Universidad de Chile

Facultad de Ciencias Químicas y Farmacéuticas

Departamento de Bioquímica y Biología Molecular

USO DE BIOINFORMÁTICA COMO HERRAMIENTA DE ANÁLISIS DE DATOS EN CÁNCER GÁSTRICO

Memoria para optar al título de Bioquímico

Francisco Javier Ossandón Cabrera

Profesor patrocinante:

Dr. Javier Puente Piccardo

Directores de memoria:

Dr. Alejandro Corvalán Rodríguez

Dr. Francisco Melo Ledermann

SANTIAGO DE CHILE

2006

DEDICATORIA

A toda mi familia, cuyo apoyo jamás me ha faltado a pesar de todas mis faltas y errores.

AGRADECIMIENTOS

Siendo esta tesis el paso final de un camino que ha sido largo y sufrido, es mucho lo que le debo a mucha gente.

Quiero agradecer muy especialmente a mi tutor Alejandro Corvalán, quien ha tenido mucha paciencia y fe en mí, alentándome constantemente a seguir adelante. También agradezco a Concepción Risueño por aceptarme como alumno en práctica y por su buena disposición a enseñarme lo necesario.

Agradezco a mi familia, tíos y primos, que me han apoyado incondicionalmente por muchos años durante mi accidentada carrera. A mi madre y a mi padre, que se han esforzado para que nada me falte, a mis hermanos y hermanas que tantas alegrías me dan. En especial a mi hermana Paula, cuyo constante aliento y preocupación es un estímulo para mí.

También quiero agradecer a mis muy queridos amigos con los cuales he compartido alegrías y pesares durante estos años. En especial a Claudio Toledo, Víctor Piña, Juan Carlos Rivera, Pablo Martínez, Alberto Abudinen, Marcelo Fortín, Óscar Besser y Rafael Vergara.

Finalmente quiero recordar a mis queridos abuelos Jorge y Flora, que ya descansan en paz. Mucho me hubiera gustado que estuvieran aquí para ver a su nieto, el “macaco”, convertirse en profesional al fin.

ABREVIATURAS

ADN : Ácido Desoxirribonucleico

CG : Cáncer Gástrico

CGAP : *Cancer Genome Anatomy Project*

COA : *Correspondence Analysis*

CpG : Dinucleótidos de Citosina-Fosfato-Guanina

EBV : Virus de Epstein-Barr

FDR : *False Discovery Rate*

FONDECYT : Fondo Nacional de Desarrollo Científico y Tecnológico

GEO : *Gene Expression Omnibus*

GST : Gen Supresor de Tumores

IM : Índice de Metilación

MS-PCR : *Methylation Specific-Polymerase Chain Reaction*

pb : Par de bases

SAGE : *Serial Analysis of Gene Expression*

SAM : *Significance Analysis of Microarrays*

ÍNDICE GENERAL

DEDICATORIA II
AGRADECIMIENTOS III
ABREVIATURAS IV
ÍNDICE GENERAL V
ÍNDICE DE FIGURAS VII
ÍNDICE DE TABLAS VIII
RESUMEN IX
SUMMARY X
I. INTRODUCCIÓN 1
1.1. Cáncer gástrico 2
1.2. Bioinformática 4
1.2.1. Agrupación jerárquica 4
1.2.2. Análisis de correspondencia 11
1.2.3. Análisis de significancia de microarreglos 12
1.3. Hipótesis 14
1.4. Objetivos 15
II. MATERIALES Y MÉTODOS 16
2.1. Bioinformática en estudio de metilación de cáncer gástrico 16
2.1.1 Casos clínicos 16
2.1.2. Análisis estadístico y bioinformático 17
2.2. Análisis bioinformático en SAGE de estómago 19

III. RESULTADOS Y DISCUSIÓN....23
3.1. Bioinformática en estudio de metilación de cáncer gástrico....23
3.1.1. Grupo de muestras en tejido fresco....24
3.1.2. Grupo de muestras en parafina....26
3.1.3. Discusión del estudio de metilación....31
3.2. Análisis bioinformático en SAGE de estómago....33
3.2.1. Patrones de expresión....34
3.2.2. Diferencias de expresión....38
3.2.3. Discusión del estudio de SAGE....41
IV. CONCLUSIONES....44

ÍNDICE DE FIGURAS

Figura 1. Comparación de los resultados de la métrica Euclidiana y la métrica de correlación de Pearson 6
Figura 2. Comparación de las variantes de Agrupación Jerárquica.....8
Figura 3. Agrupación Jerárquica para explorar correlaciones entre los casos y los perfiles de hipermetilación de genes supresores de tumores 26
Figura 4. Análisis de Agrupación Jerárquica en 83 casos de cáncer gástrico difuso.....29
Figura 5. Análisis de sobrevida Kaplan-Meier para comparación de metilación de APC y p73 en cáncer gástrico difuso.....30
Figura 6. *Support Cluster* (correlación de Pearson y *Average linkage*).....35
Figura 7. Análisis de Correspondencia (COA).....37
Figura 8. SAM para el grupo de librerías Normales vs. Tumorales 38
Figura 9. SAM para el grupo de tumores de EEUU vs. Japón 40

ÍNDICE DE TABLAS

Tabla 1. Comparación de características clínico-patológicas de casos de tejido fresco y tejido incluido en parafina utilizados en este estudio.....16
Tabla 2. Detalle de las librerías de SAGE.....19
Tabla 3. Genes supresores de tumores provenientes de literatura e inactivados en cáncer gástrico.....24
Tabla 4. Análisis de metilación en cáncer gástrico difuso en muestras de tejido fresco 24
Tabla 5. Índice de Metilación (IM) y correlaciones clínico-patológicas en cáncer gástrico difuso.....27
Tabla 6. Análisis de metilación en cáncer gástrico difuso en muestras de tejido incluido en parafina.....28
Tabla 7. Diez de los genes más significativos obtenidos por SAM entre muestras normales y tumorales.....39
Tabla 8. Diez de los genes más significativamente distintos obtenidos por SAM entre tumores de EEUU y Japón.....41

RESUMEN

El avance de la informática y el poder de proceso de los computadores han posibilitado la aparición de una disciplina emergente, la Bioinformática; cuyos objetivos son una visión amplia de los procesos biológicos y el entendimiento de grupos complejos de datos. En esta tesis, el uso de herramientas bioinformáticas en datos de hipermetilación de promotores de genes en cáncer gástrico difuso reveló asociaciones con características clínicas y 2 genes que se asocian a mal pronóstico (APC y p73). Por otro lado, el análisis de librerías de SAGE de estómago con estas herramientas mostró genes diferencialmente expresados en adenocarcinomas gástricos y diferencias étnicas en la transcriptómica de estas células neoplásicas.

SUMMARY

The progress of the informatics and the process power of the computers have made possible the raising of a new discipline, Bioinformatics; whose goals are a wider vision of the biological processes and the understanding of complex sets of data. In this thesis, the usage of bioinformatics tools in gene's promoter hipermethylation data in diffuse gastric cancer reveals associations to clinical features and 2 genes which are associated to bad outcome (APC and p73). Another analysis on stomach SAGE libraries with these tools shows differentially expressed genes in gastric adenocarcinomas and ethnic differences in the neoplasic cell transcriptomes.

I. INTRODUCCIÓN

Esta tesis tiene como propósito desarrollar el uso de herramientas bioinformáticas para la evaluación de perfiles moleculares en cáncer gástrico y se inserta en el contexto del proyecto FONDECYT 1030130.

Esta memoria se encuentra dividida en dos partes.

La primera parte consiste en un análisis bioinformático realizado en base al análisis del estado de metilación de 25 genes supresores de tumores en 104 muestras de cáncer gástrico. Estos datos fueron generados en el laboratorio en el contexto del proyecto FONDECYT 1030130 "*Patrón de Metilación de Genes Supresores de Tumores en la patogénesis del Cáncer Gástrico Difuso*".

La segunda parte corresponde a un análisis bioinformático realizado sobre datos de transcriptómica de cáncer gástrico, disponibles en sitios de dominio público. Los datos analizados provienen de una técnica novedosa llamada Análisis Seriado de la Expresión de Genes ("*Serial Analysis of Gene Expression*" o **SAGE**), una técnica costosa y cuya información generada es aún escasa, pero que aún así ha proporcionado resultados provechosos para el estudio del cáncer.

1.1. Cáncer gástrico

El cáncer gástrico (CG) representa la primera causa de muerte por enfermedades neoplásicas en nuestro país [1] y es la segunda causa de muerte por cáncer en el mundo [2]. En Latinoamérica, Chile ocupa el segundo lugar en mortalidad por CG después de Costa Rica, sumando entre ambos países casi el 50% de la mortalidad en este continente. En nuestro país ha ocurrido una constante declinación en la mortalidad por CG durante varias décadas, sin embargo ésta se detuvo en la década del ‘90 estabilizándose entre 19 y 20 muertes por cada 100.000 habitantes [3].

Los CG detectados son, en más del 90%, clasificados en Adenocarcinomas de tipo Intestinal o de tipo Difuso, según su tipo de diferenciación. El CG de tipo intestinal es un tumor bien diferenciado histológicamente, con incidencia asociada en su mayoría a gastritis atrófica severa. El CG de tipo difuso es histológicamente indiferenciado y se caracteriza por crecimiento infiltrativo e invasión peritoneal [4]. Dado que el CG difuso carece de lesiones precursoras a nivel morfológico, los métodos convencionales no son suficientes para su detección temprana, lo que redunda en la necesidad de métodos moleculares para su diagnóstico.

Desde mediados de la década de los ’80, el conocimiento de la activación de oncogenes e inactivación de genes supresores de tumores (GST) en el proceso neoplásico ha permitido proponer un modelo genético del cáncer. Este modelo plantea que la inactivación de GST estaría asociada a las primeras etapas del proceso neoplásico y que la activación de oncogenes estaría asociada a las etapas más avanzadas [5]. Así la identificación de los GST responsables del CG tendría un impacto importante en el diagnóstico precoz, ya que podrían utilizarse como marcadores de seguimiento en sujetos en riesgo de desarrollar CG.

Los GST pueden ser inactivados al menos por tres mecanismos: deleción, mutación puntual y metilación de ambos alelos del gen. La deleción consiste en la pérdida de un segmento de ADN, desde unos pocos pares de bases hasta cientos de kilobases,

afectando desde un segmento de un gen hasta un grupo de genes contiguos. La mutación puntual consiste en el cambio de entre 1 y 20 nucleótidos en la secuencia de un gen y en CG, de acuerdo a la hipótesis de "dos eventos" (*two-hits*) [6], ocurriría en forma complementaria a la deleción.

Recientemente, se ha reconocido a la metilación del ADN como un nuevo mecanismo de inactivación de GST y genéricamente se le ha denominado "hipermetilación de las regiones promotoras de genes" [7]. Este mecanismo se basa en la incorporación de un grupo metilo en el carbono 5 de la citosina (5-metil-citosina) cuando forma parte del dinucleótido CpG. Dado que los dinucleótidos CpG están concentrados a lo largo del genoma en regiones denominadas "islotes CpG" [8], y que la mayor parte de ellos se encuentran en el extremo 5´ de los genes, generalmente en las regiones promotoras o de inicio de transcripción, se ha sugerido que la hipermetilación de secuencias CpG sería un mecanismo de regulación de la expresión génica [7]. En cáncer, un creciente número de evidencias reconocen el papel de la hipermetilación de regiones promotoras como mecanismo de inactivación de GST [9].

En una revisión de la literatura se han identificado más de 30 GST asociados a CG. Estos GST están involucrados en múltiples funciones biológicas y en varios se ha descrito más de un mecanismo de inactivación, incluyendo la metilación. Sin embargo, en muy pocos de ellos existe información sobre su papel en lesiones precursoras, correlaciones clínicos-patológicas y de sobrevida en CG. Debido a los antecedentes presentados en esta tesis, se analiza el estado de metilación de algunos de estos GST específicos para definir correlaciones clínico-patológicas y de sobrevida.

Dado que la investigación gen a gen no provee suficiente información sobre las bases moleculares de CG difuso, se han creado nuevos métodos y tecnologías para el análisis de grupos complejos de datos. Estos análisis permiten ordenar una gran cantidad de información y se han vuelto cada vez más importantes en áreas de investigación biológica y biomédica.

1.2. Bioinformática

El análisis de grupos complejos de datos, en particular de datos genéticos, requiere de herramientas bioinformáticas. Algunas de las herramientas bioinformáticas actuales son la Agrupación Jerárquica (*Hierarchical Clustering*), el Análisis de Correspondencia (*Correspondence Analysis*) y el Análisis de Significancia de Microarreglos (*Significance Analysis of Microarrays*).

1.2.1. Agrupación jerárquica

La técnica de Agrupación Jerárquica, o "*Hierarchical Clustering*", provee una visión amplia de los patrones del comportamiento de todo el conjunto y ha sido utilizada en distintos ámbitos investigativos para descubrir familias de elementos con características similares. En 1998, Eisen y col. aplicaron esta técnica para el análisis de microarreglos con excelentes resultados [10].

Para entender la función de los genes como conjunto, se busca la agrupación de genes que posean un perfil de expresión similar ante diversas condiciones. Idealmente se busca que los grupos (*clusters*) encontrados sean homogéneos (baja variabilidad intra-grupo) y bien separados (alta variabilidad inter-grupos). Se define pues la correlación como medida de co-expresión. La matriz de expresión es una representación de datos de múltiples experimentos, que incluyen miles de variables simultáneas (genes).

Esta técnica se usa como método de descubrimiento de agrupaciones o asociaciones cuando no existe un conocimiento previo al respecto, y por tanto es una técnica no supervisada (que no está sujeta a una idea preconcebida).

Un valor que representa la distancia entre 2 genes o experimentos es computado al sumar las distancias entre sus respectivos vectores. Cómo este valor es normalizado o cómo la distancia es computada depende de la medida de distancia (métrica) utilizada. Hay una multitud de algoritmos disponibles, tanto coeficientes de correlación lineales,

como correlaciones no-paramétricas, o coeficientes de correlación por rangos. Todos estos algoritmos pueden ser divididos en 2 tipos.

1. **Procedimientos Aglomerativos**: Este procedimiento comienza con *n* grupos, donde cada objeto forma un grupo en sí mismo, e iterativamente se reduce el número de grupos mediante la fusión de los 2 objetos o grupos más similares en cada paso, hasta que sólo queda 1 grupo. $(n \rightarrow 1)$.
2. **Procedimientos Divisivos**: Este procedimiento empieza con 1 grupo e iterativamente divide un grupo por vez, de modo de que se reduzca la heterogeneidad. $(1 \rightarrow n)$.

1.2.1.1. Métricas

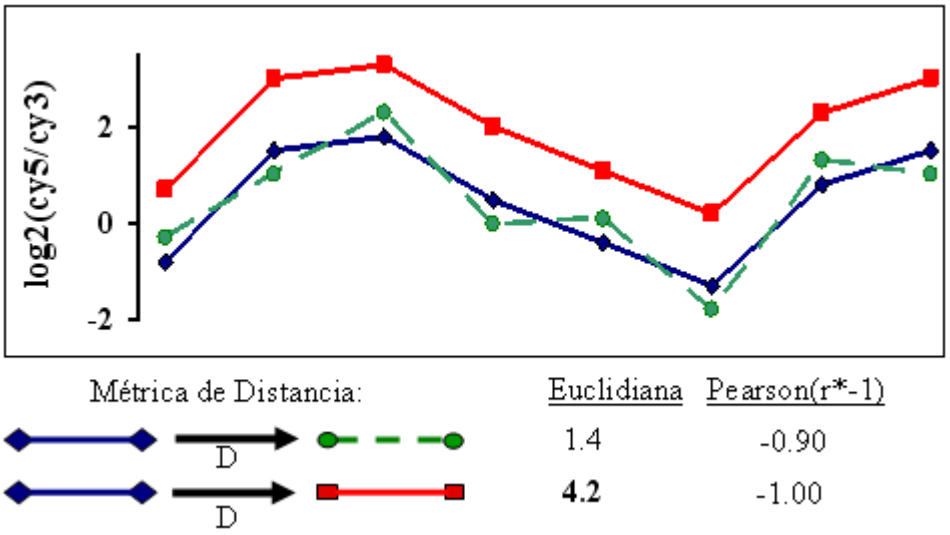
La manera de encontrar estas asociaciones se basa en el uso de una métrica como medida de similitud entre un patrón de expresión y otro, la que juega un rol primordial en las familias que se forman. Las métricas se basan en 3 elementos básicos:

1. *La distancia entre genes individuales.*
2. *La distancia entre los grupos.*
3. *La regla de detención.*

Los grupos son construidos iterativamente uniendo los dos genes o grupos más "cercanos" en cada paso. El resultado es un árbol de familias (*dendrograma*) donde cada grupo se acomoda junto a otro grupo similar. Para ejemplificar se presentan dos de las métricas más usadas, el Coeficiente de Correlación de Pearson y la distancia Euclidiana.

Un ejemplo de la diferencia de estas 2 métricas puede verse en la Figura 1.

Figura 1. **Comparación de los resultados de la métrica Euclidiana y la métrica de correlación de Pearson.** Se compara un mismo gen con otros 2 usando ambas métricas. En el primer caso los genes se reportan como parecidos por ambas métricas. En el segundo caso la métrica Euclidiana reporta que son mucho más distintos, mientras que la Correlación de Pearson reporta que son idénticos. Para la métrica Euclidiana, la distancia de las curvas es más importante, para Pearson la forma de las curvas es más importante.



1.2.1.1.1. Coeficiente de correlación de Pearson

El coeficiente lineal o de correlación de Pearson es la métrica más usada de asociación entre 2 vectores. Si **x** e **y** son vectores de *n*-componentes para los cuales se quiere calcular el grado de asociación, para pares de cantidades (x_i,y_i), $i=1,\ldots,n$ el coeficiente de correlación lineal *r* está dado por la fórmula:

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

Donde $\bar{x}$ es igual a la media del vector x, e $\bar{y}$ es igual a la media del vector y. Los valores *r* van entre -1 y 1, inclusive, donde 1 significa que 2 series son muy similares, 0 significa que son completamente independientes, y -1 significa que son opuestos. El

coeficiente de correlación es invariable ante la transformación escalar de los datos (adición, resta o multiplicación de los vectores por un factor constante).

Esto significa que el uso del coeficiente de correlación de Pearson tiene el mismo efecto que el uso de la distancia Euclidiana en datos estandarizados (centrados en la media, varianza unitaria).

1.2.1.1.2. Distancia euclidiana

La distancia Euclidiana es una medida de distancia muy usada.

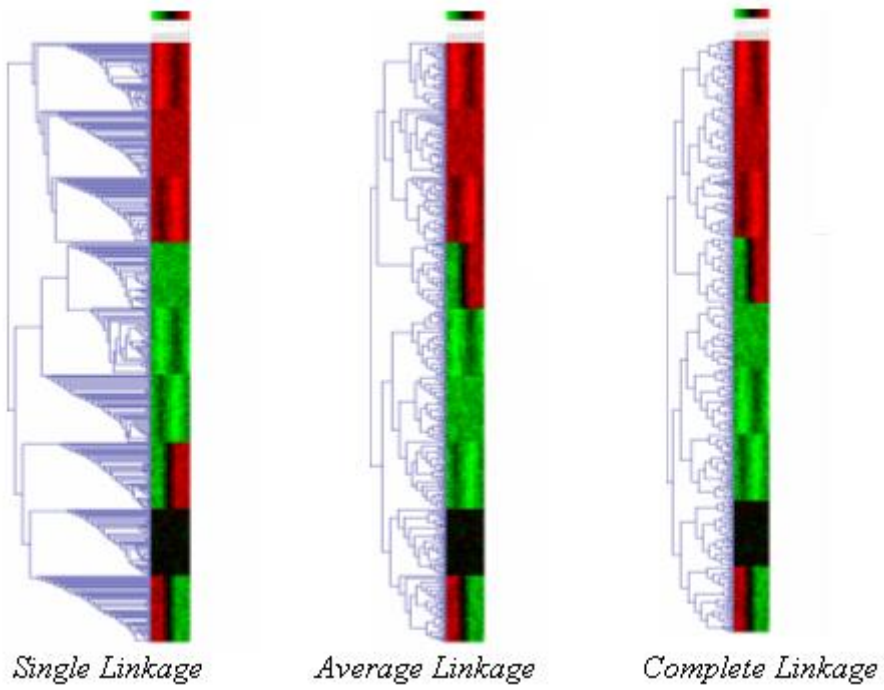
$$d_E = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

La distancia Euclidiana es sensible a la adición y multiplicación de todos los vectores por un factor constante. También es sensible a la dimensión de los valores de los vectores, por ejemplo si valores faltantes reducen la dimensión de ciertos vectores.

1.2.1.2. Reglas de asociación o "*linkage*"

Después del primer paso, cuando cada objeto representa su propio grupo, la distancia entre estos objetos es definida por la métrica escogida. Sin embargo, una vez que varios objetos han sido unidos (*linked*), una regla de asociación (*linkage*) es necesaria para determinar si 2 grupos son lo suficientemente similares para ser unidos. Hay numerosas reglas de asociación que han sido propuestas, de las cuales 3 son las más comunes. Un ejemplo del resultado de estas tres reglas de asociación con el mismo conjunto de datos se presenta en la Figura 2.

Figura 2. **Comparación de las variantes de Agrupación Jerárquica.** Estas son llamadas *Single Linkage*, *Average Linkage* y *Complete Linkage*. Se utiliza un mismo conjunto de datos y se muestran los distintos dendrogramas producidos por la sola variación del método de *Linkage*.



1.2.1.2.1. "*Single linkage*" (distancia más cercana)

Este método consiste en que la distancia entre 2 grupos es determinada por los 2 objetos más cercanos (vecinos más cercanos) entre ambos grupos. Su mayor desventaja es su tendencia a la formación de cadenas, ya que sólo una distancia pequeña al azar es suficiente para forzar la unión de 2 grupos que de otra forma serían muy distintos. De esta manera, los grupos resultantes tienden a representar cadenas elongadas, con genes individuales encadenados a los grupos.

1.2.1.2.2. "*Complete linkage*" (distancia más lejana)

En este método la distancia entre 2 grupos es determinada por la distancia entre los 2 objetos más lejanos (vecinos más lejanos) entre ambos grupos. El método usualmente se comporta bastante bien en casos donde los objetos forman naturalmente nubes de datos distintivas en un espacio multi-dimensional y tiende a crear grupos de similar tamaño y

variabilidad. Sin embargo, es inapropiado si los grupos tienden a tener cierta elongación o son de naturaleza de cadenas. Esto es debido a que una sola distancia larga al azar es suficiente para pretender que 2 grupos se unan, y los grupos tienden a ser pequeños y juntados muy tarde con un mayor valor de error.

1.2.1.2.3. "*Average linkage*" (distancia media)

En este método la distancia entre 2 grupos es calculada como la distancia promedio entre todos los pares de objetos entre ambos grupos. Es muy eficiente cuando los objetos naturalmente forman aglutinaciones distintivas, aunque se comporta igualmente bien con grupos elongados de tipo cadenas, y tiene una ligera tendencia a producir grupos de varianza similar. Debido a que la distancia entre 2 grupos cae entre la mínima de formación del *Single linkage* y la máxima de formación del *Complete linkage*, este procedimiento no muestra empíricamente tendencias a ningún extremo como las dos anteriores, y es por tanto más estable ante distribuciones de datos desconocidas. Por supuesto que si hay varias distancias iguales entre *clusters*, la secuencia de unión es crítica.

1.2.1.3. *Bootstrapping*

Es un método usado para robustecer los resultados de la Agrupación Jerárquica, que consiste en forzar al sistema para producir cambios y mostrar aquellas agrupaciones azarosas o circunstanciales. En este método la matriz de expresión original es recreada muchas veces mediante el reemplazo al azar de un dato por otro dato repetido cada vez, manteniendo el tamaño de la matriz original, y crea una Agrupación Jerárquica para cada una de estas matrices modificadas [11].

Cada uno de los dendrogramas formados es comparado con el dendrograma de la matriz de datos original, y mientras más veces se encuentre un grupo particular de la matriz original en las matrices modificadas, más fuerte es la confianza (*support*) para ese grupo. Una confianza baja implica que la estructura formada es lábil y cambiante ante la modificación de alguno de sus componentes, ya que ante este cambio otros aspectos

priman para formar una estructura distinta la mayoría de las veces. En cambio, una confianza alta indica que el cambio de uno de sus componentes no altera la estructura, pues la gran similitud de sus componentes hace que permanezcan inalterables a pesar de los cambios y el grupo se mantiene invariante en un alto porcentaje de las veces.

Como cada matriz modificada carece de alguno de los datos originales, una alta confianza para un grupo significa que su estructura no se encuentra influenciada por pequeños conjuntos de datos y el resultado original es un dendrograma más robusto con un grado de confianza conocido para cada familia. Es por esto que a esta variante también se le conoce como *Support Clustering*.

1.2.1.4. Características del método

La Agrupación Jerárquica es la estrategia de agrupamiento más usada actualmente para el análisis de expresión génica. Su mayor ventaja es que, aparte de la elección de la regla de asociación y la métrica usada para medir la distancia, no se necesita la especificación de más parámetros. El resultado es un conjunto reordenado de genes y/o experimentos, donde los vectores similares se encuentran uno junto al otro en la estructura del dendrograma, y la distancia entre los vectores y grupos se representa en el largo de las ramas de las subfamilias. Esto no sólo permite la estimación de la similitud entre vectores vecinos, sino también la distancia entre vectores distantes; lo que es útil cuando se investiga las distancias entre dos o más condiciones.

Sin embargo, este método sólo reorganiza el conjunto de datos original a un conjunto nuevo y más ordenado de vectores de datos, por tanto los grupos deben ser especificados por el usuario seleccionando a una subfamilia como grupo. Una segunda desventaja es la complejidad computacional. Conjuntos de datos grandes son difíciles o imposibles de calcular debido a la vasta cantidad de memoria necesaria para la matriz de similitudes y al tiempo de cálculo requerido.

Una característica a tener en cuenta es el hecho de que todo algoritmo producirá un árbol de familias, incluso si los datos no están realmente estructurados como tal. De este modo

los resultados deben ser analizados para comprobar si son concordantes con las características conocidas de los elementos involucrados, además de buscar el algoritmo que mejor se adapte a la realidad de los datos que se investigan.

1.2.2. Análisis de correspondencia

El Análisis de Correspondencia ("*Correspondence Analysis*" o **COA**) es un método explorativo utilizado para estudiar la asociación entre variables. Revela los principales ejes del espacio multi-dimensional que representan los datos, lo que permite su proyección en un subespacio de dos o tres dimensiones que representa la mayor parte de la varianza de estos datos. Despliega una proyección de los datos en un número reducido de dimensiones de máxima varianza, donde tanto los genes como las muestras pueden ser proyectados en el mismo espacio, revelando asociaciones entre ellos [12].

Los genes que se encuentran juntos en el espacio virtual creado tienden a tener perfiles similares, independiente de su valor absoluto, y lo mismo ocurre con las muestras. Si existen genes y muestras que se encuentran juntas en el espacio, entonces estos genes tienden a tener una alta expresión en las muestras vecinas en comparación con las muestras que se encuentran más alejadas de ellos. Por otro lado, si un conjunto de genes se encuentran en el lado opuesto en relación al origen del espacio ocupado por muestras, entonces la expresión de ese conjunto de genes se encuentra probablemente disminuida en comparación a otras muestras ubicadas cerca de esos genes. Mientras más lejos se encuentran los puntos del origen, más fuerte es la asociación entre genes y muestras.

El Análisis de Correspondencia funciona descomponiendo la matriz de los valores chi-cuadrado derivados de las filas y columnas de la matriz de expresión. Normalmente, los dos o tres primeros ejes son los más informativos en mostrar la asociación entre genes y experimentos. La cantidad de información que representa un eje dado se cuantifica por su **Inercia**, que puede ser considerada como la proporción del valor chi-cuadrado total de la matriz explicada por ese eje.

1.2.3. Análisis de significancia de microarreglos

El Análisis de Significancia de Microarreglos ("*Significance Analysis of Microarrays*" o **SAM**) se usa para seleccionar genes significativos en base a su expresión diferencial entre varios conjuntos de muestras [13]. Es útil cuando ya existe una hipótesis de que algunos genes tendrán un nivel de expresión media significativamente distinta entre varios grupos. Una característica valiosa del SAM, es que proporciona el valor estimado de la Tasa de Falso Descubrimiento ("*False Discovery Rate*" o **FDR**), que es la proporción de genes que pudieran haber sido erróneamente identificados como significativos por azar. Es posible implementar varios diseños de SAM. El siguiente es el diseño más común y el que se usa en esta tesis.

- **Dos-clases desapareado**: Es donde las muestras se categorizan en uno de dos grupos, y los sujetos son distintos entre ambos grupos. Una vez fijados los miembros de ambos grupos y realizado el análisis, se considera "positivamente significativos" a aquellos elementos cuya expresión media en el grupo B sea significativamente superior a su expresión media en el grupo A. En cambio son "negativamente significativos" aquellos cuya expresión media en el grupo A sea significativamente superior a su expresión media en el grupo B.

Una vez escogido el diseño del SAM, para cada gen se computa un valor-*d*, que corresponde al *valor d-observado*. Luego se forma un ranking donde se ordenan los genes de manera ascendente según sus valores-*d*. Entonces se mezclan al azar los valores de los genes entre los grupos A y B, de tal manera que ambos grupos sigan conteniendo el mismo número de elementos que los grupos originales. Luego se computa un nuevo valor-*d* para cada gen mezclado al azar. Posteriormente se vuelven a ordenar los genes en orden ascendente según sus valores-*d* permutados. Se repite la randomización muchas veces, de manera que cada gen posea muchos valores-*d* randomizados equivalentes al ranking de su valor-*d* observado (no permutado). El promedio de todos estos valores randomizados constituyen el *valor d-esperado* para cada gen. En un diseño de 2 clases desapareado, *d* es análogo al valor estadístico *t* en

una prueba de *t*, ya que captura la diferencia entre los niveles de expresión media de las condiciones experimentales, puestos a escala por la medición de la varianza de los datos.

SAM genera una proyección de 2 dimensiones de los valores *d-observados* contra los *d-esperados* (basados en los datos permutados). Luego se ajusta el parámetro *delta*, que es la distancia vertical entre el límite de significancia superior e inferior y la línea que representa a los valores donde *d-observado* es igual a *d-esperado*. El valor de *delta* determina el límite que un gen debe traspasar para considerar que su expresión diferencial entre los grupos es positivamente o negativamente significativa.

Además se puede aplicar un criterio de magnitud de cambio, donde un gen debe cumplir, aparte del criterio del valor *delta*, con esta condición para ser considerado significativo. Para un cambio de magnitud **F** ("*Fold Change*"):

$$\frac{MediaNoLogarítmica(B)}{MediaNoLogarítmica(A)} \geq F$$ Para genes positivamente significativos.

$$\frac{MediaNoLogarítmica(B)}{MediaNoLogarítmica(A)} \leq \frac{1}{F}$$ Para genes negativamente significativos.

1.3. Hipótesis

1. El agrupamiento de perfiles de metilación de genes supresores de tumores permitiría definir subtipos clínico-patológicos del CG de tipo difuso.

Esta hipótesis será probada mediante el análisis de Agrupamiento Jerárquico de genes relevantes en 2 grupos distintos de casos de CG. El primer grupo corresponde a 32 casos que poseen muestras disponibles de tejido de alta calidad (tejido fresco). El segundo grupo corresponde a 104 casos que poseen muestras disponibles de tejido de menor calidad (tejido fijado en formalina e incluido en parafina), pero de los cuales existe información clínico-patológica, incluyendo datos de sobrevida.

2. El análisis *in-silico* de datos de transcriptoma de librerías de casos de CG, disponibles en sitios de dominio público, es capaz de identificar potenciales nuevos genes relevantes en la patogénesis del CG de tipo difuso.

Esta hipótesis será probada mediante un análisis *in-silico* de librerías provenientes tanto de "*Cancer Genome Anatomy Project*" (**CGAP**, http://cgap.nci.nih.gov/), como de "*Gene Expression Omnibus DataSets*" (**GEO DataSets**, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=gds).

1.4. Objetivos

Objetivo General:

Evaluar el uso de herramientas bioinformáticas para identificar perfiles de expresión y potenciales nuevos marcadores asociados a la patogénesis del CG de tipo difuso.

Objetivos Específicos:

1. Agrupar los genes identificados en mucosa tumoral de pacientes de cáncer gástrico, de tejido fresco y tejido incluido en parafina, mediante técnicas bioinformáticas de Agrupación Jerárquica (*Hierarchical Clustering*), para construir grupos de genes con valor predictivo en cáncer gástrico.
2. Evaluar las herramientas bioinformáticas en librerías de SAGE de estómago, y comparar sus resultados con la búsqueda tradicional de genes por bibliografía realizada previamente.

II. MATERIALES Y MÉTODOS

2.1. Bioinformática en estudio de metilación de cáncer gástrico

2.1.1 Casos clínicos

El material correspondió a tejido fresco y tejido incluido en parafina de piezas quirúrgicas de pacientes operados por CG entre 1993 y 2003, mantenidos en el Departamento de Anatomía Patológica de Universidad Católica (tejido fresco) y el Instituto Chileno-Japonés de Enfermedades Digestivas - Hospital Clínico San Borja Arriarán (ICHJED-HCSBA), Universidad de Chile (tejido incluido en parafina). De este modo se ingresaron 32 casos de CG en tejido fresco y 104 casos de CG en tejido incluido en parafina. Ambos grupos comparten características clínico-patológicas similares y se muestran en la Tabla 1.

Tabla 1. **Comparación de características clínico-patológicas de casos de tejido fresco y tejido incluido en parafina utilizados en este estudio.** Se observan el número y porcentaje de cada característica con respecto a los totales (32 muestras para tejido fresco y 104 muestras para parafina).

Cáncer Gástrico	***Tejido fresco***		***Tejido en parafina***	
Variables Clínicas	**n**	**%**	**n**	**%**
Sexo				
Hombre	21	65,6%	66	63,5%
Mujer	11	34,4%	38	36,5%
Edad Promedio				
	65 (45 - 80 años)		59 (42 - 86 años)	
Ubicación				
Cardia	11	34,4%	40	38,5%
Medio	8	25,0%	29	27,9%
Antro	13	40,6%	35	33,6%
Estadío				
Temprano	4	12,5%	14	13,5%
Avanzado	28	87,5%	90	86,5%
Linfonodos				
Positivo	22	68,8%	75	72,1%
Negativo	10	31,3%	29	27,9%
Mucinoso/Anillo de Sello				
Positivo	13	40,6%	37	35,6%
Negativo	19	59,4%	67	64,4%

Este material ha sido identificado como Cáncer Gástrico Difuso[14], utilizando la definición de Lauren [15], y son la base del presente estudio. Las características clínicas de estos casos se obtuvieron de la revisión de fichas médicas, características patológicas de informes anátomo-patológicos, y revisión de las láminas correspondientes.

Las características clínicas consideradas fueron el sexo (mujer, hombre), edad (menor o igual a 58 años, mayor de 58 años) y sobrevida global (censura). Las características anátomo-patológicas consideradas fueron la ubicación del tumor (no antral, antral), infiltración de pared gástrica (mucosa/submucosa, muscular propia/serosa) y compromiso ganglionar (sin consignar el número de ganglios comprometidos), de acuerdo a la Unión Internacional Contra el Cáncer. Adicionalmente se incluyó la presencia del virus de Epstein-Barr (EBV) como variable anátomo-patológica en el estudio, dado que recientemente se ha descrito una fuerte asociación entre el cáncer gástrico difuso y la infección por EBV [16]. El estudio contó con la aprobación del Comité de Ética del Servicio de Salud Metropolitano Central (Hospital Clínico San Borja Arriaran).

2.1.2. Análisis estadístico y bioinformático

Los resultados de metilación proporcionados por el Laboratorio de Biología Molecular del Departamento de Anatomía Patológica de la Universidad Católica fueron agrupados según el Índice de Metilación (IM), que se define como la proporción de genes con metilación positiva dentro del total de genes analizados, y se clasificó como alto si tenía un valor mayor a 30% o bajo si era menor a 30% de acuerdo a la literatura.

Se utilizó el programa de TIGR (The Institute for Genomic Research), llamado *MultiExperiment Viewer* (**MeV**) (http://www.tm4.org/) [17] para medir la similitud entre las variables analizadas. Se utilizó la herramienta de *Support Tree* para obtener una visualización de las interacciones, que consiste en la aplicación de una Agrupación Jerárquica usando el método de *Bootstrap* (consistente en el reemplazo aleatorio de valores de la matriz de datos por otros repetidos durante 100 iteraciones y la ejecución

de una Agrupación Jerárquica para cada matriz permutada). Se utilizó la métrica Euclidiana y se consideró la máxima distancia de los componentes de los grupos entre sí (*Complete linkage*). De este modo la Agrupación Jerárquica, a diferencia de otros métodos estadísticos, otorga una visión gráfica de asociaciones particulares entre los genes estudiados y las variables clínico-patológicas entre sí, y no promedios de datos como el Índice de Metilación. Las asociaciones más cercanas observadas mediante la Agrupación Jerárquica fueron confirmadas por la prueba de chi cuadrado, con un valor significativo de $p<0,05$ (Epi Info 2000).

2.2. Análisis bioinformático en SAGE de estómago

Se utilizaron todas las librerías disponibles de SAGE (*Serial Analysis of Gene Expression*) de estómago para buscar las diferencias entre el transcriptoma de los estados normales y tumorales. A diferencia del estudio anterior, estas librerías son más generales pues surgen de muestras de CG tanto de tipo histológico "intestinal" como "difuso". Se usaron 2 librerías provenientes del sitio web de GEO DataSets (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=gds), GDS551 [18, 19] y GDS442 [20] (de la cual se excluyó la librería de metástasis de Linfonodos), y 5 librerías adicionales provenientes del sitio web de CGAP (http://cgap.nci.nih.gov/).

Todas las librerías usan tags de 10 pb y las mismas enzimas en el proceso (*Bsm*FI y *Nla*III), sin embargo es importante hacer notar que las librerías de GDS442 no contenían muestras de tejido de sujetos normales. Algunos detalles de las librerías pueden verse en la Tabla 2.

Tabla 2. Detalle de las librerías de SAGE. Se observa el número total de tags únicos y el número total de tags para todas las librerías.

Librerías Normales	*Total de tags únicos*	*Total de tags*	*Librerías Tumorales*	*Total de tags únicos*	*Total de tags*
CGAP_normal_MD_13S	15.700	45.908	CGAP_adenocarcinoma_MD_HG7	34.469	93.714
CGAP_normal_MD_14S	18.718	73.353	CGAP_adenocarcinoma_MD_HS29	22.385	58.831
GSM784(normal_B_antrum)	9.137	25.302	CGAP_adenocarcinoma_MD_G329	19.208	46.802
GSM14780(normal_B_antrum)	9.349	26.653	GSM757(GC_G234)	21.967	66.032
			GSM758(GC_xenograft_X101)	25.293	70.433
			GSM14760(GC_B_X43)	19.923	51.620
			GSM2385(GC_G189)	18.010	64.102
			GSM7800(Hiroshima_GC_S219T)	14.576	34.660
			GSM8505(Hiroshima_GC_W246T)	12.792	32.174
			GSM8867(Hiroshima_GC_W226T)	16.082	43.908
			GSM9103(Hiroshima_GC_P208T)	6.135	11.582

Todas las librerías normales consisten en una mezcla de tejidos (GSM784 y GSM14780) o muestras microdisecadas (MD) (CGAP_MD_13S y CGAP_MD_14S), y fueron producidas por el laboratorio de El-Rifai, Universidad de Virginia, EEUU.

Las librerías tumorales GSM757, GSM2385, GSM758, GSM14760, CGAP_MD_HG7, CGAP_MD_HS29 y CGAP_MD_G329 son muestras de pacientes con adenocarcinomas, obtenidas por el laboratorio de El-Rifai, Universidad de Virginia, EEUU.

Las librerías tumorales GSM7800, GSM8505, GSM8867 y GSM9103 son muestras de adenocarcinomas obtenidas por el laboratorio de Yasui y Oue, Universidad de Hiroshima, Japón.

Se generó una base de datos que contenía 121.409 tags diferentes. La frecuencia de cada tag fue normalizada al dividirla por el total de tags y multiplicada por 200.000 tags (debido a que es el formato estándar usado por CGAP). La librería GSM9103 fue removida debido a que su conteo de tags distintos (alrededor de 6 mil) era muy bajo, tal vez por una baja eficiencia. Las demás librerías abarcan entre 9 mil y 34 mil tags distintos.

Se utilizó el programa de TIGR (The Institute for Genomic Research), llamado *MultiExperiment Viewer* (**MeV**) (http://www.tm4.org/) [17], mencionado anteriormente, para ejecutar los siguientes análisis bioinformáticos:

1. *Support Tree* (ST): para buscar patrones de expresión similares en las muestras. Es una Agrupación Jerárquica normal repetida al menos 100 veces con el método de *Bootstrap* (*Support Clustering*), con el fin aumentar la robustez del resultado.
2. Análisis de Correspondencia (COA): para explorar asociaciones entre variables, genes o experimentos, que cuando están cercanos en este espacio virtual tienden a tener perfiles similares.
3. Análisis de Significancia de Microarreglos (SAM): para seleccionar genes cuya media de expresión es significativamente distinta entre 2 o más grupos.

Se seleccionaron 2.437 tags que se cumplían con el criterio de haber sido detectadas en "*todas las librerías normales*" o en "*todas las librerías tumorales*", con el propósito de reducir el ruido que genera una enorme cantidad de tags que sólo son detectados 1 vez

en 1 sola librería, y para reducir la enorme potencia computacional que sería necesaria para realizar los cálculos sobre toda la base. Este criterio se basa en la búsqueda de tags que sean consistentemente expresados en una u otra condición, y que pudieran variar entre ambas ya sea por una reducción, sobreexpresión o desaparición entre ambas condiciones.

Luego los datos de frecuencia normalizada de tags se transformaron a una escala logarítmica en base 2 para ser usada por TIGR MeV. Esta transformación logarítmica tiene por objetivo reducir la escala para compensar la desviación que se generaría por la mezcla de una gran cantidad de números pequeños, provenientes de frecuencias normalizadas de tags de 1 copia y 0 copia, y un muy reducido grupo de números grandes provenientes de los tags más abundantes. Al haber más de 3 órdenes de magnitud entre los extremos (0 a >3000), es necesario compensar mediante la transformación de la escala, ya que la mayor parte de los valores se encuentra en el rango de 0 a 100, como ocurre también con los datos de señales de microarreglos para los cuales estas herramientas fueron pensadas originalmente.

El análisis jerárquico de grupos se realizó usando el método de "*Bootstrap*" (conocido como *Support Clustering*) mediante la iteración del proceso 100 veces. Se usó la correlación de Pearson y el *Average linkage* como medidas de semejanza para la creación de los grupos.

El análisis de SAM se hizo con un delta calculado en cada caso para mantener el FDR (la probabilidad de encontrar tags significativos por azar) cercano a 0, además de usar un factor de cambio de magnitud F=10 como medida de estrictez adicional.

El análisis de correspondencia se realizó con su configuración predeterminada en el programa, ya que esencialmente es un método no supervisado que no requiere ninguna intervención por parte del usuario.

Finalmente la asociación de tags a genes se hizo usando TAGmapper como referencia (http://tagmapper.ibioinformatics.org) [21]. Esta herramienta en línea es producto de la colaboración conjunta de Pandey Lab (de la Universidad de Johns Hopkins, EEUU) y el *Institute of Bioinformatics* (India).

III. RESULTADOS Y DISCUSIÓN

3.1. Bioinformática en estudio de metilación de cáncer gástrico

Los antecedentes de la literatura indican que la hipermetilación de genes supresores de tumores (GST) en CG se encuentra asociada a la infección por el virus de Epstein-Barr (EBV) [22], los que en su gran mayoría corresponden a CG del tipo histológico difuso [16], y que la hipometilación global del genoma es más frecuente en CG del tipo histológico intestinal [23], sugiriendo dos vías independientes de carcinogénesis gástrica. Por otra parte no hay antecedentes en la literatura sobre correlaciones clínico-patológicas entre hipermetilación de GST y variables clínicas, de sobrevida, ni tampoco sobre el rol de la hipermetilación de GST como evento precursor de CG. Estos antecedentes junto con las evidencias de detección de hipermetilación de GST en sangre periférica en carcinoma colorectal, esofágico y nasofaríngeo, llevaron a plantear las siguientes hipótesis en relación con la hipermetilación de GST:

A. Estaría asociada a la patogénesis del CG difuso.
B. Estaría asociada a características clínico-patológicas, de sobrevida y a las primeras etapas del tumor (“efecto de campo”).

Para resolver estas hipótesis se propuso identificar el perfil de hipermetilación de GST en casos retrospectivos de CG. Para ello se identificaron GST relevantes en la literatura para CG que cubrían distintas vías celulares, y se logró obtener resultados consistentemente reproducibles en 25 genes que aparecen en la Tabla 3.

Tabla 3. **Genes supresores de tumores provenientes de literatura e inactivados en cáncer gástrico.** Se encuentran ordenados por su localización cromosomal y se muestra su función general.

Gen	*Localización*	*Función*	*Gen*	*Localización*	*Función*
RUNX3	1p36.11	proliferación	p14	9p21.3	ciclo celular
RIZ1	1p36.21	factor transcripcional	p15	9p21.3	ciclo celular
p73	1p36.32	apoptosis	p16	9p21.3	ciclo celular
COX-2	1q31.1	inflamación	DAPK	9q21.33	apoptosis
TIMP3	22q12.3	invasión	PTEN	10q23.31	ciclo celular
Reprimo	2q23.3	ciclo celular	MGMT	10q26.3	reparación
FHIT	3p14.2	apoptosis	GSTp1	11q13.2	reparación
BLU	3p21.31	unión a proteína	SHP1	12p13.31	apoptosis
SEMA3b	3p21.31	crecimiento axonal	3OST2	16p12.1	actividad sulfotransferasa
hMLH1	3p22.3	reparación	SOCS-1	16p13.13	crecimiento
RARbeta	3p24	diferenciación	E-Cadherina	16q22.1	adhesión
APC	5q22.2	adhesión	BRCA1	17q21.31	reparación
ER	6q25.1	diferenciación			

3.1.1. Grupo de muestras en tejido fresco

De este modo, se estudiaron 25 genes en un grupo piloto de 32 casos de CG difuso en tejido fresco. Los resultados de este análisis se muestran en la Tabla 4.

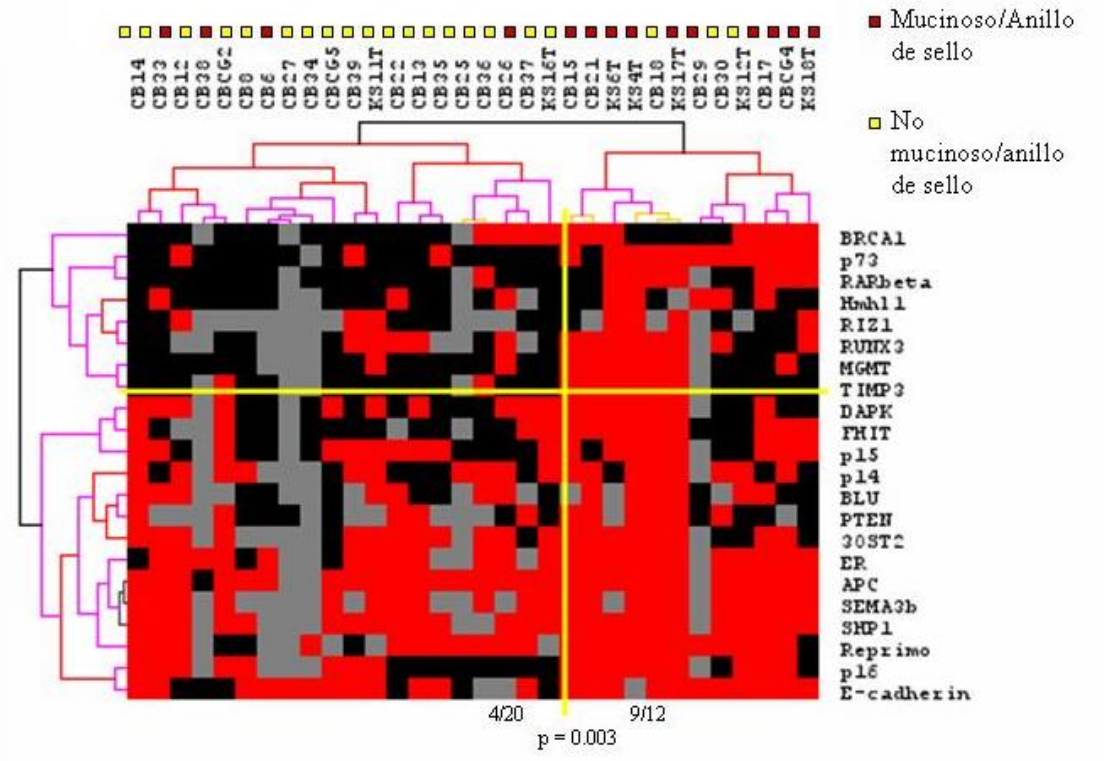
Tabla 4. **Análisis de metilación en cáncer gástrico difuso en muestras de tejido fresco.** Se muestra el número de muestras en las cuales se encontró metilado y el porcentaje equivalente con respecto al total de 32 muestras.

Gen	*Metilación*		*Gen*	*Metilación*	
	n	**%**		**n**	**%**
APC	28	87,5%	RUNX3	13	40,6%
SHP1	25	78,1%	FHIT	13	40,6%
ER	23	71,9%	PTEN	11	34,4%
E-Cadherina	23	71,9%	BRCA1	11	34,4%
Reprimo	22	68,8%	MGMT	9	28,1%
SEMA3B	21	65,6%	hMLH1	8	25,0%
3OST2	20	62,5%	TIMP3	8	25,0%
p14	18	56,3%	RAR-beta	8	25,0%
p15	18	56,3%	RIZ1	7	21,9%
p16	17	53,1%	SOCS	3	9,4%
DAPK	17	53,1%	COX2	3	9,4%
p73	14	43,8%	GSTp1	0	0,0%
BLU	13	40,6%			

Se observa que 24 (96%) de los 25 genes mostraron hipermetilación en al menos 3 o más casos estudiados, mientras 11 genes estuvieron metilados en al menos el 50% de los casos.

Para explorar correlaciones entre los casos y los perfiles de hipermetilación de GST se utilizó el *Support Clustering* descrito anteriormente (Agrupación Jerárquica con método de *Bootstrap*) [17]; sin embargo, SOCS, COX-2 y GSTp1 fueron excluidos del análisis debido a que su número de casos metilados era muy bajo (3, 3 y 0 casos respectivamente), por lo que no proporcionaban ninguna información asociativa. Se utilizó la métrica Euclidiana y *Complete linkage*. Los resultados de este análisis se muestran en la Figura 3. Se observa que los casos se agrupan en 2 ramas, de un alto grado de confianza, que difieren en la proporción de casos según la variedad histológica mucinoso/anillo de sello (p=0.003), una forma de CG difuso de muy mal pronóstico [24]. Por otro lado, los genes se agrupan en un bloque de alto grado de metilación y otro de bajo grado de metilación. No es posible sin embargo, obtener más detalles en las asociaciones debido a la baja confianza de los grupos internos a estas ramas principales, lo que indica una ausencia de otros elementos discriminadores importantes más allá de los macrogrupos encontrados.

Figura 3. **Agrupación Jerárquica para explorar correlaciones entre los casos y los perfiles de hipermetilación de genes supresores de tumores.** SOCS, COX-2 y GSTp1 fueron excluidos de la figura. Se utilizó la métrica Euclidiana y *Complete linkage*. Dentro de la matriz se ve una escala binaria (0/1), el color rojo indica presencia de metilación, el color negro indica ausencia de metilación, y el color gris indica un dato faltante. Se observa que los casos se agrupan en dos ramas de alta confianza (ver escala en la Figura 6) para genes y experimentos; sin embargo, casi todos los grupos internos son de muy baja confianza, por lo que se descarta un análisis más minucioso de los dendrogramas.



3.1.2. Grupo de muestras en parafina

Para conocer el significado clínico de los perfiles de hipermetilación de los GST identificados en CG difuso, once genes (APC, Reprimo, E-Cadherina, p14, p15, p16, FHIT, p73, BRCA1, MGMT y hMLH1) fueron evaluados en el grupo de 104 casos de CG del banco de tumores del ICHJED-HCSBA. Los resultados de estos análisis fueron informativos sólo en 83 tumores de los 104 iniciales, debido a la reconocida mala calidad del tejido incluido en parafina. Sin embargo, la mala calidad se compensa por el acceso a grandes números de muestras. Los resultados de estas correlaciones se muestran en la Tabla 5.

Tabla 5. **Índice de Metilación (IM) y correlaciones clínico-patológicas en cáncer gástrico difuso.** Se ha definido el límite del índice de metilación en 30% de los genes estudiados, mediante el uso de tejido incluido en parafina.

Variables Clínicas	*IM < 30*		*IM > 30*		
	n	**%**	**n**	**%**	**p test**
Género					
Mujer	15	32%	15	42%	0,36
Hombre	32	68%	21	58%	
Edad					
<58 años	22	47%	16	44%	0,83
>58 años	25	53%	20	56%	
Censura					
Vivo	20	43%	15	42%	0,54
Muerto	19	40%	19	53%	
Ubicación					
No antral	26	58%	22	65%	0,53
Antral	19	42%	12	35%	
Estado					
Incipiente	7	15%	5	14%	0,87
Avanzado	39	83%	31	86%	
Linfonodos					
Negativo	10	21%	6	17%	0,49
Positivo	29	62%	26	72%	
Virus Epstein-Barr					
Negativo	34	72%	22	61%	0,07
Positivo	8	17%	13	36%	

Se definió el Índice de Metilación (IM) como el número de genes metilados *versus* el número de genes analizados. El promedio del índice de metilación fue de 0,23 y utilizando un punto de corte de 30% por caso, se observó que la única asociación clínico-patológica con índice de metilación >30% fue la ausencia de infección por EBV (p=0.07). El estudio de sobrevida no demostró asociaciones con el IM, considerando o no la presencia de EBV.

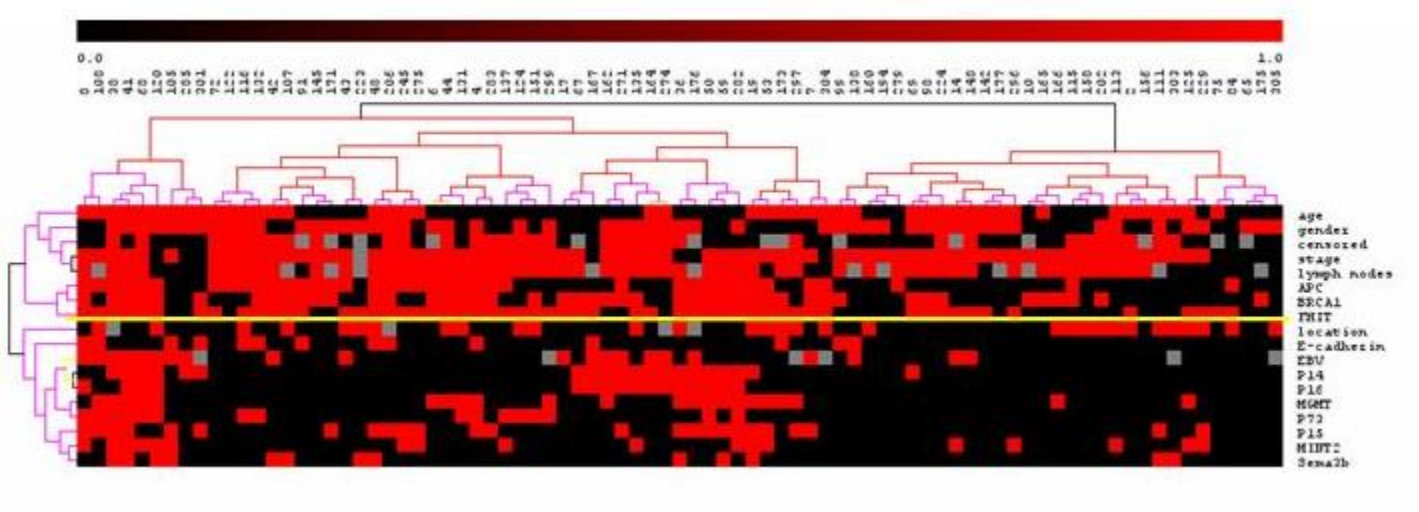
Las correlaciones de genes individuales en tumor con variables clínico-patológicas demostró que en general los genes más frecuentemente metilados en casos de tejido fresco también estaban metilados en muestras de tejido incluido en parafina (APC, FHIT y BRCA1), con la excepción de E-Cadherina y Reprimo (Tabla 6).

Tabla 6. **Análisis de metilación en cáncer gástrico difuso en muestras de tejido incluido en parafina.** Se muestra el número de muestras en las cuales se encontró metilado y el porcentaje equivalente con respecto al total de 83 muestras.

Gen	*Tumor*	
	Positivos	**%**
BRCA1	63	76,8%
FHIT	61	74,4%
APC	52	63,4%
p16	40	48,8%
p14	40	48,8%
p15	32	39,0%
E-Cadherina	28	34,1%
p73	28	34,1%
Reprimo	26	31,7%
MGMT	25	30,5%
hMLH1	10	12,2%

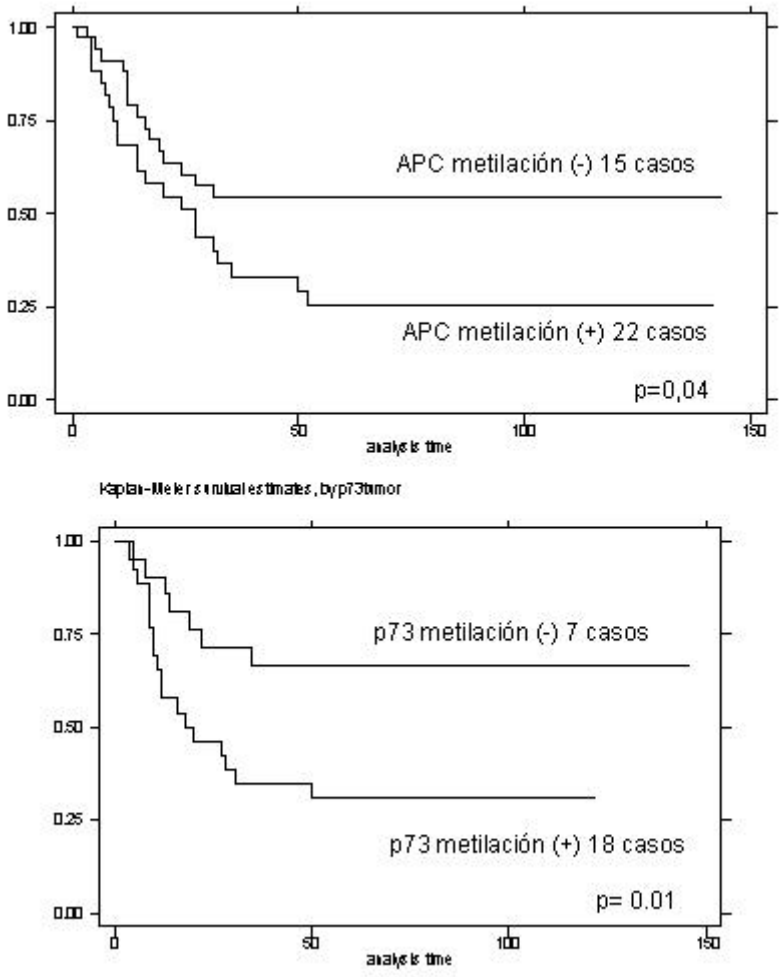
Dado que los análisis de índice de metilación >30% no demostraron asociaciones significativas, se decidió explorar asociaciones específicas entre genes y variables clínicas utilizando el método bioinformático de Agrupación Jerárquica [17], con métrica Euclidiana y *Complete linkage*. Como ya se ha mencionado esta metodología permite tener una visión gráfica de las asociaciones entre los genes y las variables clínico-patológicas estudiadas, y sus resultados pueden verse en la Figura 4. Se observan dos grupos, un grupo superior con casi todas las variables clínico-patológicas (edad, género, censura, estadío y presencia de linfonodos), excepto ubicación, junto con FHIT, BRCA1 y APC; y un grupo inferior con 8 genes junto con ubicación e infección por EBV. Además se observa también una fuerte asociación entre la mutilación de p14 y p16 e infección por EBV.

Figura 4. **Análisis de Agrupación Jerárquica en 83 casos de cáncer gástrico difuso.** Se utilizó la métrica Euclidiana y *Complete linkage*. Pueden observarse dos grupos de alta confianza (ver escala en la Figura 6), uno superior con todas las variables clínico-patológicas (excepto ubicación) junto con FHIT, BRCA1 y APC, y uno inferior con 8 genes junto con ubicación e infección por EBV.



Las asociaciones más cercanas observadas en el dendrograma fueron confirmadas por la prueba de chi cuadrado, con un valor significativo de $p<0,05$ (Epi Info 2000). Los análisis de correlaciones clínico-patológicas indicaron que la metilación de FHIT estaba asociada a pacientes mayores de 58 años ($p = 0.042$), que las metilaciones de APC y p73 estaban asociadas a peor pronóstico ($p = 0.04$ y $p = 0.01$ respectivamente) (Figura 5) y que la metilación de p14 y p16 estaban asociadas a la infección por EBV ($p = 0,0006$ y $p = 0,0004$ respectivamente). De este modo, fue confirmada la importancia de 2 de 3 de los genes predichos por el dendrograma como asociados a variables clínicas importante (FHIT y APC), y la asociación entre la metilación de p14, p16 y la infección por el virus de Epstein-Barr.

Figura 5. **Análisis de sobrevida Kaplan-Meier para comparación de metilación de APC y p73 en cáncer gástrico difuso.** El eje Y representa el número de pacientes vivos, que va disminuyendo a medida que aumenta el tiempo, reflejado en el eje X. Se observa una menor sobrevida de los pacientes que poseen APC o p73 metilado cuando se comparan con pacientes que no tienen metilados estos genes.



En este sentido es interesante mencionar que durante el desarrollo de este proyecto se completó un estudio de genotipos de EBV en tumores CG-EBV positivo. Los resultados identificaron una sola cepa asociada a CG en dos países latinoamericanos (Chile y Colombia) a pesar de la existencia de múltiples cepas en la población sana [25]. Esta información indica una alta selectividad viral en la transformación neoplásica de la célula gástrica. En paralelo, un estudio adicional reveló que las interacciones entre esta cepa única de EBV y las proteínas del huésped en casos de CG, presentaban una

exclusión entre p53 (por inactivación funcional) y p16 (por hipermetilación), los 2 genes más importantes en la patogénesis del CG asociado a EBV [26, 27]. La implicancia de estos resultados es que a pesar de que la transformación celular está restringida a una sola cepa de EBV, el impacto en la sobrevida de los pacientes CG-EBV positivo dependería de la proteína celular (p53 o p16) con la cual interactúa EBV.

3.1.3. Discusión del estudio de metilación

Después de un seguimiento de 25 genes en más de 100 casos de CG se ha logrado identificar un perfil de genes asociados a la variedad más agresiva de CG difuso, y 2 genes asociados a mal pronóstico (APC y p73).

Los datos indican que el uso del método de análisis gráfico de Agrupación Jerárquica es capaz de identificar asociaciones potencialmente significativas entre variables clínico-patológicas y metilación de GST en estudios con gran cantidad de biomarcadores, en particular sobrevida y metilación de APC, e infección por EBV y metilación de p14 y p16. Es interesante la observación de que estos mismos resultados no demuestran asociaciones al analizarlos por medio del Índice de Metilación, una aproximación clásica que mide promedio de datos y no datos particulares.

Con respecto a las frecuencias de metilación de los distintos genes en estudio, llama la atención la baja frecuencia relativa de E-Cadherina, ya que se ha descrito metilada en casi el 80% de los casos de CG difuso [28]. Probablemente, el rendimiento del ensayo MS-PCR para E-Cadherina está disminuido por el tamaño de la región amplificada (120 pb) y el tipo de muestra utilizada, tejido incluido en parafina [29]. En este sentido, los tres genes que con mayor frecuencia aparecieron como metilados (FHIT, APC y BRCA1) corresponden a tamaños entre 67 y 98 pb. Esta información puede ser relevante al momento de interpretar los resultados porque pueden estar influidos por el tamaño de la secuencia utilizada para la identificación de genes particulares.

Una proyección del estudio sería aplicar la metilación de GST en la búsqueda de marcadores de detección precoz en CG difuso. En este sentido se ha reportado que el

estado de metilación de los genes APC, c-met, y p53 puede ser detectado en sueros de pacientes con CG [30].

3.2. Análisis bioinformático en SAGE de estómago

El CG posee la segunda mortalidad más alta en el mundo, pero existen marcadas variaciones geográficas. Los índices más altos se han registrado en países asiáticos (Japón y Corea), europeos (Islandia, Portugal y Bulgaria) y latinoamericanos (Chile y Costa Rica). La menor mortalidad se ha reportado en Canadá y en Estados Unidos (EEUU). Existe evidencia emergente que apoya el concepto de que la etnicidad puede jugar un rol en el pronóstico o en la caracterización de subgrupos de pacientes con CG [31, 32]. La sobrevida a 5 años estratificada por etapas es marcadamente mayor en Japón y Corea que en EEUU. Este índice de sobrevida puede reflejar diferencias en el criterio de diagnóstico, mejores métodos de clasificación y cirugías más radicales. Sin embargo, diferencias en la biología del tumor podrían contribuir a los resultados dispares entre los países de Oriente y Occidente. Esta última perspectiva se ha basado en estudios en asiáticos, que han recibido un diagnóstico de CG en EEUU teniendo una enfermedad menos avanzada que los no-asiáticos [31]. Aunque esto no ha podido ser confirmado por Gill y col. [32], Theuer y col. [31] han sugerido que las diferencias relacionadas con la etnicidad en la biología del tumor no pueden ser completamente excluidas como potencial factor contribuyente a la disparidad de sobrevida surgida en el CG entre el los países de Oriente y Occidente. Para obtener mayor información, tanto de las diferencias de los transcriptomas entre tejido gástrico normal y el CG como de las diferencias entre el CG del Oriente y Occidente, se combinaron varias librerías de SAGE normales y tumorales de dominio público; tanto de *Gen Expression Omnibus Data Sets* (**GEO DataSets**) [33, 34], como de *Cancer Genome Anatomy Project* (**CGAP**) [35]). El Análisis Seriado de la Expresión de Genes ("*Serial Analysis of Gene Expression*" o **SAGE**) es un método de útil para crear perfiles de expresión que permiten la caracterización global, objetiva y cuantitativa de los transcriptomas [36].

Tres son las características principales de la metodología de SAGE:

1. Existe una secuencia corta o larga (14 ó 21 pb), llamada Tag de SAGE, que contiene suficiente información para identificar de manera inequívoca a un transcrito la mayor parte del tiempo. Esto debido a que este tag se obtiene de una única posición 3´ dentro de cada transcrito.
2. Los tags de secuencias pueden ser unidos para formar moléculas en serie que pueden ser clonadas y secuenciadas.
3. Cuantificación del número de veces que un tag en particular (llamado Número de Frecuencia de Tag de SAGE) es observado, lo que provee el nivel de expresión estimada del transcrito correspondiente.

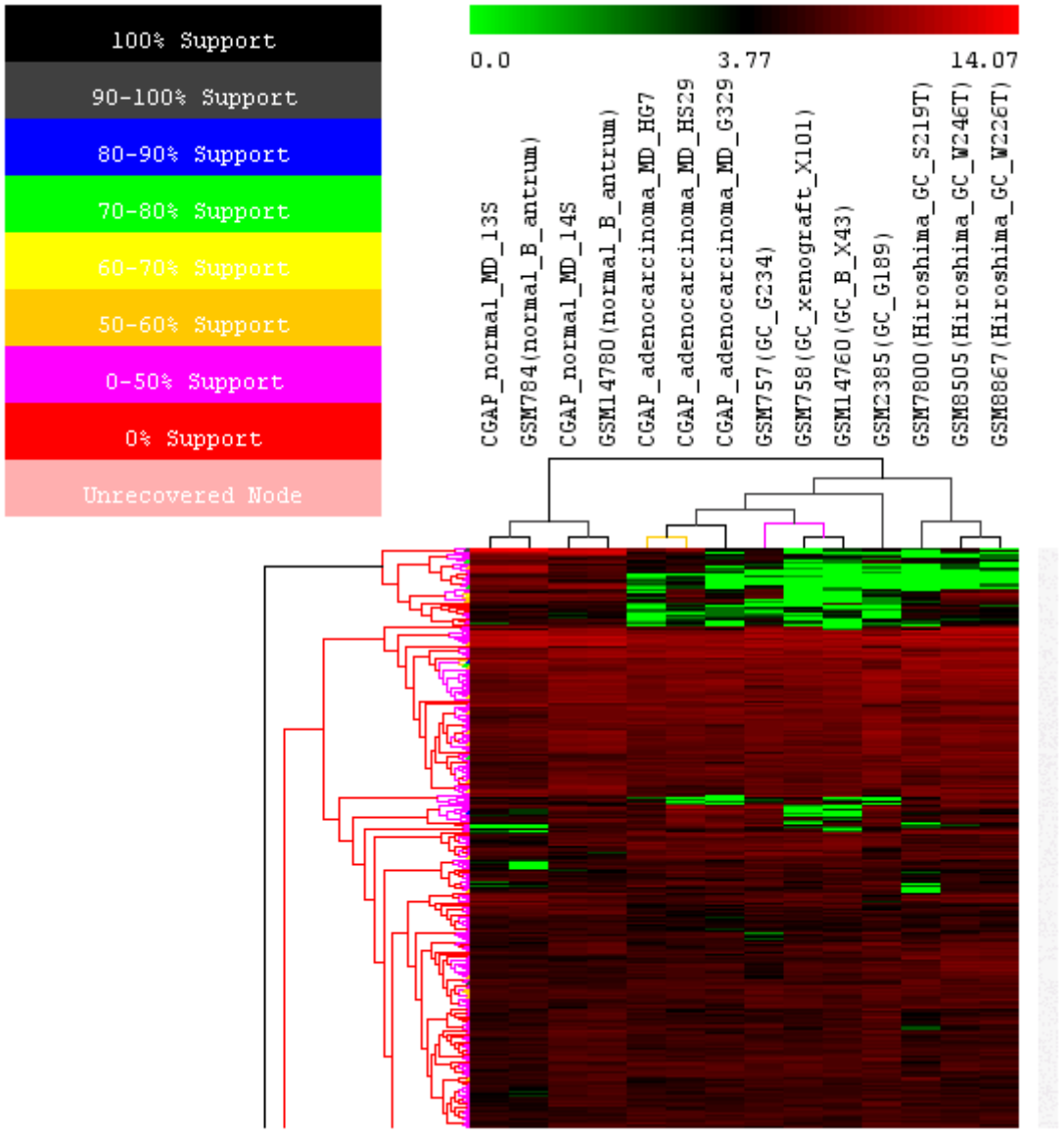
Una ventaja considerable de SAGE es que, una vez normalizado, es posible comparar directamente los niveles de tags generados en un solo experimento (Librería) con cualquier otro experimento disponible. El análisis simultáneo de todas las librerías de estómago encontradas proporciona detalles sobre sus semejanzas y diferencias.

3.2.1. Patrones de expresión

Después de seleccionar los datos a analizar de la base de datos de SAGE de estómago creada, se realizó el *Support Cluster* y los resultados obtenidos se muestran en la Figura 6. El patrón de expresión de los tags resultante muestra una organización altamente variable que se refleja en su baja confianza. En cambio, los experimentos presentan una organización estructurada con un alto grado de confianza en sus distintas ramas (barras negras equivalentes a 90%-100% de confianza), dado por una gran cantidad de elementos discriminadores (tags).

En el árbol de librerías se distinguen familias y subfamilias marcadamente distintas. Las dos ramas principales corresponden a una división entre librerías normales y tumorales. La rama normal muestra dos grupos con muestras de tejido con y sin microdisección mezclados indistintamente. La familia tumoral muestra dos subfamilias, una contiene todas las librerías provenientes de EEUU y la otra contiene todas las librerías provenientes de Japón.

Figura 6. ***Support Cluster*** **(correlación de Pearson y** ***Average linkage*****).** En el extremo superior izquierdo se encuentra el código de colores que representa el grado de confianza (*support*) para cada grupo. Sólo se muestra el límite superior de la figura, ya que ésta es extremadamente larga y aquí sólo se analiza el dendrograma de los experimentos. Puede verse fácilmente la división entre las librerías normales y las tumorales, así como la agrupación de las librerías tumorales japonesas en una familia separada del resto. El dendrograma es además apoyado por una altísima confianza para casi toda su estructura.



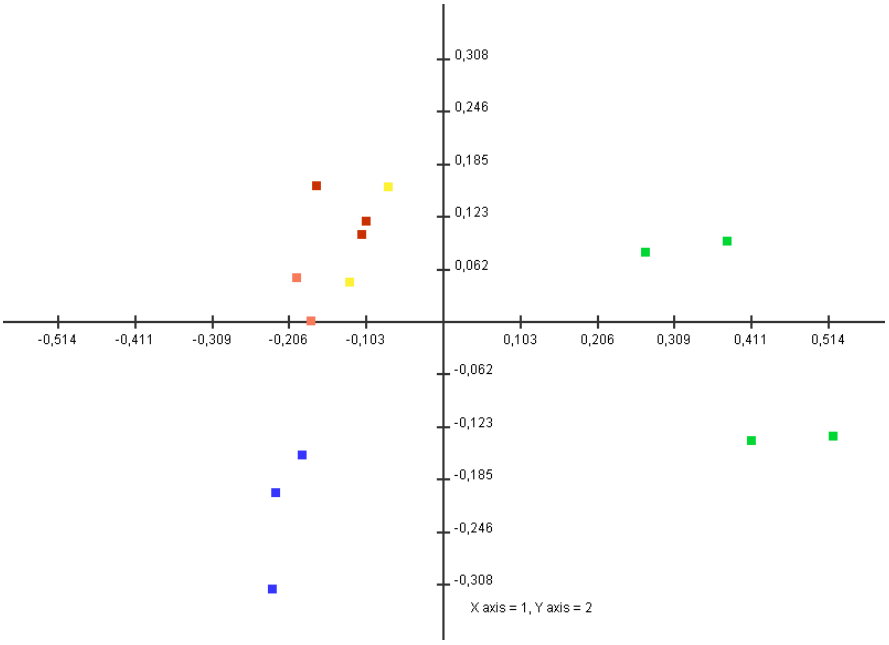
El grupo de tumores japoneses contiene una pareja central (GSM8505 y GSM8867) provenientes de tumores bien diferenciados histológicamente, y la tercera librería (GSM7800), que proviene de un tumor pobremente diferenciado, en la periferia.

Los tumores de EEUU poseen dos grupos distintivos; el primero corresponde a las tres librerías microdisecadas, y el segundo a las librerías de muestras sin microdisecar y aquellas expandidas por injerto en ratones (xenotrasplante), con GSM2385 como el elemento más distanciado del resto. Los tumores de EEUU son una mezcla de adenocarcinomas moderada y pobremente diferenciados, histologías que aparentemente no presentan diferencias sustantivas en este ordenamiento.

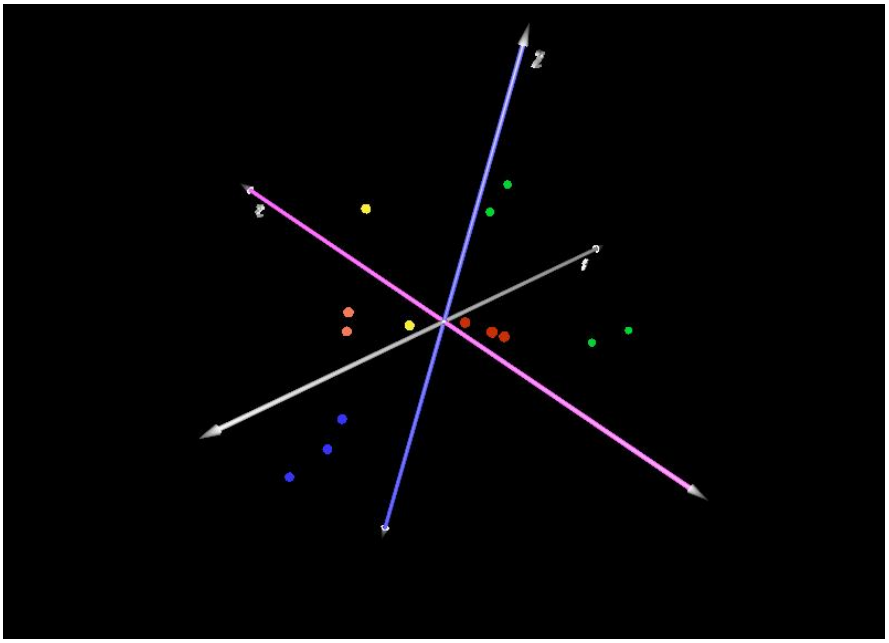
A continuación se realizó el análisis de correspondencia como método independiente para comprobar estos macrogrupos encontrados. La Figura 7 muestra los resultados en 2 dimensiones y 3 dimensiones, que representan el 56% de la inercia total.

Figura 7. **Análisis de Correspondencia (COA).** Sólo se muestran las librerías, en favor de la claridad de la imagen. Los puntos verdes representan a las librerías normales, los puntos azules a las librerías japonesas, los puntos rojos son las librerías de EEUU microdisecadas, los puntos naranjos son aquellas librerías realizadas mediante xenotrasplante, y los puntos amarillos son aquellas librerías que no tuvieron microdisección. A) Perspectiva en 2 dimensiones. B) Perspectiva tridimensional, el eje-Y se encuentra ligeramente rotado hacia abajo y a la derecha.

A)



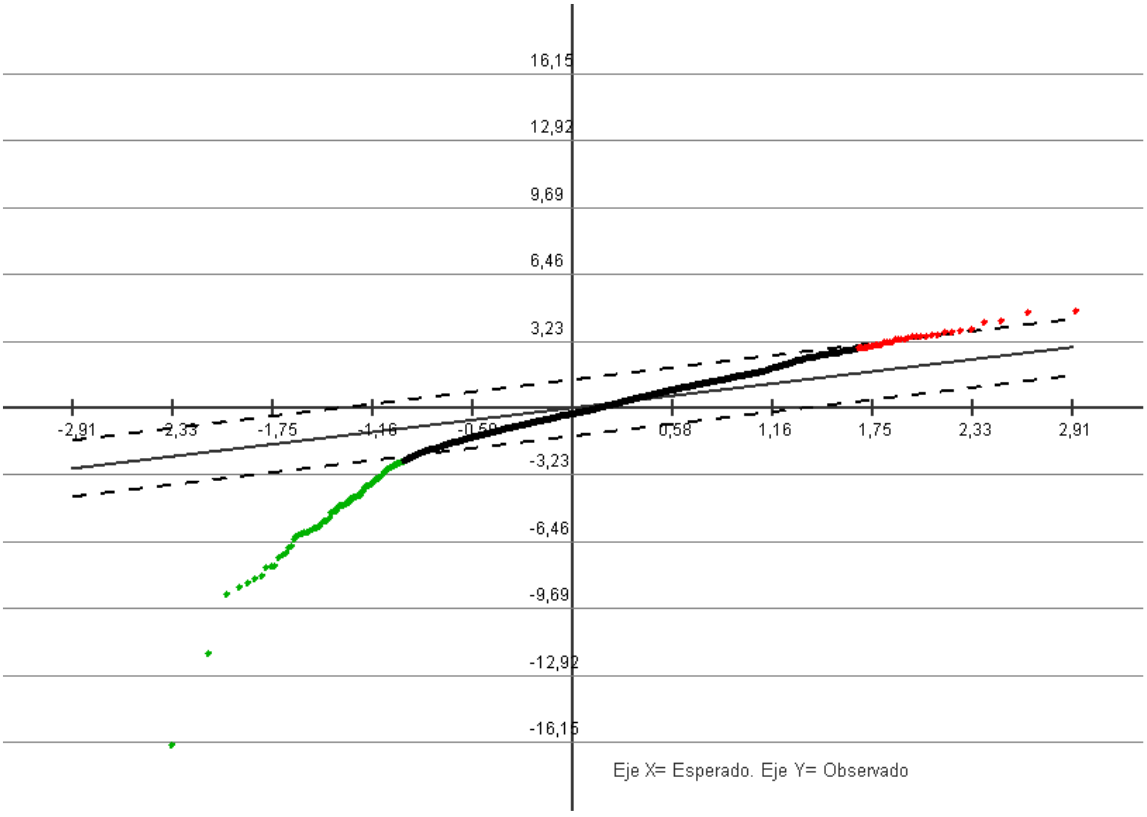
B)



3.2.2. Diferencias de expresión

El análisis de SAM se hizo con un *delta*=1,3888398, además de los criterios mencionados anteriormente. Esto reveló 90 tags diferencialmente expresados entre las librerías normales y tumorales que se muestran en la Figura 8, donde predominan aquellos tags cuya expresión cae en el tumor.

Figura 8. **SAM para el grupo de librerías Normales vs. Tumorales.** Posee un *delta* = 1,3888398, FDR ~0, 1001 permutaciones únicas y F=10. A la izquierda y en color verde, los tags significativos con mayor expresión en las librerías normales; a la derecha y en color rojo, los tags significativos con una mayor expresión en las librerías tumorales.



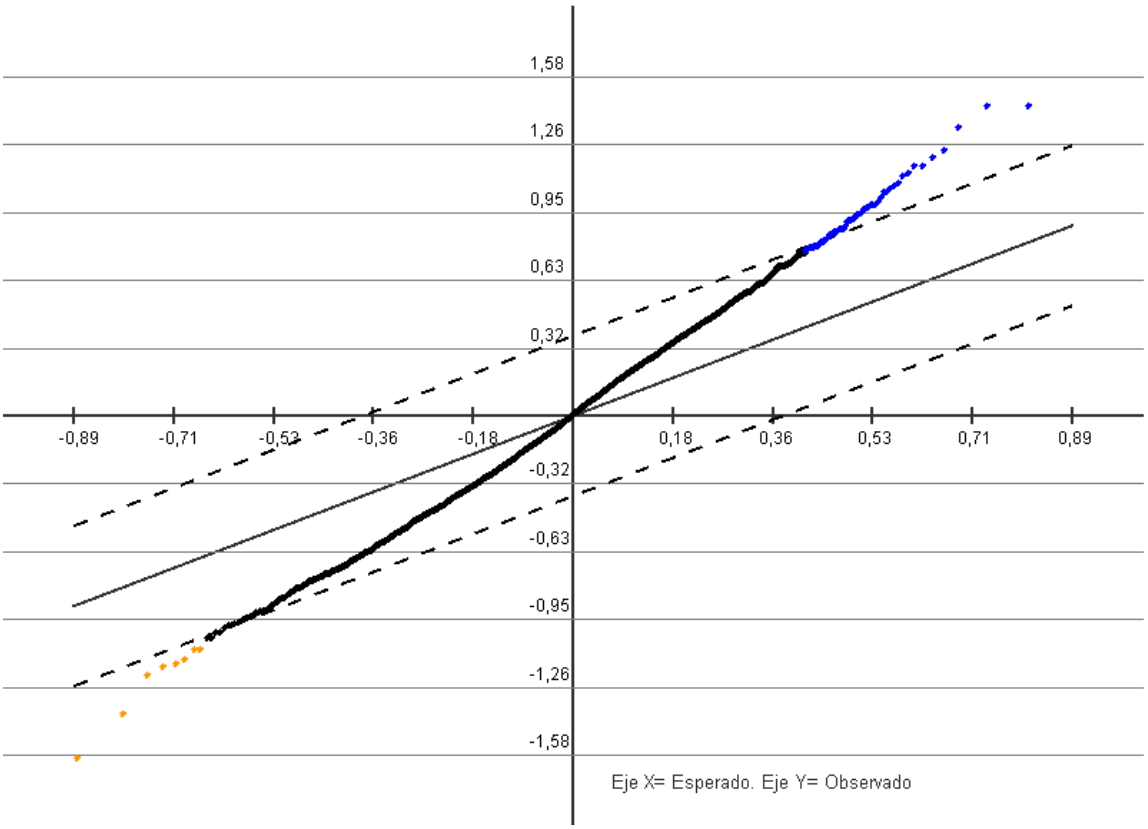
Del total de tags significativamente distintos en su nivel de expresión, la Tabla 7 detalla los cinco tags de mayor expresión en las librerías normales y los cinco tags de mayor expresión en las librerías tumorales.

Tabla 7. **Diez de los genes más significativos obtenidos por SAM entre muestras normales y tumorales.** Se muestran cinco genes con mayor expresión en tumor y cinco con mayor expresión en muestras normales.

Tags	*Símbolo del Gen*	*Nombre de Proteína*	*Nº de librerías normales*	*Promedio en normal (Tags por c/200.000 tags totales)*	*Nº de librerías tumorales*	*Promedio en tumor (Tags por c/200.000 tags totales)*
CCTGGTCCCA	KRT7	Keratin 7	1	0,68	10	121,82
TGGCCATCTG	PP1201	Transmembrane BAX inhibitor motif containing 1	1	1,09	10	40,79
ATGCGGGAGA	BCLP	Transmembrane protein 54	1	0,68	10	24,22
CAGGAGGAGT	GRP58	Glucose regulated protein, 58kDa	1	2,18	10	44,47
TGGCCCCACC	PKM2	Pyruvate kinase, muscle	1	3,95	10	75,82
GAGAACCACT	GIF	Gastric intrinsic factor (vitamin B synthesis)	4	103,22	0	0
GGAACGCAAG	ATP4B	ATPase, H+/K+ exchanging, beta polypeptide	4	172,66	2	1,23
CAGTGCCTCT	ATP5J2	ATP synthase, H+ transporting, mitochondrial F0 complex, subunit f, isoform 2	4	64,27	1	0,34
TTTAGGATGA	GDDR	Down-regulated in gastric cancer GDDR	4	530,55	3	1,98
CTGACTGTGC	ATP4A	ATPase, H+/K+ exchanging, alpha polypeptide	4	455,02	3	3,48

Luego, para ver si existían elementos discriminatorios entre los tumores de EEUU y los de Japón, se realizó nuevamente una selección de tags, esta vez basados en el criterio de "*presente en todos los tumores EEUU*" o "*presente en todos los tumores japoneses*", lo que resultó en una selección de 3.952 tags. Posteriormente se realizó un SAM con un *delta*=0,37466514, además de los criterios mencionados anteriormente. Esto resultó en 54 tags diferencialmente expresados como se muestra en la Figura 9.

Figura 9. **SAM para el grupo de tumores de EEUU vs. Japón.** Posee un *delta* = 0,37466514, FDR ~0, 120 permutaciones únicas y F=10). A la izquierda en color naranjo, los tags significativos con mayor expresión en las librerías tumorales de EEUU; a la derecha y en color azul, los tags significativos con mayor expresión en las librerías tumorales de Japón.



Del total de tags significativamente distintos en su nivel de expresión, la Tabla 8 detalla los cinco tags de mayor expresión en las librerías de tumores de EEUU y los cinco tags de mayor expresión en las librerías de Japón.

Tabla 8. **Diez de los genes más significativos obtenidos por SAM entre tumores de EEUU y Japón.** Se muestran cinco genes con mayor expresión en tumores de EEUU y cinco genes con mayor expresión en tumores de Japón.

Tags	*Símbolo del Gen*	*Nombre de Proteína*	*Nº de librerías tumorales de EEUU*	*Promedio en tumores de EEUU (Tags por c/200.000 tags totales)*	*Nº de librerías tumorales de Japón*	*Promedio en tumores de Japón (Tags por c/200.000 tags totales)*
CACCTATTGG		Similar to OK/SW-CL.16	1	0,30	3	60,85
TGATTGGTGG	PDGFRA	Platelet-derived growth factor receptor, alpha polypeptide	3	1,88	3	115,05
GGCTGGGTTT	HLX1	H2.0-like homeo box 1 (Drosophila)	2	1,04	3	59,13
CCATCGTCCT		Asociación inespecífica	3	2,39	3	77,42
TCCGTCCGGA	RPL13	Ribosomal protein L13	3	1,36	3	39,56
ACTGTATTTT	GPCR5A	G protein-coupled receptor, family C, group 5, member A	7	67,77	0	0
CTTCCTTGCC	KRT17	Keratin 17	7	220,64	0	0
TAATTTGCAT	EMP1	Epithelial membrane protein 1	7	43,26	0	0
TGGAGAATGT	ITGB1	Integrin, beta 1 (fibronectin receptor, beta polypeptide, antigen CD29 includes MDF2, MSK12)	7	27,25	0	0
CTTTATTCCA	BOMB	BH3-only member B protein	7	84,84	1	2,07

3.2.3. Discusión del estudio de SAGE

El análisis bioinformático de las librerías de SAGE permitió identificar 90 tags diferencialmente expresados entre muestras normales y tumorales, con una asociación

entre experimentos que apuntan a diferencias sutiles entre los tumores. La rama normal muestra dos *clusters* con muestras de tejido con y sin microdisección mezclados indistintamente, lo que sugiere que la heterogeneidad de un tejido normal no disminuye con la microdisección, debido probablemente a que en cada muestra, las células se encuentran en un ciclo celular distinto o ejerciendo múltiples acciones distintas, en comparación a las células tumorales que dejan de cumplir sus roles y se enfocan en la desdiferenciación y proliferación. La agrupación cercana de los tejidos tumorales microdisecados refleja un aumento en la pureza de la muestra y por tanto una mayor homogeneidad y representatividad. La cercanía de las muestras injertadas en animales (técnica denominada xenotrasplante o *xenograft*) también muestra mayor homogeneidad, pero difiere de las microdisecadas al agruparse en distintos subgrupos; esto sugiere cambios sutiles en el transcriptoma dado por un trasfondo genético distinto (el microambiente proporcionado por el animal). En cambio, las librerías sin microdisección se encontraron más dispersas, probablemente dado por una mayor contaminación y heterogeneidad de la muestra con células adyacentes no tumorales.

Tanto el *Support Clustering* como el COA, ambos análisis de carácter no supervisado, son altamente sugerentes de un perfil de expresión distinto para los 3 grupos que surgieron. Pudiera ser que la sumatoria de los diferentes niveles de expresión, que incluyen algunos genes puntuales mencionados, influye en la conocida mayor sobrevida y menor agresividad de los tumores de pacientes asiáticos como los de Japón. Sin embargo, aún no existen librerías normales japonesas de estómago para poder completar esta comparación étnica.

Al buscar diferencias específicas entre ambos tipos de tumores es posible observar que los tags que muestran diferencias significativas de expresión son en su mayoría de un mayor nivel de expresión en tumores japoneses comparados con los norteamericanos. De esta manera, parece que la expresión promedio de las muestras norteamericanas es inferior a la de las muestras japonesas, lo que puede deberse a una represión en la expresión génica más amplia.

Epidemiológicamente se han observado diferencias entre ambos tipos de tumores, pero ésta podría ser la primera vez que se proporciona una evidencia genética específica que apunta a un transcriptoma distinto entre ambos tipos de tumores. Aún falta comprobar los niveles de los genes asociados a estos tags que se han descubierto, e investigar que rol que cada uno cumple en la patogénesis del CG.

Estos resultados abren la posibilidad de que genes específicos puedan ser candidatos potenciales para terapias específicas orientadas según las características étnicas del CG.

IV. CONCLUSIONES

El primer estudio utilizó información del proyecto FONDECYT 1030130 diseñado en sus inicios de la manera clásica, mediante la investigación bibliográfica de posibles candidatos para estudios más extensos. Las características del segundo estudio, que es exclusivamente bioinformático, lograron que fuera capaz de encontrar un orden a través de enormes volúmenes de información, e indicó inmediatamente nuevos candidatos interesantes para estudiar en CG.

Ninguno de los genes relevantes que aparecieron en el segundo estudio fueron usados en el primer estudio. Los al menos 20 genes de mayor potencial según el análisis de SAGE no fueron cubiertos. La desventaja de la selección de genes para estudio por medio de literatura es que sólo se confirman genes ya conocidos, pero no aparecen nuevos.

El uso de la Bioinformática para análisis de librerías SAGE abre nuevos caminos en el entendimiento de los complejos mecanismos tumorales como un todo interrelacionado. El estudio de SAGE proporcionó candidatos nuevos de sobreexpresión y represión génica en el CG. También abrió una aproximación para identificar diferencias moleculares asociadas a aspectos étnicos del CG

A pesar de que estas herramientas han sido diseñadas para su uso en experimentos de volúmenes importantes de información, su aplicación en proyectos de números más pequeños también aporta información adicional relevante y permiten la identificación de candidatos con mejores potenciales, lo que ahorra el tiempo, esfuerzo y costo que significa probar cada uno de los candidatos individuales.

La Agrupación Jerárquica fue una herramienta valiosa en la identificación de asociaciones y de los genes más interesantes de estudiar, y la aplicación del método de *Bootstrap* consiguió distinguir las asociaciones importantes de las azarosas. El Análisis

de Correspondencia usado en el segundo estudio también reveló asociaciones y, a pesar de basarse en principios distintos, fue concordante con los resultados obtenidos mediante la Agrupación Jerárquica. El SAM fue una herramienta indispensable en la identificación precisa de los mejores candidatos, lo cual permite enfocarse directamente sobre los mejores candidatos, y probar estos mediante técnicas clásicas.

A pesar de que ya existían estudios de SAGE previos para cada una de las librerías usadas, nunca se habían analizado tantas al mismo tiempo. La aparición de indicios étnicos en el transcriptoma del CG es un elemento novedoso en este tipo de estudios.

La asociación de genes a variables clínicas del estudio de metilación de CG ha sido aceptada para publicación en la Revista Médica Chilena y se encuentra actualmente en prensa [37].

El estudio de SAGE fue presentado en formato póster en el **XXXVIII Congreso de la Sociedad de Genética de Chile** (SOCHIGEN), realizado en Noviembre del año 2005 en Puerto Varas, bajo el título de "*SAGE de Estómago Muestra Distintos Patrones Raciales*". Este trabajo también fue expuesto en formato póster (*Gene Expression Profile in Gastric Carcinoma*) en el **2006 Gastrointestinal Cancers Symposium** organizado por la ASCO (*American Association for Cancer Research*, EEUU), realizado en Enero del año 2006 en San Francisco, EEUU. En ambos eventos llamó la atención y tuvo un positivo recibimiento. Actualmente se prepara su envío para ser evaluado y publicado en una revista internacional.

V. BIBLIOGRAFÍA

1. ORELLANA C., TORRES S., DERIO L., PRIETO M. Cancer care in Chile. The Lancet Oncology, 4(11): 653-6, 2003 Nov.
2. PISANI P., PARKIN D. M., BRAY F., FERLAY J. Estimates of the worldwide mortality from 25 cancers in 1990. International Journal of Cancer, 83(1): 18-29, 1999 Sep 24.
3. SERRA I., BÁEZ S., SERRA J., CALVO A., DECINTI E. Evolución epidemiológica reciente del cáncer gástrico en Chile y el mundo. Revista Chilena de Cirugía, 49(54-63, 1997
4. DUNBIER A., GUILFORD P. Hereditary diffuse gastric cancer. Advances in Cancer Research, 83(55-65, 2001
5. BISHOP J. M. Molecular themes in oncogenesis. Cell, 64(2): 235-48, 1991 Jan 25.
6. KNUDSON A. G., JR. Hereditary cancer, oncogenes, and antioncogenes. Cancer Research, 45(4): 1437-43, 1985 Apr.
7. ISSA J. P. The epigenetics of colorectal cancer. Annals of the New York Academy of Sciences, 910(140-53; discussion 153-5, 2000 Jun.
8. BIRD A. P. CpG-rich islands and the function of DNA methylation. Nature, 321(6067): 209-13, 1986 May 15-21.
9. TOYOTA M., AHUJA N., SUZUKI H., ITOH F., OHE-TOYOTA M., IMAI K., BAYLIN S. B., ISSA J. P. Aberrant methylation in gastric cancer associated with the CpG island methylator phenotype. Cancer Research, 59(21): 5438-42, 1999 Nov 1.
10. EISEN M. B., SPELLMAN P. T., BROWN P. O., BOTSTEIN D. Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences of the United States of America, 95(25): 14863-8, 1998 Dec 8.
11. GRAUR D., WEN-HSIUNG L. Fundamentals of Molecular Evolution. Second Edition. Sunderland, MA. Sinauer Associates, 209-210, 1999.
12. FELLENBERG K., HAUSER N. C., BRORS B., NEUTZNER A., HOHEISEL J. D., VINGRON M. Correspondence analysis applied to microarray data. Proceedings of the National Academy of Sciences of the United States of America, 98(19): 10781-6, 2001 Sep 11.
13. TUSHER V. G., TIBSHIRANI R., CHU G. Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences of the United States of America, 98(9): 5116-21, 2001 Apr 24.
14. CORVALAN A., AKIBA S., VALENZUELA M. T., CUMSILLE M. A., KORIYAMA C., ARGANDONA J., BACKHOUSE C., BAL M., MENA F., PALMA M., EIZURU Y. [Clinical and molecular features of cardial gastric cancer associated to Epstein Barr virus]. Revista Médica de Chile, 133(7): 753-60, 2005 Jul.
15. LAUREN P. The Two Histological Main Types of Gastric Carcinoma: Diffuse and So-Called Intestinal-Type Carcinoma. An Attempt at a Histo-Clinical Classification. Acta Pathologica et Microbiologica Scandinavica, 64(31-49, 1965
16. CORVALAN A., KORIYAMA C., AKIBA S., EIZURU Y., BACKHOUSE C., PALMA M., ARGANDONA J., TOKUNAGA M. Epstein-Barr virus in gastric carcinoma is associated with location in the cardia and with a diffuse histology: a study in one area of Chile. International Journal of Cancer, 94(4): 527-30, 2001 Nov.
17. SAEED A. I., SHAROV V., WHITE J., LI J., LIANG W., BHAGABATI N., BRAISTED J., KLAPA M., CURRIER T., THIAGARAJAN M., STURN A., SNUFFIN

M., REZANTSEV A., POPOV D., RYLTSOV A., KOSTUKOVICH E., BORISOVSKY I., LIU Z., VINSAVICH A., TRUSH V., QUACKENBUSH J. TM4: a free, open-source system for microarray data management and analysis. Biotechniques, 34(2): 374-8, 2003 Feb.

18. EL-RIFAI W., MOSKALUK C. A., ABDRABBO M. K., HARPER J., YOSHIDA C., RIGGINS G. J., FRIERSON H. F., JR., POWELL S. M. Gastric cancers overexpress S100A calcium-binding proteins. Cancer Research, 62(23): 6823-6, 2002 Dec 1.
19. KOON N., ZAIKA A., MOSKALUK C. A., FRIERSON H. F., KNUUTILA S., POWELL S. M., EL-RIFAI W. Clustering of molecular alterations in gastroesophageal carcinomas. Neoplasia, 6(2): 143-9, 2004 Mar-Apr.
20. OUE N., HAMAI Y., MITANI Y., MATSUMURA S., OSHIMO Y., AUNG P. P., KURAOKA K., NAKAYAMA H., YASUI W. Gene expression profile of gastric carcinoma: identification of genes and tags potentially involved in invasion, metastasis, and carcinogenesis by serial analysis of gene expression. Cancer Research, 64(7): 2397-405, 2004 Apr 1.
21. BALA P., GEORGANTAS R. W., 3RD, SUDHIR D., SURESH M., SHANKER K., VRUSHABENDRA B. M., CIVIN C. I., PANDEY A. TAGmapper: a web-based tool for mapping SAGE tags. Gene, 364(123-9, 2005 Dec 30.
22. KANG G. H., LEE S., KIM W. H., LEE H. W., KIM J. C., RHYU M. G., RO J. Y. Epstein-barr virus-positive gastric carcinoma demonstrates frequent aberrant methylation of multiple genes and constitutes CpG island methylator phenotype-positive gastric carcinoma. The American Journal of Pathology, 160(3): 787-94, 2002 Mar.
23. CRAVO M., PINTO R., FIDALGO P., CHAVES P., GLORIA L., NOBRE-LEITAO C., COSTA MIRA F. Global DNA hypomethylation occurs in the early stages of intestinal type gastric carcinoma. Gut, 39(3): 434-8, 1996 Sep.
24. HENSON D. E., DITTUS C., YOUNES M., NGUYEN H., ALBORES-SAAVEDRA J. Differential trends in the intestinal and diffuse types of gastric carcinoma in the United States, 1973-2000: increase in the signet ring cell type. Archives of Pathology & Laboratory Medicine, 128(7): 765-70, 2004 Jul.
25. CORVALAN A., DING S., KORIYAMA C., CARRASCAL E., CARRASQUILLA G., BACKHOUSE C., URZUA L., ARGANDONA J., PALMA M., EIZURU Y., AKIBA S. Association of a distinctive strain of Epstein-Barr virus with gastric cancer. International Journal of Cancer, 118(7): 1736-42, 2006 Apr 1.
26. SCHNEIDER B. G., GULLEY M. L., EAGAN P., BRAVO J. C., MERA R., GERADTS J. Loss of p16/CDKN2A tumor suppressor protein in gastric adenocarcinoma is associated with Epstein-Barr virus and anatomic location in the body of the stomach. Human Pathology, 31(1): 45-50, 2000 Jan.
27. KORIYAMA C., AKIBA S., ITOH T., KIJIMA Y., SUEYOSHI K., CORVALAN A., HERRERA-GOEPFER R., EIZURU Y. Prognostic significance of Epstein-Barr virus involvement in gastric carcinoma in Japan. International Journal of Molecular Medicine, 10(5): 635-9, 2002 Nov.
28. TAN L. W., DOBROVIC A. Methylation analysis of formalin-fixed, paraffin-embedded sections using a nontoxic DNA extraction protocol. Biotechniques, 31(6): 1354, 1356-7, 2001 Dec.
29. WANG J. Y., HSIEH J. S., CHEN C. C., TZOU W. S., CHENG T. L., CHEN F. M., HUANG T. J., HUANG Y. S., HUANG S. Y., YANG T., LIN S. R. Alterations of APC, c-met, and p53 genes in tumor tissue and serum of patients with gastric cancers. The Journal of Surgical Research, 120(2): 242-8, 2004 Aug.

30. WANG J. Y., HSIEH J. S., CHANG M. Y., HUANG T. J., CHEN F. M., CHENG T. L., ALEXANDERSEN K., HUANG Y. S., TZOU W. S., LIN S. R. Molecular detection of APC, K- ras, and p53 mutations in the serum of colorectal cancer patients as circulating biomarkers. World Journal of Surgery, 28(7): 721-6, 2004 Jul.
31. THEUER C. P., KUROSAKI T., ZIOGAS A., BUTLER J., ANTON-CULVER H. Asian patients with gastric carcinoma in the United States exhibit unique clinical features and superior overall and cancer specific survival rates. Cancer, 89(9): 1883-92, 2000 Nov 1.
32. GILL S., SHAH A., LE N., COOK E. F., YOSHIDA E. M. Asian ethnicity-related differences in gastric cancer presentation and outcome among patients treated at a canadian cancer center. Journal of Clinical Oncology, 21(11): 2070-6, 2003 Jun 1.
33. EDGAR R., DOMRACHEV M., LASH A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research, 30(1): 207-10, 2002 Jan 1.
34. BARRETT T., SUZEK T. O., TROUP D. B., WILHITE S. E., NGAU W. C., LEDOUX P., RUDNEV D., LASH A. E., FUJIBUCHI W., EDGAR R. NCBI GEO: mining millions of expression profiles--database and tools. Nucleic Acids Research, 33(Database issue): D562-6, 2005 Jan 1.
35. RIGGINS G. J., STRAUSBERG R. L. Genome and genetic resources from the Cancer Genome Anatomy Project. Human Molecular Genetics, 10(7): 663-7, 2001 Apr.
36. TUTEJA R., TUTEJA N. Serial analysis of gene expression (SAGE): application in cancer research. Medical Science Monitor, 10(6): RA132-40, 2004 Jun.
37. ZAVALA L., LUENGO V., OSSANDON F., RIQUELME E., BACKHOUSE E., PALMA M., ARGANDONA J., CUMSILLE M. A., CORVALAN A. Identificación de asociaciones clínico-patológicas e hipermetilación de genes supresores de tumores en Cáncer Gástrico Difuso a través de análisis de Hierarchical Clustering. Revista Médica de Chile, 2006 En prensa.