Resumen

Uno de los análisis más comunes en investigación genómica es la comparación de perfiles de expresión génica (o transcriptómica) para entender la relación entre los genes (o genotipo) y el fenotipo. El análisis del transcriptoma se ha llevado a cabo tradicionalmente utilizando *microarrays* y cada vez con mayor frecuencia desde 2008 mediante secuenciación del ARN (RNA-seq). Un objetivo fundamental en este tipo de estudios es identificar aquellos genes cuya expresión cambia entre condiciones, en otras palabras, seleccionar las variables más relevantes (genes) en términos de variabilidad entre-condiciones. El problema de selección de variables, normalmente conocido en transcriptómica como "análisis de expresión diferencial", se puede abordar desde el punto de vista univariante o multivariante, siempre teniendo en cuenta la complejidad del diseño experimental. Uno de los retos a los que los bioestadísticos se enfrentan al tratar de resolver este problema es la llamada "maldición de la dimensión": se tienen que analizar cientos o incluso miles de variables con muy pocas observaciones disponibles normalmente. Por tanto, es esencial proporcionar a los investigadores herramientas estadísticas eficientes para llevar a cabo esta tarea.

Un enfoque típico en selección de variables es contrastar la hipótesis nula de igualdad del nivel de expresión medio entre dos condiciones para un gen particular y hacerlo repetidamente para todos los genes del conjunto de datos transcriptómicos. Sin embargo, la transcriptómica puede conllevar también diseños experimentales

multifactoriales (múltiples tratamientos, varios estados de desarrollo, series temporales...). La primera parte de esta tesis se ha dedicado al problema de selección de variables en diseños multifactoriales cuando se usan métodos multivariantes para modelizar los perfiles de expresión génica en *microarrays*. En particular, elegimos como punto de partida la técnica multivariante ASCA-genes para proponer algunas estrategias de selección de genes diferencialmente expresados en un contexto multivariante, que fueron evaluadas bajo distintos escenarios biológicos.

La tecnología RNA-seq apareció al comienzo de esta tesis y ahora se aplica ampliamente en transcriptómica. Los datos de RNA-seq son en esencia diferentes a los datos de *microarrays* y esto ha motivado el desarrollo de nuevos métodos estadísticos para estudiar la expresión diferencial en esta tecnología. Por tanto, la segunda parte de la tesis se centra exclusivamente en experimentos de RNA-seq. Primero, desarrollamos una colección de procedimientos para determinar la calidad de las medidas de RNA-seq, identificar los sesgos potenciales de la tecnología y procesar los datos para reducir el impacto del ruido técnico en los resultados estadísticos. En segudo lugar, abordamos el problema de selección de variables para el caso de comparación de dos grupos. Dado que existe cierta controversia en la distribución teórica que siguen los datos de RNA-seq, optamos por investigar técnicas no paramétricas dirigidas por los datos para vencer las limitaciones de las hipótesis paramétricas y propusimos una estrategia que es eficiente a la hora de controlar la tasa de falsos positivos. Se desarrollaron dos metodologías, NOISeq para réplicas técnicas y NOISeqBIO para réplicas biológicas, y se compararon con los métodos más punteros.