

La introducción de la secuenciación de ARN a nivel de célula única (scRNA-seq) en el ámbito de la transcriptómica ha redefinido nuestro entendimiento de la diversidad celular, arrojando luz sobre los mecanismos subyacentes a la heterogeneidad tisular. No obstante, al inicio de esta tesis, las limitaciones de esta tecnología obstaculizaban su aplicación en el estudio de procesos complejos, entre ellos el splicing alternativo. A pesar de ello, los patrones de splicing a nivel celular planteaban incógnitas que esta tecnología tenía el potencial de resolver: ¿es posible observar, a nivel celular, la misma diversidad de isoformas que se detecta mediante RNA-seq a nivel de tejido? ¿Qué función desempeñan las isoformas alternativas en la constitución de la identidad celular?

El objetivo de esta tesis es desbloquear el potencial del scRNA-seq para el análisis de isoformas alternativas, abordando sus dificultades técnicas y analíticas mediante el desarrollo de nuevas metodologías computacionales. Para lograrlo, se trazó una hoja de ruta con tres objetivos. Primero, se establecieron cuatro requisitos para el estudio de las isoformas mediante scRNA-seq, llevando a cabo una revisión de la literatura existente para evaluar su cumplimiento. Tras completar este marco con simulaciones computacionales, se identificaron las debilidades y fortalezas de los métodos de scRNA-seq y las herramientas computacionales disponibles. Durante la segunda etapa de la investigación, estos conocimientos se utilizaron para diseñar un protocolo óptimo de procesamiento de datos de scRNA-seq. En concreto, se integraron datos de lecturas largas a nivel de tejido con datos de scRNA-seq para garantizar una identificación adecuada de las isoformas así como su cuantificación a nivel celular. Este proceso permitió ampliar las estrategias computacionales disponibles para la reconstrucción de transcriptomas a partir de lecturas largas, mejoras que fueron implementadas en SQANTI3, software de referencia en transcriptómica. Por último, los datos procesados se utilizaron para desarrollar un nuevo método de análisis de co-expresión de isoformas a fin de desentrañar redes de regulación del splicing alternativo implicadas en la constitución de la identidad celular.

Dada la elevada variabilidad de los datos de scRNA-seq, este método se basa en la utilización de una estrategia de correlación basada en percentiles que atenúa el ruido técnico y permite la identificación de grupos de isoformas co-expresadas. Una vez configurada la red de co-expresión, se introdujo una nueva estrategia de análisis para la detección de patrones de co-utilización de isoformas que suceden de forma independiente a la expresión a nivel de gen, denominada co-Differential Isoform Usage. Este enfoque facilita la identificación de una capa de regulación de la identidad celular atribuible únicamente a mecanismos post-transcripcionales. Para una interpretación biológica más profunda, se aplicó una estrategia de anotación computacional de motivos y dominios funcionales en las isoformas definidas con lecturas largas, revelando las propiedades biológicas de las isoformas involucradas en la red de co-expresión. Estas investigaciones culminan en el lanzamiento de acorde, un paquete de R que encapsula las diferentes metodologías desarrolladas en esta tesis, potenciando la reproducibilidad de sus resultados y proporcionando una nueva herramienta para explorar la biología de las isoformas alternativas a nivel de célula única.

En resumen, esta tesis describe una serie de esfuerzos destinados a desbloquear el potencial de los datos de scRNA-seq para avanzar en la comprensión del splicing alternativo. Desde un contexto de escasez de herramientas y conocimiento previo, se han desarrollado soluciones de análisis innovadoras que permiten la aplicación de scRNA-seq al estudio de las isoformas alternativas, proporcionando recursos innovadores para profundizar en la regulación post-transcripcional y la función celular.