Se analizaron cinco escenarios, 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 3er cuartil + 1.5 IQR y por debajo de 1er cuartil-1.5IQR.
2. Los valores perdidos fueron sustituidos por la mediana de los valores de un mismo grupo.
3. Los valores se escalaron al rango [0,1].
4. Se ejecutó un test de Wilcoxon para detectar diferencias significativas en las expresiones de factores entre los diferentes grupos definidos por cada escenario.
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 cuál 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 jerárquico 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 escenario**

Control vs III

* Existen diferencias significativas en los niveles de expresión de los siguientes factores

|  |  |  |
| --- | --- | --- |
| p-values < 0.05 | p-values < 0.01 | p-values < 0.001 |
| ESRP2 | SF3TV2 | CA150 |
| RBM17 | U11 | CUGBP |
| SAM68TV1 |  | FBP11 |
| U4 |  | MAGOH |
|  |  | NOVA1 |
|  |  | PRP8 |
|  |  | PSF |
|  |  | PTB |
|  |  | RAVER1 |
|  |  | RBM3 |
|  |  | RBM22 |
|  |  | RBM45 |
|  |  | SC35 |
|  |  | SF3BTV1 |
|  |  | SKIP |
|  |  | SND1 |
|  |  | SRM160 |
|  |  | SRSF3 |
|  |  | SRSF4 |
|  |  | SRSF5 |
|  |  | SRSF6 |
|  |  | SRSF9 |
|  |  | SRSF10 |
|  |  | TIA1 |
|  |  | TRA2A |
|  |  | TRA2B |
|  |  | U1 |
|  |  | U2 |
|  |  | U2AF1 |
|  |  | U2AF2 |
|  |  | U4ATAC |
|  |  | U5 |
|  |  | U6 |
|  |  | U6ATAC |
|  |  | U12 |

* Existen altas correlaciones entre par de factores que se expresan significativamente diferentes cuando se analizan por sí solos, lo cual evidencia la existencia de distribuciones conjuntas.

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):

SF3BTV1 0.03571428571428571

SC35 0.0718045112781955

SRSF3 0.09323308270676693

RBM3 0.1368421052631579

RBM22 0.13721804511278196

U2AF1 0.15

SND1 0.1864661654135338

PTB 0.20112781954887218

RAVER1 0.20827067669172933

….

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

* Se encontraron más de 1900 combinaciones de factores que producen modelos de clasificación con AUC de 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 U2AF1, SC35, RBM3 y SRM160. El resto de factores no hacen falta para lograr una separación perfecta de los grupos.

* Se encontraron 1900 combinaciones de factores que permiten hacer un clúster jerárquico con un AUC de 1.

En este estudio hay que tener en cuenta que las poblaciones son pequeñas, tanto para los controles como para las muestras tumorales del grupo III. Al existir tantas diferencias en los niveles de expresiones de muchos factores, entonces es lógico que se encuentren tantas combinaciones de factores que producen modelos con buenos AUC. Hay que tener cuidado con las conclusiones que se sacan en estos estudios que tienen estos tipos de características.

Teniendo en cuenta que los modelos LASSO son modelos que trabajan muy bien en situaciones donde se tienen pocas muestras y muchos atributos (problemas que en estadística se conocen como Fat-Short problem), entonces me inclino más por los resultados arrojados por este tipo de modelo. En otras palabras, los factores más importantes son U2AF1, SC35, RBM3 y SRM160; estos son los que logran una mejor separación lineal entre los grupos.

Control vs IV

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 |
| CELF4 | U6ATAC | CA150, ESRP2 |
| nsR100 | U12 | CUGBP |
| RBM17 |  | FBP11 |
| U4 |  | MAGOH |
|  |  | NOVA1 |
|  |  | PRP8 |
|  |  | PSF |
|  |  | PTB |
|  |  | RAVER1 |
|  |  | RBM3 |
|  |  | RBM22 |
|  |  | RBM45, SAM68TV1 |
|  |  | SC35 |
|  |  | SF3BTV1, SF3TV2 |
|  |  | SKIP |
|  |  | SND1 |
|  |  | SRM160 |
|  |  | SRSF3 |
|  |  | SRSF4 |
|  |  | SRSF5 |
|  |  | SRSF6 |
|  |  | SRSF9 |
|  |  | SRSF10 |
|  |  | TIA1 |
|  |  | TRA2A |
|  |  | TRA2B |
|  |  | U1 |
|  |  | U2 |
|  |  | U2AF1 |
|  |  | U2AF2, U4 |
|  |  | U4ATAC |
|  |  | U5 |
|  |  | U6 |
|  |  | U11 |

Podemos observar que la mayoría de los factores que se expresaban significativamente diferentes en el escenario de Control vs III se repiten en este escenario. Nuevos factores aparecen expresados significativamente, tales como CELF4 y nsR100. Otros factores pasaron a estar expresados con diferencias más significativas, tales como SAM68TV1 y SF3TV2.

* Existe una alta correlación entre par de factores que se expresan significativamente por separados.

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

* Según los algoritmos de selección de características ejecutados, en promedio los 10 factores que discriminan mejor a los grupos son (ordenados de mayor a menor relevancia):

SRSF3 0.0347985347985348

RBM22 0.04761904761904762

RBM3 0.058608058608058615

U5 0.11921411921411924

SRSF6 0.1451881451881452

TIA1 0.17490842490842493

PTB 0.1978021978021978

CA150 0.2017982017982018

SAM68TV1 0.2797635697635698

TRA2A 0.3196403596403597

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

* Se encontraron cientos de combinaciones de factores que producen modelos de clasificación con AUC de 1.0.

Según el modelo LASSO los factores más importantes son RBM22, SRSF6, U5 y RBM3.

* Se encontraron 110 combinaciones de factores que permiten hacer un clúster jerárquico con un AUC de 1.

En este escenario vuelve a ocurrir una situación parecida al estudio Control vs III. Son demasiados los factores que se expresan significativamente diferentes entre los dos grupos, y a la vez muy pocas muestras. Se debe tener cuidado con las conclusiones que se saquen.

Control vs Tumor (grupos III y IV unidos)

-Existen diferencias significativas en los niveles de expresión de los siguientes factores

|  |  |  |
| --- | --- | --- |
| p-values < 0.05 | p-values < 0.01 | p-values < 0.001 |
| CELF4 | nsR100 | CA150, ESRP2 |
|  | RBM17 | CUGBP |
|  |  | FBP11 |
|  |  | MAGOH |
|  |  | NOVA1 |
|  |  | PRP8 |
|  |  | PSF |
|  |  | PTB |
|  |  | RAVER1 |
|  |  | RBM3 |
|  |  | RBM22 |
|  |  | RBM45, SAM68TV1 |
|  |  | SC35 |
|  |  | SF3BTV1, SF3TV2 |
|  |  | SKIP |
|  |  | SND1 |
|  |  | SRM160 |
|  |  | SRSF3 |
|  |  | SRSF4 |
|  |  | SRSF5 |
|  |  | SRSF6 |
|  |  | SRSF9 |
|  |  | SRSF10 |
|  |  | TIA1 |
|  |  | TRA2A |
|  |  | TRA2B |
|  |  | U1 |
|  |  | U2 |
|  |  | U2AF1 |
|  |  | U2AF2, U4 |
|  |  | U4ATAC |
|  |  | U5, U12 |
|  |  | U6, U6ATAC, U4 |
|  |  | U11 |

Podemos observar que la mayoría de los factores que se expresaban significativamente diferentes en los escenarios Control vs III y Control vs IV se repiten en este escenario. Varios factores pasaron de estar con diferencias significativas moderadas a con diferencias altas, esto es debido al incremento del número de muestras.

* Existe una alta correlación entre par de factores que se expresan significativamente por separados.

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

* Según los algoritmos de selección de características ejecutados, en promedio los 10 factores que discriminan mejor a los grupos son (ordenados de mayor a menor relevancia):

RBM22 0.014285714285714287

SRSF3 0.028571428571428574

RBM3 0.08571428571428572

PTB 0.14285714285714288

SRSF6 0.21428571428571427

TIA1 0.2285714285714286

CA150 0.2571428571428571

SRSF10 0.26857142857142857

SC35 0.32142857142857145

FBP11 0.35000000000000003

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

* Se encontraron miles de combinaciones de factores que producen modelos de clasificación con AUC de 1.0.

Según el modelo LASSO los factores más importantes son RBM3, SRSF6, SRSF10, RBM22 y U11 produciendo un modelo lineal con un AUC de 0.998.

* Se encontraron 394 combinaciones de factores que permiten hacer un clúster jerárquico con un AUC de 1.

Como era de esperar, en este escenario vuelve a ocurrir una situación parecida a los estudios Control vs III y Control vs IV. Se debe tener cuidado con las conclusiones que se puedan sacar.

Tumor III vs Tumor IV

-Existen diferencias significativas en los niveles de expresión de los siguientes factores

|  |  |  |
| --- | --- | --- |
| p-values < 0.05 | p-values < 0.01 | p-values < 0.001 |
| SND1 | RAVER1 | PSF |
| SRSF4 | U2AF1 | SRSF6 |
| SRSF5 | U12 | U5 |
| SRSF9 | RAVER1 |  |
| U4ATAC | U2 |  |
|  |  |  |

* Según los algoritmos de selección de características ejecutados, en promedio los 8 factores que discriminan mejor a los grupos son (ordenados de mayor a menor relevancia):

U5 0.01380952380952381

PSF 0.05326530612244897

U12 0.09163265306122448

RAVER1 0.10020408163265306

SRSF6 0.1301360544217687

U2 0.2358503401360544

SND1 0.2453741496598639

U2AF1 0.2460544217687074

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

* Se encontraron miles de combinaciones de factores que producen modelos de clasificación con AUC de entre 0.95-1.0.

Según el modelo LASSO, el cual tiene un AUC de 0.972, los factores más importantes son U5, PSF y SRM160.

* Se encontraron 4 combinaciones de factores que permiten hacer un clúster jerárquico con un AUC de 1, y más de 400 combinaciones que producen clusterings con AUC en el rango 0.9-1.0

Grupos de sanos

* En la mayoría de factores no existen diferencias significativas en los niveles de expresión, excepto en algunos pocos que se encontraron diferencias a un nivel de significancia de 0.05.
* Según los algoritmos de selección de características ejecutados, en promedio los 10 factores que discriminan mejor a los grupos son (ordenados de mayor a menor relevancia):

SKIP 0.020204081632653057

SF3BTV1 0.08530612244897959

RBM22 0.10938775510204081

U6ATAC 0.1559183673469388

ESRP2 0.32367346938775515

SND1 0.32489795918367353

NOVA1 0.3377551020408163

TIA1 0.37163265306122445

PSF 0.3981632653061225

* Se encontraron pocos modelos que logran un AUC en el rango 0.8-0.9.

Según el modelo LASSO, el cual obtuvo un AUC de 0.79, los factores más importantes según la zona de la muestra son:

Para W: TIA1, NOVA1

Para M: U2

Para B: SF3TV1 (este es el grupo que mejor se separa)

Para C: SKIP

* Se encontraron 98 combinaciones de factores que permiten hacer un clúster jerárquico con un AUC aproximadamente de 0.85.

Como era de esperar, en este escenario no se logran separar con buena precisión los grupos de muestras sanas. Hay que tener en cuenta que solo se tienen 14 muestras, por lo que es complejo sacar conclusiones estadísticas con tan pocos datos.