Se realizaron tres estudios, en todos los casos se realizaron los siguientes pasos.

1. Eliminación de outliers. Teniendo en cuenta los grupos de muestras, se eliminaron todos aquellos valores que estaban por encima de 1.5 X 3er cuartil y por debajo de 1.5 X 1er cuartil.
2. En este estudio no había valores perdidos. Los outliers removidos fueron sustituidos por la mediana de los valores de un mismo grupo.
3. Los valores se escalaron a la escala [0,1].
4. Se ejecutó un test de Wilcoxon para detectar diferencias significativas en las expresiones de factores entre los diferentes grupos.
5. Se detectaron correlaciones lineales entre par de factores. Solo se generó el scatter-plot de correlaciones mayores que 0.9 y con una p-value menor que 0.01.
6. Se ejecutaron varios algoritmos de selección de características supervisados de tipo filter, es decir que no dependen de un clasificador. Se genera un ranking de importancia de características.
7. Se ejecutaron varios modelos para clasificar los grupos determinando cual subconjunto de factores produce un modelo con mayor AUC. En este caso se añadió un modelo LASSO que hace una selección de características implícita.
8. Se ejecutó un algoritmo de clúster jeráquico determinando el mejor subconjunto de factores.

Para todos los casos se hizo una validación cruzada Leave-One-Out-Cross-Validation, lo cual permite hacer una estimación bastante precisa. Se intentó encontrar la mejor configuración de parámetros para cada algoritmo.

**Resultados en cada estudio**

NormalesP1 vs NormalesP2

1. Existen diferencias significativas en los niveles de expresión de los siguientes factores

|  |  |  |
| --- | --- | --- |
| p-values < 0.05 | p-values < 0.01 | p-values < 0.001 |
| RBM22 | CA150 | SKIP |
| SC35 | FBP11 | U2 |
| SF3BTV1 | nsR100 |  |
| TIA1 | PSF |  |
| U1 | RBM45 |  |
|  | U2AF1 |  |
|  | U4ATAC |  |
|  | U5 |  |

1. Según los algoritmos de selección de características ejecutados, en promedio los factores que discriminan mejor a los grupos son (ordenados de mayor a menor relevancia):

SKIP 0.012987012987012988

U2 0.04761904761904762

U4atac 0.07359307359307361

SC35 0.0778787878787879

U5 0.12229437229437232

U2AF1 0.14670995670995673

CA150 0.17160173160173162

Estos factores están dentro del grupo que fueron encontrados como más significativos.

1. Se encontraron más de mil combinaciones de factores que producen modelos de clasificación con AUC bastante altos (0.9-1.0).

También se ejecutó un algoritmo LASSO, y el resultado fue de un AUC igual a 1.0 y según este modelo los factores más importantes son SKIP, U2 y PSF, el resto de factores no hacen falta para lograr una separación perfecta de los grupos.

1. Se encontraron más de mil combinaciones de factores que permiten hacer un clúster jerárquico con un AUC de 1.

NormalesP1 vs TumoralesP1

1. Existen diferencias significativas en los niveles de expresión de los siguientes factores

|  |  |  |
| --- | --- | --- |
| p-values < 0.05 | p-values < 0.01 | p-values < 0.001 |
| CUGBP | nsR100 | NOVA1 |
| RAVER1 | SRSF1 | SAM68TV1 |
| RBM17 | U6 | SRSF9 |
| SC35 |  |  |
| SRSF5 |  |  |
| SRSF10 |  |  |
| TRA2A |  |  |
| U1 |  |  |
| U4 |  |  |
| U5 |  |  |
| U11 |  |  |
| U12 |  |  |

1. Existe una alta correlación entre los factores:

* SAM68TV1 y SRSF1
* SAM68TV1 y SRSF9
* SAM68TV1 y TRA2A
* SAM68TV1 y U1
* SRSF1 y U4
* SRSF9 y U1
* SRSF9 y U11
* TRA2A y U11
* U11 y U5
* U1 y U11
* U1 y U5
* U4 y U6

Estos factores por separado se expresan significativamente diferente en ambos grupos, y esta alta correlación indica que son factores con una distribución conjunta relevante.

1. Según los algoritmos de selección de características ejecutados, en promedio los factores que discriminan mejor a los grupos son (ordenados de mayor a menor relevancia):

SRSF9 0.0

NOVA1 0.07224489795918368

U4 0.09448979591836737

SRSF5 0.14959183673469392

SAM68TV1 0.21346938775510202

SRSF1 0.22306122448979587

U6 0.2857142857142857

RAVER1 0.36734693877551017

RBM17 0.36734693877551017

nSR100 0.47020408163265304

Estos factores están dentro del grupo que fueron encontrados como más significativos.

1. No se encontró ninguna combinación de factores que produzca un modelo de clasificación con un AUC mayor de 0.8.

También se ejecutó un algoritmo LASSO, y el resultado fue de un AUC igual a 0.74 y según este modelo los factores más importantes son SRSF9 y NOVA1, lo cual coincide con los factores que tienen mayores diferencias significativas.

1. No se encontró ninguna combinación de factores que permita hacer un clúster jerárquico con un AUC mayor que 0.8.

NormalesP2 vs TumoralesP1

1. Existen diferencias significativas en los niveles de expresión de los siguientes factores

|  |  |  |
| --- | --- | --- |
| p-values < 0.05 | p-values < 0.01 | p-values < 0.001 |
| RBM22 | NOVA1 | SC35 |
| RBM45 | PSF | SKIP |
| SF3TV1 | SRSF9 |  |
|  | U2 |  |
|  | U2AF1 |  |
|  | U4ATAC |  |

1. Existe una alta correlación entre los factores:

* PSF y SKIP
* SC35 y SKIP
* SF3TV1 y SKIP

Estos factores por separado se expresan significativamente diferente en ambos grupos, y esta alta correlación indica que son factores con una distribución conjunta relevante.

1. Según los algoritmos de selección de características ejecutados, en promedio los factores que discriminan mejor a los grupos son (ordenados de mayor a menor relevancia):

SC35 0.02380952380952381

SKIP 0.061688311688311695

U2AF1 0.11012987012987015

U4atac 0.22298701298701293

U2 0.23489177489177487

Estos factores están dentro del grupo que fueron encontrados como más significativos.

1. Se encontraron muchísimas combinaciones de factores que producen modelos de clasificación con AUC bastante altos (0.9-1.0).

También se ejecutó un algoritmo LASSO, y el resultado fue de un AUC igual a 0.94 y según este modelo los factores más importantes son SC35, U4ATAC, U2AF1, U2 y SRSF9 (todos estos se expresan significativamente diferentes en ambos grupos, ver tabla).

1. Se encontraron 77 combinaciones de factores que permiten hacer un clúster jerárquico con un AUC de 1.

Conclusiones

* El factor SRSF9 está saliendo como uno de los más significativos para diferenciar entre muestras normales (sean de la población 1 ó 2) y tumorales, así como se evidencia que este factor no tiene cambios en muestras normales pertenecientes a poblaciones diferentes.
* U2 y SKIP permiten observar una diferenciación entre las muestras normales de la Pob1 y la Pob2, así como entre las muestras normales de la Pobl2 y las muestras tumorales de la Pob1. Si observamos los gráficos de estos factores, donde se expresan significativamente diferentes es en las muestras normales de la población 2. De todos modos, hay que tener en cuenta que estos resultados pueden estar sesgados por el hecho que de la población 2 solamente existen 5 muestras. Aunque la lógica nos dice que, si estos factores fueran comunes entre muestras normales independientemente de la población que sea, entonces estos factores deberían haber salido como significativos en la comparativa NormalesP1 vs TumoralesP1, y no fue el caso. Quizás una conclusión es que estos factores se ven mucho menos expresados en las muestras extraídas de la población 1, indicando que aunque sean muestra normales adyacentes a las tumorales, estos factores se expresan mucho menos en pacientes con tumores…