# SVMRFE-GSEA: Módulo de Aprendizaje Supervisado en GSEA

**Javier Murillo, Elizabeth Tapia, Serge Guillaume, Pilar Bulacio.**
***Bioinformática, Cifasis, Conicet | Cemagref Montpellier***
***Rosario, Argentina | Montpellier, Francia***
***E-mail: {murillo,bulacio,tapia}@cifasis-conicet.gov.ar***

**Resumen**
*El presente trabajo presenta un módulo de software que amplia y fortalece las capacidades de GSEA. El principal objetivo de SVMRFE-GSEA es mejorar el análisis de selección de genes y hacerlo más robusto. Para ello, SVMRFE-GSEA completa los resultados de las métricas de GSEA con un aprendizaje automatizado que considera la interacción entre genes. Resultados experimentales sobre el conjunto de datos Diabetes y Leukemia muestran un primer aporte de la herramienta.*

**Palabras Clave**

GSEA, Análisis conjunto de genes, aprendizaje automatizado, SVM, RFE, Support Vector Machine.

## I - Introducción

Dentro de la investigación genómica, el desarrollo de herramientas robustas para el análisis de ADN mediante microarreglos es de vital importancia.

El objetivo principal de GSEA (Gene Set Enrichment Analysis) [1,2] es determinar si los miembros de un conjunto de genes están correlacionados con su clase genotípica distintiva. GSEA introduce una estrategia analítica que detecta cambios modestos pero coordinados en la expresión del grupo de genes. Con este fin, GSEA ordena los genes en una lista rankeada de acuerdo a su expresión diferencial entre clases y la lista resultante, es analizada se compara con el conjunto de genes predefinidos que forman parte de una recopilación hecha por los desarrolladores de GSEA.
SVMRFE-GSEA es un módulo de software que amplía el análisis de GSEA mediante un método de aprendizaje automatizado. El módulo aporta a GSEA un mecanismo más robusto para realizar el análisis de los datos, e introduce en éste técnicas exitosas de investigaciones previas [3,4,5].

Para fortalecer la lista rankeada obtenida mediante GSEA, SVMRFE-GSEA utiliza Support Vector Machine—Recursive Feature Elimination (SVM-RFE), que es una técnica que se ha utilizado por primera vez de manera exitosa para el estudio del cáncer en [6].

Como resultado SVMRFE-GSEA devuelve una lista rankeada “alternativa” a la devuelta por las métricas de GSEA, para luego combinarlas en una lista única y que ésta pueda ser analizada con todo el poder estadístico original de software.

El paper está estructurado de la siguiente forma: luego de esta introducción, la sección 2 describe el método SVMRFE-GSEA implementado. En la sección 3 se muestran los módulos que se modificaron en GSEA, y se explican las clases introducidas. La sección 4 compara los resultados de simulaciones SVMRFE-GSEA y en GSEA con los datos Diabetes y Leukemia [9,10]. Finalmente, en la sección 5, se exponen las conclusiones.

## II. SVMRFE-GSEA

La fortaleza de GSEA reside en su componente estadístico sólido para validar resultados y en su base de datos de conjuntos predefinidos de genes. Nuestro aporte viene dado por la modificación del método de construcción de la lista rankeada de los genes del conjunto de datos.

El método propuesto se basa en la idea de que los genes funcionan en forma conjunta. Mientras GSEA utiliza métricas para ponderar cada gen del conjunto de datos de acuerdo a la información que este aporta para la distinción de las clases genotípicas; SVMRFE-GSEA construye la lista rankeada mediante SVM-RFE que tiene en cuenta la interacción entre los genes.

Sin extendernos demasiado SVM-RFE genera una lista rankeada a través de los siguientes pasos:

(1) Comienzo: Lista Rankeada final Q = [ ] ; Conjunto seleccionado S = [1..d] ;

(2) Repetir hasta que toda la lista haya sido rankeada:

(a) Entrenar una SVM con todos los datos de entrenamiento y variables en S;
(b) Calcula los pesos de cada variable en S de a cuerdo a SVM
(c) Busca la variable e con ranking mínimo;
(d) Actualiza Q : Q = Q[e;Q] ;
(e) Actualiza S : S = S - [e] ;

(3) Salida: Lista rankeada Q

Los datos utilizados para realizar la eliminación recursiva son 0.2, es decir que se elimina 20% de los datos en cada iteración, el parámetro C de la SVM es de 1 y se usaron 4 folds.

Para hacer más robusto a GSEA, se construye una lista $L_m$ que se producto de la solución original de la lista rankeada propuesta por GSEA y la nueva lista generada mediante SVM-RFE,

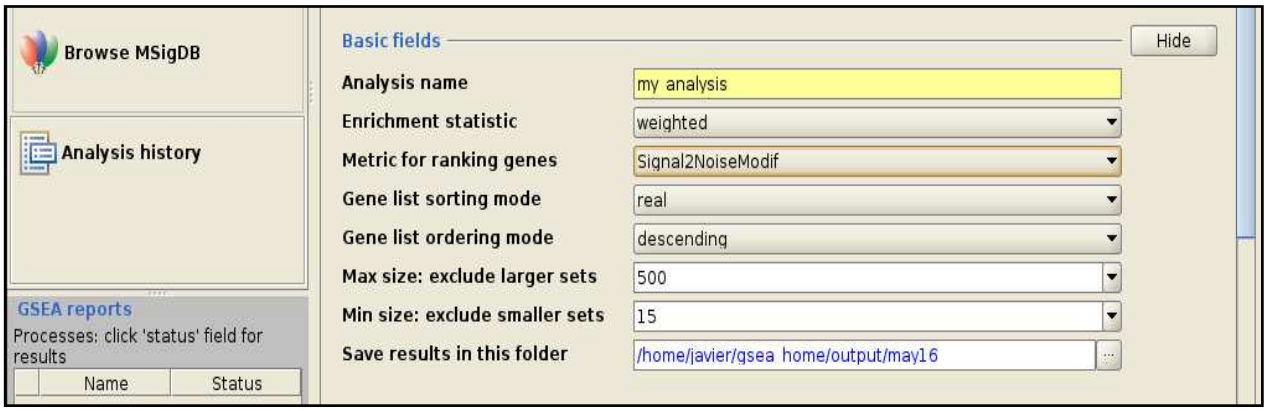
$$L_m = Q \times L \qquad \text{(Ecuación 1)}$$

Para generar la lista Q se eligió un método de backward-elimination, es decir, ir eliminando genes del conjunto total, en lugar de un típico forward-selection en el cual se realizaría el procedimiento inverso, pues, como se sabe, los genes actúan en conjunto y el hecho de ir agregando de a uno y evaluando su desempeño, hace que, posiblemente, se pierda esta aporte conjunto, cosa que no pasa en backward-elimination [5].

Esta nueva manera de obtener la lista rankeada sigue la línea de GSEA, pues hace hincapié en que los genes actúan de manera conjunta, al hacer que la lista rankeada se construya mediante un algoritmo que tenga en cuenta el trabajo conjunto de los genes[1].

## III. Módulos modificados

Se modificó el módulo xtool.gsea donde se agregó una nueva métrica “Signal2NoiseModif” y los módulos Metrics.java y Math.java para implementarla. Se agrego la clase Mlearning que implementa funciones de eliminación de genes y el método de aprendizaje supervisado para el cálculo del error.




## IV. Resultados

Se utilizaron dos conjuntos de datos: Diabetes, Transcriptional profiles of smooth muscle biopsies of diabetic and normal individuals [11]; Leukemia, Transcriptional profiles from leukemias - ALL and AML [12].

Comparación de los datos de Diabetes [9] en SVMRFE-GSEA y GSEA.

*Table: Gene sets enriched in phenotype NGT (17 samples)* ***[plain text format]***

| | GS | GS DESC | SIZE | ES | NES | NOM p-val | FDR q-val | FWER p-val |
|---|---|---|---|---|---|---|---|---|
| 1 | P53_DOWN | Details ... | 18 | 0.69 | 2.09 | 0.000 | 0.009 | 0.011 |
| 2 | VOXPHOS | Details ... | 83 | 0.62 | 1.95 | 0.007 | 0.030 | 0.058 |
| 3 | Electron_Transport_Chain | Details ... | 85 | 0.61 | 1.89 | 0.018 | 0.044 | 0.122 |
| 4 | GNF_FEMALE_GENES | Details ... | 85 | 0.43 | 1.56 | 0.049 | 0.831 | 0.866 |
| 5 | human_mitoDB_6_2002 | Details ... | 408 | 0.35 | 1.50 | 0.049 | 0.968 | 0.933 |
| 6 | mitochondr | Details ... | 431 | 0.32 | 1.39 | 0.081 | 1.000 | 0.990 |
| 7 | MAP00190_Oxidative_phosphorylation | Details ... | 43 | 0.40 | 1.38 | 0.149 | 1.000 | 0.994 |
| 8 | deathPathway | Details ... | 32 | 0.34 | 1.32 | 0.096 | 1.000 | 0.998 |
| 9 | tollPathway | Details ... | 33 | 0.36 | 1.30 | 0.107 | 1.000 | 0.998 |
| 10 | calcineurinPathway | Details ... | 18 | 0.46 | 1.27 | 0.217 | 1.000 | 1.000 |

*Table: Gene sets enriched in phenotype DMT (17 samples)* ***[plain text format]***

| | GS | GS DESC | SIZE | ES | NES | NOM p-val | FDR q-val | FWER p-val |
|---|---|---|---|---|---|---|---|---|
| 1 | MAP00120_Bile_acid_biosynthesis | Details ... | 22 | -0.50 | -1.73 | 0.007 | 0.996 | 0.461 |
| 2 | ucalpainPathway | Details ... | 15 | -0.61 | -1.65 | 0.027 | 1.000 | 0.690 |
| 3 | ST_T_Cell_Signal_Transduction | Details ... | 42 | -0.43 | -1.59 | 0.026 | 1.000 | 0.830 |
| 4 | MAP00561_Glycerolipid_metabolism | Details ... | 46 | -0.41 | -1.53 | 0.019 | 1.000 | 0.920 |
| 5 | ST_MONOCYTE_AD_PATHWAY | Details ... | 27 | -0.45 | -1.51 | 0.036 | 1.000 | 0.941 |
| 6 | nkcellsPathway | Details ... | 18 | -0.48 | -1.51 | 0.039 | 1.000 | 0.942 |
| 7 | MAP00500_Starch_and_sucrose_metabolism | Details ... | 21 | -0.46 | -1.49 | 0.046 | 1.000 | 0.952 |
| 8 | SA_B_CELL_RECEPTOR_COMPLEXES | Details ... | 24 | -0.50 | -1.48 | 0.043 | 1.000 | 0.954 |
| 9 | pitx2Pathway | Details ... | 16 | -0.56 | -1.43 | 0.099 | 1.000 | 0.985 |
| 10 | mcalpainPathway | Details ... | 23 | -0.45 | -1.41 | 0.094 | 1.000 | 0.988 |

Los resultados obtenidos para Diabetes mediante SVM-RFE, con iguales parámetros que GSEA son:

*Table: Gene sets enriched in phenotype NGT (17 samples)* ***[plain text format]***

| | GS follow link to MSigDB | GS DETAILS | SIZE | ES | NES | NOM p-val | FDR q-val | FWER p-val |
|---|---|---|---|---|---|---|---|---|
| 1 | P53_DOWN | Details ... | 17 | 0.72 | 1.67 | 0.015 | 0.337 | 0.322 |
| 2 | VOXPHOS | Details ... | 83 | 0.73 | 1.64 | 0.005 | 0.268 | 0.438 |
| 3 | ELECTRON_TRANSPORT_CHAIN | Details ... | 85 | 0.72 | 1.57 | 0.016 | 0.447 | 0.702 |
| 4 | INSULIN_2F_DOWN | Details ... | 31 | 0.59 | 1.52 | 0.036 | 0.538 | 0.827 |
| 5 | NFKB_REDUCED | Details ... | 21 | 0.58 | 1.38 | 0.101 | 1.000 | 0.983 |
| 6 | GNF_FEMALE_GENES | Details ... | 85 | 0.55 | 1.37 | 0.090 | 1.000 | 0.983 |
| 7 | IL1RPATHWAY | Details ... | 32 | 0.47 | 1.35 | 0.092 | 1.000 | 0.995 |
| 8 | TNFR2PATHWAY | Details ... | 18 | 0.51 | 1.34 | 0.096 | 1.000 | 0.995 |
| 9 | FETAL_LIVER_HS_ENRICHED_TF_JP | Details ... | 81 | 0.44 | 1.33 | 0.068 | 1.000 | 0.998 |
| 10 | HUMAN_MITODB_6_2002 | Details ... | 405 | 0.45 | 1.33 | 0.071 | 1.000 | 0.998 |

Table: Gene sets enriched in phenotype DMT (17 samples) [plain text format]

| | GS follow link to MSigDB | GS DETAILS | SIZE | ES | NES | NOM p-val | FDR q-val | FWER p-val |
|---|---|---|---|---|---|---|---|---|
| 1 | ST_MONOCYTE_AD_PATHWAY | Details ... | 27 | -0.63 | -1.75 | 0.000 | 0.209 | 0.153 |
| 2 | SA_B_CELL_RECEPTOR_COMPLEXES | Details ... | 24 | -0.71 | -1.73 | 0.000 | 0.134 | 0.188 |
| 3 | MAP00120_BILE_ACID_BIOSYNTHESIS | Details ... | 22 | -0.61 | -1.64 | 0.005 | 0.295 | 0.440 |
| 4 | ST_T_CELL_SIGNAL_TRANSDUCTION | Details ... | 42 | -0.56 | -1.60 | 0.005 | 0.350 | 0.543 |
| 5 | INTEGRINPATHWAY | Details ... | 33 | -0.67 | -1.60 | 0.005 | 0.305 | 0.572 |
| 6 | MCALPAINPATHWAY | Details ... | 23 | -0.65 | -1.54 | 0.030 | 0.509 | 0.788 |
| 7 | UCALPAINPATHWAY | Details ... | 15 | -0.72 | -1.52 | 0.032 | 0.566 | 0.842 |
| 8 | ST_GRANULE_CELL_SURVIVAL_PATHWAY | Details ... | 27 | -0.54 | -1.50 | 0.015 | 0.623 | 0.902 |
| 9 | MAP00150_ANDROGEN_AND_ESTROGEN_METABOLISM | Details ... | 17 | -0.61 | -1.47 | 0.066 | 0.704 | 0.938 |
| 10 | ST_FAS_SIGNALING_PATHWAY | Details ... | 62 | -0.47 | -1.44 | 0.020 | 0.799 | 0.965 |

Como puede verse en estas graficas, para el caso del genotipo NGT, los 3 primeros genes más relevantes son coincidentes, y GNF_Female_Genes y Human_mitoDB_6_2002 aparecen en distinto orden. Se observa además que los valores de FDR (false discovery rate) para SVM-RFE se encuentran por encima de 0.25, esto puede deberse a la no optimización de los parámetros del método SVM o a una tasa de eliminación alta (0.2). Actualmente se está trabajando en el ajuste de estos parámetros.

Para el caso del genotipo DMT, a pesar de no coincidir los primeros lugares, se encuentran coincidencia en distintos órdenes, SVM-RFE encuentra dos genes con un FDR menor a 0.25.

Comparación de los datos de Leukemia [10] en SVMRFE-GSEA y GSEA. [10]

Table: Gene sets enriched in phenotype ALL (24 samples) [plain text format]

| | GS | GS DESC | SIZE | ES | NES | NOM p-val | FDR q-val | FWER p-val |
|---|---|---|---|---|---|---|---|---|
| 1 | chr6q21 | Details ... | 31 | 0.66 | 1.99 | 0.002 | 0.039 | 0.032 |
| 2 | chr13q14 | Details ... | 31 | 0.57 | 1.86 | 0.000 | 0.090 | 0.133 |
| 3 | chr5q31 | Details ... | 59 | 0.50 | 1.85 | 0.000 | 0.065 | 0.141 |
| 4 | chr17q23 | Details ... | 39 | 0.53 | 1.79 | 0.011 | 0.094 | 0.234 |
| 5 | chr14q32 | Details ... | 64 | 0.47 | 1.75 | 0.008 | 0.112 | 0.306 |
| 6 | chr1q42 | Details ... | 32 | 0.49 | 1.67 | 0.033 | 0.190 | 0.481 |
| 7 | chr3q13 | Details ... | 28 | 0.46 | 1.54 | 0.042 | 0.455 | 0.735 |
| 8 | chr6 | Details ... | 456 | 0.32 | 1.54 | 0.007 | 0.417 | 0.745 |
| 9 | chr10q24 | Details ... | 43 | 0.45 | 1.52 | 0.015 | 0.416 | 0.765 |
| 10 | chr10p15 | Details ... | 17 | 0.49 | 1.48 | 0.054 | 0.520 | 0.829 |

Table: Gene sets enriched in phenotype AML (24 samples) [plain text format]

| | GS | GS DESC | SIZE | ES | NES | NOM p-val | FDR q-val | FWER p-val |
|---|---|---|---|---|---|---|---|---|
| 1 | chr15q25 | Details ... | 29 | -0.43 | -1.44 | 0.055 | 1.000 | 0.861 |
| 2 | chr1q24 | Details ... | 30 | -0.42 | -1.40 | 0.076 | 1.000 | 0.889 |
| 3 | chr7q36 | Details ... | 25 | -0.40 | -1.27 | 0.172 | 1.000 | 0.963 |
| 4 | chr19p13 | Details ... | 211 | -0.28 | -1.26 | 0.158 | 1.000 | 0.967 |
| 5 | chr15q22 | Details ... | 20 | -0.38 | -1.24 | 0.192 | 1.000 | 0.971 |
| 6 | chr17q25 | Details ... | 61 | -0.31 | -1.22 | 0.187 | 1.000 | 0.979 |
| 7 | chr4q13 | Details ... | 37 | -0.34 | -1.21 | 0.203 | 1.000 | 0.982 |
| 8 | chr2q35 | Details ... | 31 | -0.35 | -1.20 | 0.247 | 1.000 | 0.983 |
| 9 | chr16p11 | Details ... | 40 | -0.36 | -1.20 | 0.267 | 1.000 | 0.983 |
| 10 | chr7q34 | Details ... | 17 | -0.41 | -1.19 | 0.263 | 1.000 | 0.983 |

Los resultados obtenidos para Leukemia mediante SVM-RFE, con iguales parámetros que GSEA son:

Table: Gene sets enriched in phenotype 0 (24 samples) [pla

| | GS follow link to MSigDB | GS DETAILS | SIZE | ES | NES | NOM p-val | FDR q-val | FWER p-val |
|---|---|---|---|---|---|---|---|---|
| 1 | CHR6Q21 | Details ... | 30 | 0.79 | 1.88 | 0.000 | 0.037 | 0.030 |
| 2 | CHR14Q32 | Details ... | 58 | 0.61 | 1.68 | 0.000 | 0.289 | 0.335 |
| 3 | CHR10P15 | Details ... | 16 | 0.69 | 1.67 | 0.015 | 0.208 | 0.357 |
| 4 | CHR17Q23 | Details ... | 39 | 0.65 | 1.64 | 0.005 | 0.214 | 0.433 |
| 5 | CHR10Q25 | Details ... | 23 | 0.66 | 1.62 | 0.005 | 0.213 | 0.493 |
| 6 | CHR13Q14 | Details ... | 30 | 0.60 | 1.62 | 0.020 | 0.189 | 0.510 |
| 7 | CHR5Q31 | Details ... | 58 | 0.54 | 1.57 | 0.005 | 0.269 | 0.637 |
| 8 | CHR3Q26 | Details ... | 24 | 0.57 | 1.56 | 0.014 | 0.249 | 0.652 |
| 9 | CHR6Q25 | Details ... | 22 | 0.60 | 1.52 | 0.044 | 0.324 | 0.757 |
| 10 | CHR10Q24 | Details ... | 39 | 0.53 | 1.52 | 0.000 | 0.303 | 0.770 |

Table: Gene sets enriched in phenotype 1 (24 samples) [plair

| | GS follow link to MSigDB | GS DETAILS | SIZE | ES | NES | NOM p-val | FDR q-val | FWER p-val |
|---|---|---|---|---|---|---|---|---|
| 1 | CHR7Q34 | Details ... | 17 | -0.64 | -1.59 | 0.023 | 0.755 | 0.585 |
| 2 | CHR1Q24 | Details ... | 29 | -0.54 | -1.48 | 0.045 | 0.977 | 0.803 |
| 3 | CHR15Q25 | Details ... | 29 | -0.52 | -1.48 | 0.033 | 0.676 | 0.813 |
| 4 | CHR4Q13 | Details ... | 37 | -0.52 | -1.40 | 0.138 | 0.919 | 0.915 |
| 5 | CHR7Q31 | Details ... | 27 | -0.54 | -1.34 | 0.123 | 1.000 | 0.962 |
| 6 | CHR2Q35 | Details ... | 30 | -0.49 | -1.33 | 0.078 | 0.978 | 0.967 |
| 7 | CHR11Q24 | Details ... | 19 | -0.50 | -1.31 | 0.110 | 0.929 | 0.980 |
| 8 | CHR15Q22 | Details ... | 20 | -0.48 | -1.26 | 0.183 | 1.000 | 0.988 |
| 9 | CHR11P13 | Details ... | 15 | -0.52 | -1.24 | 0.214 | 1.000 | 0.990 |
| 10 | CHR20P11 | Details ... | 17 | -0.50 | -1.24 | 0.189 | 0.938 | 0.990 |

Para el caso del genotipo ALL, llamado 0 por SVM-RFE, puede observarse que la primer posición, lo que hace que el gen sea el más importante según el software, para la distinción de clase es coincidente, con un FDR menor a 0.25 en ambos casos, de los restantes cinco genes encontrados por GSEA con un FDR menor a 0.25, cuatro de ellos figuran en las primeras 10 posiciones de SVM-RFE.

Nuevamente los resultados no son idénticos, pero similares, y una mayor coincidencia podría lograrse mediante un mejor ajuste de los parámetros del modelo de eliminación.

Los resúmenes comparativos son:

**Diabetes:**

**SVM-RFE**

Enrichment in phenotype: NGT (17 samples)

- 115 / 318 gene sets are upregulated in phenotype **NGT**
- 0 gene sets are significant at FDR < 25%
- 1 gene sets are significantly enriched at nominal pvalue < 1%
- 4 gene sets are significantly enriched at nominal pvalue < 5%

Enrichment in phenotype: DMT (17 samples)

- 203 / 318 gene sets are upregulated in phenotype **DMT**
- 2 gene sets are significantly enriched at FDR < 25%
- 5 gene sets are significantly enriched at nominal pvalue < 1%
- 10 gene sets are significantly enriched at nominal pvalue < 5%

Gene markers for the NGT *versus* DMT comparison

- The dataset has 15054 features (genes)
- # of markers for phenotype **NGT**: 7991 (53.1% ) with correlation area 56.3%
- # of markers for phenotype **DMT**: 7063 (46.9% ) with correlation area 43.7%

**GSEA**

Enrichment in phenotype: NGT (17 samples)

- 124 / 318 gene sets had enrichment in phenotype **NGT**
- 8 gene set(s) are significant at nominal pvalue < 5%
- 3 gene set(s) are significant at FDR < 25%

Enrichment in phenotype: DMT (17 samples)

- 194 / 318 gene sets had enrichment in phenotype **DMT**
- 12 gene set(s) are significant at nominal pvalue < 5%
- 0 gene set(s) are significant at FDR < 25%

Gene markers for the NGT *versus* DMT comparison

- The dataset had 15056 features (genes)
- # of markers for phenotype **NGT**: 7993 (53.1% ) with correlation area 53.4%
- # of markers for phenotype **DMT**: 7063 (46.9% ) with correlation area 46.6%

**Leukemia:**

Enrichment in phenotype: 0 (24 samples)

- 119 / 181 gene sets are upregulated in phenotype **0**
- 6 gene sets are significant at FDR < 25%
- 6 gene sets are significantly enriched at nominal pvalue < 1%
- 15 gene sets are significantly enriched at nominal pvalue < 5%

Enrichment in phenotype: 1 (24 samples)

- 62 / 181 gene sets are upregulated in phenotype **1**
- 0 gene sets are significantly enriched at FDR < 25%
- 0 gene sets are significantly enriched at nominal pvalue < 1%
- 3 gene sets are significantly enriched at nominal pvalue < 5%

Gene markers for the 0 *versus* 1 comparison

- The dataset has 10053 features (genes)
- # of markers for phenotype **0**: 5460 (54.3% ) with correlation area 57.4%
- # of markers for phenotype **1**: 4593 (45.7% ) with correlation area 42.6%

Enrichment in phenotype: ALL (24 samples)

- 124 / 182 gene sets had enrichment in phenotype **ALL**
- 24 gene set(s) are significant at nominal pvalue < 5%
- 6 gene set(s) are significant at FDR < 25%

Enrichment in phenotype: AML (24 samples)

- 58 / 182 gene sets had enrichment in phenotype **AML**
- 2 gene set(s) are significant at nominal pvalue < 5%
- 0 gene set(s) are significant at FDR < 25%

Gene markers for the ALL *versus* AML comparison

- The dataset had 10056 features (genes)
- # of markers for phenotype **ALL**: 5460 (54.3% ) with correlation area 60.5%
- # of markers for phenotype **AML**: 4596 (45.7% ) with correlation area 39.5%

## V. Conclusión

En esta versión inicial del software puede verse que los resultados obtenidos por SVM-RFE son coincidentes con GSEA. Este hecho refuerza los resultados del mismo, pues ahora no solo se tiene en cuenta la métrica individual, sino que también se hace uso de un método de aprendizaje supervisado que tiene en cuenta el aporte conjunto de los genes. Las dos fuentes de información son combinadas lo que hace que los resultados finales sean más robustos. Se está trabajando para ajustar los parámetros del modelo de manera de obtener mejores resultados.

## Referencias

**[1]**Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P.: Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome- Wide Expression Profiles. PNAS 102(43), 15545–15550 (2005)

**[2]**Mootha, V. K., Lindgren, C. M., Eriksson, K. F.,Subbramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., et al. (2003) Nat. Genet. 34, 267-273.

**[3]**Xin Zhou and David P. Tuck, MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on DNA microarray data, Department of Pathology, Yale University School of Medicine, New Haven, Connecticut 06510, USA , Bioinformatics , 2007

**[4]**Xuegong Zhang, Lu Xin , Qian Shi, Xiu-qin Xu, Hon-chiu E Leung , Lyndsay N Harris , James D Iglehart, Alexander Miron , Jun Liu S , H Wong Wing . Recursivo SVM característica de selección y clasificación de muestras para la espectrometría de masa y de datos de microarrays , BMC Bioinformatics, 2006; 7: 197-197

**[5]**Kaibo Duan and Jagath C. Rajapakse, SVM-RFE Peak Selection for cancer classification with mass spectrometry data, Bioinformatics research Center Nanyang Technological University, Singapure.

**[6]**Isabelle Guyon+, Jason Weston+, Stephen Barnhill, M.D.+ and Vladimir Vapnik*, Gene Selection for Cancer Classification using Support Vector Machines , +Barnhill Bioinformatics, Savannah, Georgia, USA, * AT&T Labs, Red Bank, New Jersey, USA .

**[7]**Justin Bedo, Karsten Borgwardt, Arthur Gretton and Alex Smola, Data Mining in bioinformatics, NICTA Statistical Machine Learning Program, Australia (2010).

**[8]**GSEA web repository, Leukemia, Diabetes and P53 dataset, http://www.broadinstitute.org /gsea/datasets.jsp

**[9]**http://www.broadinstitute.org/gsea/resources/gsea_pnas_results/diabetes_C2.Gsea/index.html

**[10]**http://www.broadinstitute.org/gsea/resources/gsea_pnas_results/leukemia_C1.Gsea/index.html

**[11]**Mootha et al. (2003) Nat Genet 34(3): 267-73 , 15505 genes y 34 ejemplos.

**[12]**Armstrong et al. (2002) Nat Genet 30(1): 41-7, 10060 genes y 48 ejemplos.