ARNs no codificantes (Olivas y col, 1997). Para evaluar si estas regiones se traducen finalmente a proteínas existen múltiples estrategias, dentro de las cuales la espectrometría de masas representa una alternativa atractiva (Washburn y col, 2001; Oshiro y col, 2002). Con respecto a las regiones intergénicas que poseen un gen anotado en su hebra opuesta y que se demostraron que se transcriben, podrían corresponder a ARNs antisentido, transcritos que son abundantes y que tienen un rol importante durante la regulación génica en eucariontes (RIKEN y col, 2005). Finalmente, para corroborar que los

productos de PCR provienen desde las regiones predichas, entre los estudios que seguirán a este trabajo, está contemplado secuenciar los amplicones.

4.- CONCLUSIONES

En este trabajo se comprobaron las hipótesis planteadas en la sección 1.5. Se demostró que mediante la metodología HGA, la información genómica es de gran utilidad para la asignación de tags producidos por SAGE. Se obtuvo un incremento significativo en las asignaciones no ambiguas de tags experimentales, además de demostrar que gran cantidad de tags experimentales de SAGE provienen desde regiones actualmente anotadas como intergénicas, cADNs parcialmente digeridos, hebras opuestas de transcritos, intrones, límites exón-intrón y de ARNs no codificantes (Tabla 5).

También se comprobó que la utilización de información genómica para la asignación de tags experimentales de SAGE permite el descubrimiento de nuevos genes. En este trabajo se demostró que 8 regiones actualmente anotadas como intergénicas en el genoma de levadura, se transcriben bajo condiciones de fermentación alcohólica.

Además se desarrolló un servidor web basado en las metodologías creadas en este trabajo. Este servidor permitirá realizar la más completa asignación de tags experimentales de SAGE, además de facilitar el descubrimiento de nuevos genes y ayudar a entender la expresión génica a nivel cromosómico.

Es importante mencionar que la metodología inventada en este trabajo ha sido publicada en una revista de bioinformática (Malig y col, 2006) y se escribió un documento sobre el servidor web, el cual fue recientemente aceptado en la revista Nucleic Acids Research. Finalmente, el WiCell Research Institute planea utilizar la metodología HGA para mapear sus librerías de SAGE, uno de los proyectos más grandes y novedosos en el estudio de la expresión génica en células madres embrionarias humanas.

5.- MATERIALES Y METODOS

5.1.- Fuente de la secuencia genómica

La secuencia completa del genoma de *Saccharomyces cerevisiae* fue obtenida en Julio del año 2005, desde el sitio: Saccharomyces genome database (SGD) (ftp://ftp.yeastgenome.org/yeast/). Este archivo incluye los 16 cromosomas nucleares y el cromosoma mitocondrial. El archivo original que contiene estas secuencias se encuentra disponible como material suplementario en el sitio web http://dna.bio.puc.cl/HGA-yeast.html.

5.2.- Fuente de anotación genómica

Se usó la anotación genómica, versión del 26 de Julio del 2005, del sitio SGD (ftp://ftp.yeastgenome.org/yeast/). Este archivo fue filtrado de modo de sólo seleccionar los registros que contuvieran en el campo 'feature type' una de las siguientes palabras: ORF, intrón, rRNA, tRNA, snoRNA, snRNA y ncRNA. La tabla original que se ocupó en este trabajo para la anotación del genoma se encuentra disponible como material suplementario en el sitio web http://dna.bio.puc.cl/HGA-yeast.html.

5.3.- Construcción de un mapa de restricción genómico virtual para la extracción de una librería de tags de SAGE potenciales

Se escribieron varios programas computacionales, en lenguajes C++, ANSI C y Perl, para desarrollar tareas específicas. Primero, en el computador y utilizando un programa computacional llamado *subsequence,* la secuencia nucleotídica de cada

cromosoma fue fragmentada en todas las secuencias sobrelapadas posibles de un largo de 14 nucleótidos. Este proceso fue llevado a cabo para ambas hebras del ADN y los resultados se concatenaron en un sólo archivo. Segundo, las secuencias de 14 nts fueron filtradas, seleccionando sólo aquellas que tuvieran el patrón CATG (secuencia de reconocimiento de la enzima ancla, NlaIII) en su extremo 5'. Este procedimiento produjo un total de 76.516 tags, cantidad que representa el total de tags teóricos producidos por una digestión genómica completa, al utilizar la combinación de enzimas de restricción NlaIII y BsmFI. Durante este proceso, la posición y hebra donde cada tag se encontraba en el genoma fue registrada por medio de un nuevo programa computacional llamado *pattern*. Finalmente, se realizó una comparación pareada de todas las secuencias de tags y la frecuencia de ocurrencia de cada secuencia fue registrada con un software llamado *freqtag*. Todos estos programas computacionales se ejecutan en sistemas operativos LINUX y están disponibles como material suplementario en el sitio: http://dna.bio.puc.cl/HGA-yeast.html.

5.4.- Asignación de marcos abiertos de lectura y regiones no traducidas

Sólo los registros que contienen la palabra 'ORF', en el campo 'feature type' de la tabla de anotación genómica, fueron considerados para la anotación y predicción de los UTRs 3' y 5'. Esta restricción produjo un total de 6,591 ORFs candidatos para las asignaciones de los UTRs. Primero, los UTRs 3' fueron asignados a todos aquellos genes que tuvieran un UTR 3' predicho según el trabajo de Graber y colaboradores del año 2002. Este trabajo contiene un total de 3.141 predicciones, de las cuales 204 no fueron asignadas. Estas asignaciones fueron descartadas porque correspondían a pseudogenes, elementos transponibles o asignaciones

donde no se encontraba ningún ORF en las proximidades de la predicción. Por lo tanto, un total de 2.937 UTRs 3' fueron asignados basados en la predicción de Graber y colaboradores del año 2002. Los UTRs 3' para los restantes 3.654 ORFs fueron asignados siguiendo el siguiente procedimiento: 1) Un largo fijo de 370 nts fue asignado si la primera posición del siguiente ORF, anotado en el mismo cromosoma y hebra en la dirección 3', está a más de 370 nts de distancia desde el fin del ORF que se está evaluando. Se escogió este largo debido a que más del 95% de las longitudes de los UTRs 3' predichos en levadura es igual o menor que 370 nts (Figura 5). Un total de 2.752 UTRs 3' fueron asignados con un largo fijo de 370 nts. 2) Si el número total de nucleótidos disponibles entre el fin del ORF evaluado y la posición inicial del ORF río abajo es igual o menor que 370 nts, entonces la asignación del UTR 3' corresponde a la región entre el fin del ORF evaluado y la posición previa al inicio del ORF río abajo. 824 UTRs 3' fueron asignados de esta manera. 3) Para los ORFs donde su extremo terminal estuviera contenido dentro de otro ORF, no se asignó ningún UTR 3'. A un total de 78 ORFs no se les asignó un UTR 3'. Después de todas las asignaciones de los UTRs 3', las anotaciones de los UTRs 5' se realizó de la siguiente manera: debido a que la mayoría de los UTRs 5' de levadura son desconocidos, se asignó un largo fijo de 100 nts cuando existe una distancia superior a esta longitud entre el codón de inicio del ORF que se está evaluando y el fin del UTR 3' del ORF río arriba. En los casos donde el fin del UTR 3' del ORF río arriba se encuentra a una distancia menor de 100 nts, la posición del nucleótido siguiente al fin del UTR 3' fue considerada como la posición inicial del UTR 5'. Se escogió un largo fijo de 100 nts debido a que más del 95% de los tags experimentales que mapean en UTRs 5' de levadura se observaron a menos de 100 nts de distancia del codón de inicio del ORF inmediatamente río abajo (Zhang & Dietrich 2005). Un total de 5.112 UTRs 5' de largo 100 nts fueron asignados. Mientras que un total de 1.214 UTRs 5' fueron

asignados con un largo menor de 100 nts y mayor que cero. En los casos donde no existen nucleótidos disponibles para la asignación de un UTR 5' no se asignó ningún UTR 5'. Estos casos correspondieron a 265 ORFs.

5.5.- Asignación de los tags genómicos virtuales de SAGE a los transcritos anotados

Basados en sus posiciones genómicas, todos los tags genómicos virtuales de SAGE fueron mapeados contra los elementos anotados en el genoma. Todas las asignaciones de los tags con los transcritos fueron registradas, tanto las completas como las incompletas. Una asignación completa se obtiene cuando el tag virtual está totalmente contenido dentro del transcrito. En cambio, una asignación parcial fue definida sólo si la condición anterior no se cumple y si el nucleótido más 5' del tag estaba contenido dentro del transcrito; de otro modo el tag virtual es definido como intergénico. El mismo criterio descrito anteriormente fue utilizado para definir aquellos tags que mapean a intrones. Un total de 775 intrones están actualmente anotados dentro de genes conocidos dentro del genoma de levadura. Los tags virtuales que mapean en el límite entre un exón y un intrón fueron anotados como 'tags límite exón-intrón'. También se generaron splicing virtuales en el computador de modo de obtener los tags potenciales que se podrían obtener como consecuencia de este proceso. Sólo 13 nuevos tags potenciales, secuencias que no se encuentran en el genoma de levadura, se generaron como resultado del procedimiento anterior.

5.6.- Asignación de tags genómicos virtuales intergénicos de SAGE a hebras opuestas de transcritos

Todos los tags virtuales definidos como intergénicos en el paso anterior fueron inspeccionados por su ocurrencia en la hebra opuesta de un transcrito anotado. Todos los tags virtuales localizados dentro de una región intergénica pero contenidos completamente en la hebra opuesta de un transcrito anotado fueron clasificados como tal. Cabe mencionar que estos tags podrían ser importantes para el descubrimiento de nuevos elementos de ARN de interferencia (Shena, 1995; Quere, 2004).

5.7.- Asignación de tags genómicos virtuales de SAGE que están localizados río abajo de regiones poli-A dentro de transcritos

Los tags virtuales localizados dentro de un transcrito, distintos del tag más 3', fueron observados con una alta frecuencia experimental debido al apareamiento del primer oligo-d(T) a regiones de poli(A) internas durante la transcripción reversa (Nam, 2002). Por lo tanto, para todos los tags virtuales que mapean dentro de un transcrito, la actual posición dentro de éste desde el extremo 3' fue registrada. Luego, todas las regiones de 8 o más adeninas consecutivas que fueron encontradas dentro de cualquier transcrito anotado fueron grabadas. Finalmente, todos aquellos tags que mapearon dentro de un transcrito y que se encontraban en o sobre la segunda posición desde el extremo 3', donde una región de poli(A) interna estaba localizada río arriba y a una distancia menor de 800 nucleótidos, sin otro tag entre ellos, fueron definidos como 'tag cercano a poli(A)'.

5.8.- Construcción de una librería de tags experimentales producidos por SAGE en levadura

Se compiló toda la información experimental disponible de experimentos de SAGE en levadura. Esta información incluye tres trabajos independientes (Velculescu, 1997; Kal, 1999; Varela, 2005), publicaciones que totalizan los 8 puntos experimentales distintos que se ocuparon en este trabajo (Tabla 2).

5.9.- Asignación de los tags experimentales de SAGE a la librería de tags virtuales de SAGE

El proceso de asignación de tags experimentales de SAGE a las tags virtuales fue realizada al asignar a el tag experimental la anotación del tag virtual con alta confianza, cuando esto es posible, de otro modo el tag experimental fue asignado a múltiples regiones del genoma con una confianza indefinida.

5.10.- Construcción del servidor y base de datos SAGExplore

El servidor SAGExplorer se construyó en lenguaje HTML y PHP, mientras que el manejo de datos desde las bases de datos se lleva a cabo en lenguaje MySQL. Finalmente las aplicaciones gráficas se programaron en JpGraph.

5.11.- Fermentaciones vínicas

5.11.1.- Cepa de levadura y medio de cultivo

Durante las fermentaciones vínicas se utilizó la cepa comercial de *Saccharomyces cerevisiae* EC1118, la cual es ampliamente utilizada en fermentaciones alcohólicas. Los cultivos iniciales fueron crecidos en medio YPD a 28ºC bajo condiciones aeróbicas. Se utilizó en las fermentaciones realizadas en bioreactores el mosto artificial MS300, el cual simula un jugo de uva estándar. Sin embrago, el medio MS300 fue modificado al aumentar la concentración total de azúcar a 240 g/l e incluyendo partes iguales de glucosa y fructosa (Varela, 2004). El contenido de nitrógeno asimilable fue 300 mg N/l suministrado en forma de amonio y aminoácidos.

5.11.2.- Condiciones de cultivo

Un bioreactor de 20 litros (Bioengineering, Wald, Switzerland) fué inoculado con una densidad inicial de 1 x 10^6 células/ml. Las células fueron lavadas con NaCl al 0,9% para eliminar cualquier remanente de nitrógeno que pudiera quedar del medio anterior a la inoculación. La temperatura fue mantenida a 28ºC y el pH a 3,5. Se introdujo nitrógeno al medio por 30 minutos antes de la inoculación para eliminar el oxigeno del medio. La agitación se mantuvo a 100 rpm.

5.11.3.- Técnicas analíticas

La evolución de dióxido de carbono en los bioreactores fue determinada con un transductor de flujo volumétrico. Muestras de cultivo fueron tomadas

periódicamente para establecer el estatus de la fermentación, de modo de compararlas con los valores obtenidos por la fermentación de Varela (Varela, 2005). Estas muestras fueron analizadas para determinar el peso celular seco, el número de células y las concentraciones de glucosa, fructosa. El peso celular seco fue estimado al filtrar las células y lavarlas dos veces con agua destilada, para luego secarlas con un peso constante a 85°C. El número de células fue estimado microscópicamente usando una cámara Neubauer (Brand, Wertheim, Alemania). Las concentraciones de glucosa, fructosa y etanol fueron medidas por cromatografía líquida de alto desempeño (HPLC).

Es importante destacar que luego de todas las mediciones, los valores obtenidos en este trabajo y los obtenidos por Varela, 2005 son casi idénticos (resultados no mostrados).

5.12.- RT-PCR

El transcriptoma de levadura fue analizado durante la fase estacionaria tardía del proceso de fermentación alcohólica (Varela, 2005), ya que es en ésta fase donde los tags intergénicos de interés muestran mayor número de observaciones. El ARN total fue extraído desde las muestras usando el reactivo Trizol (Invitrogen, Carlsbad, CA, USA) según las instrucciones del fabricante. Este RNA fue utilizado para sintetizar ADN de doble hebra con el kit ImProm-II Reverse Transcription System (Promega, Madison, WI, USA), siguiendo el protocolo del proveedor usando hexaoligonucleotidos de secuencia aleatoria en el caso del RT-PCR tradicional o partidores secuencia-específicos en el caso del RT-PCR específico de hebra. Adicionalmente, un control negativo se realizó para cada muestra al sustituir la transcriptasa reversa por agua durante el protocolo de la transcripción reversa.

Para la amplificación por PCR se utilizó dos microlitros de cada reacción de transcripción reversa como templado, amortiguador de PCR 1X; 0,2 uM de cada partidor (Tabla 10); 0,2 mM de dNTPs; 1,5 mM de MgCl2 y 1 U de *Taq* polimerasa en un volumen final de reacción de 20 uL.

Las condiciones de PCR fueron las siguientes: denaturación inicial a 94°C por 5 minutos; seguido por 30 ciclos de: denaturación a 94°C por 1 minuto; annealing a 58°C por 40 segundos; extensión a 70°C por 50 segundos. La elongación final fue realizada a 72°C por 5 minutos. Los productos de PCR fueron resueltos en geles de 1,5% de agarosa y visualizados al teñirlos con bromuro de etidio.

Tabla 10: Partidores utilizados en este trabajo.

ID[1]	Tag estudiado	Partidor sentido[2]	Partidor antisentido[3]	Tamaño[4]
I1	CATGAGGCTACCTA	5'-GTTTTCTTCGCCCGCTCCGTG-3'	5'-TAGCTACCACGCGTATTC-3'	183
I2	CATGTAGTTGCTCC	5'-CTGAGTATTCCCACAGTC-3'	5'-GGTATCGGTCAGTATCAC-3'	154
I3	CATGAGAGGTGATC	5'-CTTGCGGTAGTACGTGTG-3'	5'-GTAAGCTCGCTACGTGAC-3'	115
I4	CATGGAATTTATAG	5'-ATGTGTCGCCATTCACTACC-3'	5'-ATCACCATCGACCGACTAGA-3'	144
I5	CATGGCGACTTGAT	5'-AGTGCCCAAACACGGGAT-3'	5'-TCGCCATGTCCGTTCTCT-3'	173
O1	CATGGCCAATGATA	5'-TGTGTTGCAGTAGCAGTC-3'	5'-CGCTAAGCTCCTTCTATC-3'*	186
O2	CATGCCCACGTAAG	5'-TCTCAATCCGCTCTTGTTCC-3'	5'-TGACTCCGAAGGCATTACAG-3'*	212
O3	CATGGGTAATCGAA	5'-GGAGACCTCTCGCACGTATG-3'	5'-GTCACCACGGCGAACTGGAT-3'*	198
O4	CATGGCCAGGACAA	5'-GTCGTCATTCTCACCAGTAG-3'	5'-CACGTCAAGGTTAGCAAGTC-3'*	147
O5	CATGATAATGAGGA	5'-AACCTCAGTGGCAGTCTT-3'	5'-GTTCCACCTTCTCCTCAT-3'*	175

[1] Nombre de identificación del tag.
[2] Partidores que se aparean en regiones río arriba del tag estudiado.
[3] Partidores que se aparean en regiones río abajo del tag estudiado.
[4] Tamaño predicho del producto de PCR.
* Primers utilizados durante la síntesis de DNA de doble hebra por la transcripción reversa específica de hebra.

6.- BIBLIOGRAFIA

Adams, M.D., S.E. Celniker, R.A. Holt, C.A. Evans, J.D. Gocayne, P.G. Amanatides, S.E. Scherer, P.W. Li, R.A. Hoskins, R.F. Galle, R.A. George, S.E. Lewis, y S. Richards. 2000. The genome sequence of Drosophila melanogaster. Science **287**: 2185-2195.

Akmaev, V.R. y Wang, C.J. 2004. Correction of sequence-based artefacts in serial analysis of gene expression. Bioinformatics **20**: 1254-1263.

Bauer, F., Pretorius, I. 2000. Yeast stress response y fermentation efficiency: how to survive the making wine – a review. Afr. J. Enol. Viticult. **21**: 27-51.

Boheler, K.R. y M.D. Stern. 2003. The new role of SAGE in gene discovery. Trends Biotechnol. **21**: 55-57.

Burkhard R. Twilight zone of protein sequence alignments. 1999. Protein Engineering. **12(2)**: 85–94.

Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., Hermes, M., van Asperen, R., Boon, K., Voute, P., Heisterkamp, S., van Kampen, A y Versteeg, R. 2001. The human transcriptome map: Clustering of highly expressed genes in chromosomal domains. Science. **291**: 1289-1292.

Chen, J., M. Sun, S. Lee, G. Zhou, J.D. Rowley, y S.M. Wang. 2002. Identifying Novel Transcripts y Novel Genes in the Human Genome by Using Novel SAGE Tags. Proc. Natl. Acad. Sci. U.S.A. **99**: 12257-12262.

Chothia,C. y Lesk,A.M. 1986 EMBO J. **5**: 823–826.

Claverie J.M. 2005. Fewer Genes, More Noncoding RNA. Science 309: 1529-1530

Divina P, Forejt J. 2004. The Mouse SAGE Site: database of public mouse SAGE libraries Nucleic Acids Res. **32**(1): D482-3.

Doolittle,R.F. 1981. Science **214**: 149–159.

Doolittle,R.F. 1986. Of URFs y ORFs: a primer on how to analyze derived amino acid sequences. University Science Books, Mill Valley, CA, USA.

Graber, J.H., McAllister, G.D. y Smith, T.F. 2002. Probabilistic prediction of Saccharomyces cerevisiae mRNA 3'-processing sites. Nucleic Acids Res. **30**: 1851-1858.

Harbers, M. y P. Carninci. 2005. Tag-based approaches for transcritome research y genome annotation. Nat. Methods **7**: 495-502.

Havilio, M., E.Y. Levanon, G. Lerman, M. Kupiec, y E. Eisenberg. 2005. Evidence for abundant transcription of non-coding regions in the Saccharomyces cerevisiae genome. BMC Genomics **6**: 93-101.

Kal, A.J., A.J. van Zonneveld, V. Benes, M. van den Berg, M.G. Koerkamp, K. Albermann, N. Strack, J.M. Ruijter, A. Richter, B. Dujon, y col. 1999. Dynamics of gene expression revealed by comparison of seriado analysis of gene expression transcript profiles from yeast grown on two different carbon sources. Mol. Biol. Cell **10**: 1859-1872.

Kessler, M., Zeng, Q., Hogan, S., Cook, R., Morales, A. y Cottarel, G. 2003. Systematic Discovery of New Genes in the Saccharomyces cerevisiae Genome. Genome Res. **13**: 264-271

Kuai, L., F. Fang, Butler J.S, y F. Sherman. 2004. Polyadenylation of rRNA in Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. USA **101**: 8581-8586.

Lash, A.E., Tolstoshev, C.M., Wagner, L., Schuler, G.D., Strausberg, R.L., Riggins, G.J. y Altschul, S.F. 2000. SAGEmap: a public gene expression resource. Genome Res. **10**: 1051-1060.

Lee, S., Clark, T., Chen, J., Zhou, G., Scott, L.R., Rowley, J.D. y Wang, S.M. 2002. Correct identification of genes from serial analysis of gene expression tag sequences. Genomics **79**: 598-602.

Malig, R., Varela, C., Agosin, E. y Melo, F. 2006. Accurate and unambiguous tag-to-gene mapping in serial analysis of gene expresión. BMC Bioinformatics. 2006; 7: 487-504.

McCutcheon, J.P. y Eddy, S.R. 2003. Computational identification of non-coding RNAs in Saccharomyces cerevisiae by comparative genomics. Nucleic Acids Res. **31(14)**: 4119-28.

Nam, D., S. Lee, G. Zhou, X. Cao, C. Wang, T. Clark, J. Chen, Rowley J, y M. Wang. 2002. Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly-A priming during reverse transcription. Proc. Natl. Acad. Sci. USA **99**: 6152-6156.

Olivas, W., Muhlrad, D. y Parker, R. 1997. Analysis of the yeast genome: identification of new non-coding and small ORF-containing RNAs. Nucleic Acids Res. **25**: 4619-4625.

Oshiro, G., Wodicka, L., Washburn, M., Yates, J., Lockhart, D. y Winzeler E. 2002. Parallel identification of new genes in *Saccharomyces cerevisiae*. Genome Res. **12**: 1210-1220.

Peng, W.T., Robinson, M.D., Mnaimneh, S., Krogan, N.J., Cagney, G., Morris, Q., Davierwala, A.P., Grigull, J., Yang, X., Zhang, W., Mitsakakis, N., Ryan, O.W., Datta, N., Jojic, V., Pal, C., Canadien, V., Richards, D., Beattie, B., Wu, L.F., Altschuler, S.J., Roweis, S., Frey, B.J., Emili, A., Greenblatt, J.F., Hughes, T.R. 2003. A panoramic view of yeast noncoding RNA processing. Cell **113(7)**: 919-33.

Quere, R., L. Manchon, M. Lejeune, O. Clement, F. Pierrat, B. Bonafoux, T. Commes, D. Piquemal, y J. Marti. 2004. Mining SAGE data allows large-scale, sensitive screening of antisense transcript expression. *Nucleic Acids Res.* **32**: E163.

RIKEN Genome Exploration Research Group, Genome Science Group (Genome Network Project Core Group) y el Consorcio FANTOM y col. 2005. Antisense Transcription in the Mammalian Transcriptome. Science **309**: 1564-1566.

Rossignol, T., Dulau, L., Julián, A. y Blondin B. 2003. Genome-wide monitoring of wine yeast gene expresión during alcoholic fermentation. Yeast **20**: 1369-1385.

Saha, S., Sparks, A., Rago, C., Akmaev, V., Wang, C., Vogelstein, B., Kinzler, K., Velculescu, V. 2002. Using the transcriptome to annotate the genome. Nature Biotechnol. **20**: 508-512.

Schattner P, et al. 2004. Genome-wide searching for pseudouridylation guide snoRNAs: analysis of the Saccharomyces cerevisiae genome. Nucleic Acids Res. **32(14)**: 4281-96.

Schuler, G.D. 1997. Pieces of the puzzle: Expressed sequences tags y the catalog of human genes. J. Mol. Med. **75**: 694-698.

Shena, M., Shalon, D., Davis, R. y Brown, P. 1995. Quantitative monitoring of gene expression patters with a complementary DNA microarray [see comments]. Science **270**: 467-470.

Stolc, V., Samanta, M.P., Tongprasit, W., Sethi, H., Liang, S., Nelson, D.C., Hegeman, A., Nelson, C., Rancour, D., Bednarek, S., Ulrich, E.L., Zhao, Q., Wrobel, R.L., Newman, C.S., Fox, B.G., Phillips, G.N. Jr., Markley, J.L., Sussman, M.R. 2005. Identification of transcribed sequences in Arabidopsis thaliana by using high-resolution genome tiling arrays. Proc Natl Acad Sci U S A. **102(12)**: 4453-8.

Sun, M., G. Zhou, S. Lee, J. Chen, R.Z. Shi, y S.M. Wang. 2004. SAGE is far More Sensitive than EST for Detecting low-abundance Transcripts. BMC Genomics **5**: 1-4.

Tuteja R, Tuteja N. 2004. Serial Analysis of Gene Expression: Applications in Human Studies. J. Biomed. Biotechnol. **2**: 113-120.

Tuteja R, Tuteja N. 2004 Serial analysis of gene expression: unraveling the bioinformatics tools. BioEssays. **26**: 916-922.

van Kampen, A.H., van Schaik, B.D., Pauws, E., Michiels, E.M., Ruijter, J.M., Caron, H.N., Versteeg, R., Heisterkamp, S.H., Leunissen, J.A., Baas, F. et al. 2000. USAGE: a web-based approach towards the analysis of SAGE data. Serial Analysis of Gene Expression. Bioinformatics. **16**:899-905.

Varela, C., Pizarro, F. y Agosin, E. 2004. Biomass content governs fermentation rate in nitrogen-deficient wine must. Appl. Environ. Microbiol. 70: 3392-3400.

Varela, C., J. Cardenas, F. Melo, y E. Agosin. 2005. Quantitative analysis of wine yeast gene expression profiles under winemaking conditions. Yeast **22**: 369-383.

Velculescu, V.E., Zhang, L., Vogelstein, B. y Kinzler, K.W. 1995. Serial analysis of gene expression. Science **270**: 484-487.

Velculescu, V.E., L. Zhang, W. Zhou, J. Vogelstein, M.A. Basrai, D.E. Bassett Jr, P. Hieter, B. Vogelstein, y K.W. Kinzler. 1997. Characterization of the yeast transcritome. Cell **88**: 243-251.

Wahl, M., Heinzmann, U., y Imai, K. 2004. LongSAGE analysis significantly improves genome annotation: identifications of novel genes y alternative transcripts in the mouse. Bioinformatics **21**: 1393-1400.

Walter G. 1998. Yeast Physiology y Biotechnology. Wiley: New York.

Washburn, M.P., Wolters, D. y Yates, 3rd, J.R. 2001. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat. Biotechnol. **19**: 242-247.

Zhang, Z. y F.S. Dietrich. 2005. Mapping of transcription start sites in Saccharomyces cerevisiae using 5' SAGE. Nucleic Acids Res. **33**: 2838-2851.

Zuckerkyl,E. y Pauling,L. 1965. Evolutionary Divergence y Convergence in Proteins. Academic Press, New York; London. 97–166.