Resum

Una de les anàlisis més comunes en investigació genòmica és la comparació de perfils d'expressió gènica (o transcriptòmica) per entendre la relació entre els gens (o genotip) i el fenotip. L'anàlisi del transcriptoma s'ha dut a terme tradicionalment utilitzant *microarrays* i cada vegada amb més freqüència des de 2008 mitjançant la seqüenciació de l'ARN (RNA-seq). Un objectiu fonamental en aquest tipus d'estudis és identificar aquells gens l'expressió dels quals canvia entre condicions, en altres paraules, seleccionar les variables més rellevants (gens) en termes de variabilitat entre-condicions. El problema de selecció de variables, normalment conegut en transcriptòmica com a "anàlisi d'expressió diferencial", es pot abordar des del punt de vista univariant o multivariant, sempre tenint en compte la complexitat del disseny experimental. Un dels reptes a què s'enfronten els bioestadístics en tractar de resoldre aquest problema és l'anomenada "maldició de la dimensió": s'han d'analitzar centenars o milers de variables amb molt poques observacions disponibles normalment. Per tant, és essencial proporcionar als investigadors ferramentes estadístiques eficients per a dur a terme aquesta tasca.

Un enfocament típic en selecció de variables és contrastar la hipòtesi nul·la d'igualtat de nivells d'expressió mitjans entre dues condicions per a un gen particular i fer-ho repetidament per a tots els gens del conjunt de dades transcriptòmiques. No obstant això, la transcriptòmica pot comportar també dissenys experimentals

multifactorials (múltiples tractaments, diversos estats de desenvolupament, sèries temporals...). La primera part d'aquesta tesi s'ha dedicat al problema de selecció de variables en dissenys multifactorials quan s'usen mètodes multivariants per a modelitzar els perfils d'expressió gènica en *microarrays*. En particular, triàrem com a punt de partida la tècnica multivariant ASCA-genes per a proposar algunes estratègies de selecció de gens diferencialment expressats en un context multivariant, que van ser avaluades sota diferents escenaris biològics.

La tecnologia RNA-seq va aparèixer al començament d'aquesta tesi i ara s'aplica àmpliament en transcriptòmica. Les dades de RNA-seq són en essència diferents a les dades de *microarrays* i açò ha motivat el desenvolupament de nous mètodes estadístics per a estudiar l'expressió diferencial en aquesta tecnologia. Així doncs, la segona part de la tesi se centra exclusivament en experiments de RNA-seq. En primer lloc, vam desenvolupar una col·lecció de procediments per a determinar la qualitat de les mesures de RNA-seq, identificar biaixos potencials de la tecnologia i processar les dades per a reduir l'impacte del soroll tècnic en els resultats estadístics. En segon lloc, es va abordar el problema de selecció de variables per al cas de comparació de dos grups. Donat que existeix certa controvèrsia en la distribució teòrica que segueixen les dades de RNA-seq, vam optar per investigar tècniques no paramètriques dirigides per les dades per a véncer les limitacions de les hipòtesis paramètriques i vam proposar una estratègia que és eficient a l'hora de controlar la tasa de falsos positius. Es desenvoluparen dues metodologies, NOISeq per a rèpliques tècniques i NOISeqBIO per a rèpliques biològiques, i es compararen amb els mètodes més capdavanters.