Se realizó el estudio sobre muestras pareadas (dataset Net-pareados), se realizaron los siguientes pasos.

1. Se ejecutó un test de Wilcoxon para detectar diferencias significativas en las expresiones de factores entre los diferentes grupos.
2. Eliminación de outliers. Teniendo en cuenta los grupos de muestras, se eliminaron todos aquellos valores que estaban por encima de 1.5 XIQ+ 3er cuartil y por debajo de 1er cuartil -1.5 XIQ.
3. En este estudio no había valores perdidos. Los outliers removidos fueron sustituidos por el método “*bagging fits*”.
4. Se realizó un *Power Transformation* mediante el método *Yeo-Johnson*. Esto reduce la heteroscedasticity en los datos.
5. Se centraron los datos, de esta manera nos centramos más en las diferencias que en las similitudes.
6. Los valores se escalaron mediante Autoscaling, de esta manera a todos los predictores (factores de splicing) se les da la misma importancia.
7. 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.
8. Se ejecutaron varios modelos para clasificar los grupos determinando cual subconjunto de factores produce un modelo con mayor AUC (umbral de 0.75).
9. Se ejecutó un algoritmo de clúster jeráquico determinando el mejor subconjunto de factores (umbral de 0.75).

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**

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 |
| PRP8 | SRSF9 | NOVA1 |
| SAM68TV1 |  |  |
| SRSF1 |  |  |
| SRSF5 |  |  |
| U6 |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

1. Según los algoritmos de selección de características ejecutados, en promedio los 6 factores que discriminan mejor a los grupos son (ordenados de mayor a menor relevancia):

NOVA1

SRSF9

U6

RAVER1

SRSF5

SC35

Estos factores están dentro del grupo que fueron encontrados como más significativos, excepto RAVER1.

1. Se puso un umbral de 0.75 de tal manera que se registraran solo aquellos modelos de clustering que logran un AUC mayor que este umbral. Solo se detectaron 15 modelos que tienen todos los factores significativos, siendo el modelo de mayor AUC igual a 0.80. Mientras que se encontraron casi 1000 combinaciones, que combinan factores significativos con no significativos, que producen clusterings con AUC entre 0.75 y 0.84. Con estos valores de AUC los heatmaps que se generan no son perfectos.
2. Se encontraron miles de combinaciones que producen modelos de clasificación con AUC mayores de 0.75. Los más relevantes de todos fueron el modelo de RandomForest:

TRA2B NOVA1 SRSF9 RAVER1 SRSF5 SC35 nSR100 CUGBP U2AF1 U4atac RBM45 ESRP2 -> AUC=0.9349112426035503

Este modelo tiene el inconveniente que incluye factores que en el estudio a priori (punto 1) no aparecen como significativos. De los 12 factores que incluye, solo tres aparecen en la lista de los significativos (NOVA1, SRSF9, SRSF5). Además, según la importancia que le da RandomForest a las variables, de las 12, las cinco más importante son SRSF9, SRSF5, NOVA1, RAVER1 y SC35. Estos cinco factores coinciden con la lista de factores que según los algoritmos de selección de características (punto 2) son los más discriminativos.