Transcriptómica, Regulación Genómica y Epigenómica

Resumen

La asignatura aborda el análisis de datos de transcriptómica y proteómica, analizando las tecnologías disponibles, la cuantificación de la expresión y métodos para el análisis estadístico de la expresión diferencial. Además, se verán métodos de análisis funcional, estudios de la regulación genómica y epigenómica, análisis multimodal de datos de célula única y métodos de clasificación supervisada y no supervisada (clustering) aplicados a datos ómicos de bulk y de célula única.

Obtendremos la capacidad de analizar de manera cuantitativa datos de transcriptómica y proteómica tanto a nivel de tejido como de célula única, e integrarlo con técnicas para el estudio de la expresión de la transcripción, tales como la modificación de histonas y la actividad de la cromatina y los factores de transcripción.

Sandra Mingo Ramírez

UAM - 2024/25

29 de enero de 2025 20:09

Universidad Autónoma de Madrid
Bioinformática y Biología Computacional

Código en Github

Índice general

Capítulo I

Diseño experimental y principios estadísticos del análisis de datos ómicos

El transcriptoma permite estudiar cómo se expresan los ARNs. Después va el proteoma, el cual se centra en el estudio de las proteínas. El último paso es el del metaboloma. Genómica, transcriptómica y proteómica se pueden analizar con NGS y arrays, mientras que proteómica y metabolómica se estudia con espectometría de masas. Para la proteómica, aunque la espectometría de masas tiene mayor detalle, la secuenciación es más escalable, por lo que se está popularizando. Concretamente hay dos casas comerciales que lo permiten: Olink y Somalogic.

I.1. Pipeline de un experimento ómico

Esto aplica a la cuantificación de la expresión con NGS, con espectometría de masas o de metabolitos con espectometría. La primera parte es la pregunta biológica. Un experimento de ómicas se basa en una pregunta clara de qué es lo que se busca en los datos. Esta pregunta guía la plataforma a utilizar. Por ejemplo, la proteómica se puede estudiar por secuenciación o por espectometría. Si tenemos una cohorte humana grande y queremos datos abundantes, quizás la mejor opción puede ser la secuenciación. En general, la tecnología nunca debe guiar, se debe elegir en función de la pregunta. Después hay que definir el diseño experimental. Una vez hecho el experimento, se analiza la imagen, se preprocesan los datos, se normalizan y se analizan. Dentro del análisis de datos, dentro de las ómicas para la cuantificación tienen tres pasos importantes: identificación de genes diferencialmente expresados, análisis de cluster y métodos de ingeniería reversa. Tras esto, se estandariza y se guardan los datos y finalmente se integran los datos y se interpreta biológicamente. Una base de datos importante es la base de datos GEO, Gene Expression Omnibus.

I.2. Diseño experimental

Esta parte es esencial, ya que los experimentos son muy caros. Se trata de utilizar ciertos principios para que el coste sea el menor posible, y a la vez poder extraer toda la información posible con ese coste. En otras palabras: minimizar el coste y maximizar la información obtenida.

Para hacer un buen diseño experimental, hay dos cosas esenciales:

1. **Pregunta biológica:** es imprescindible saber qué se está buscando para generar un experimento; ver si es data-driven o hypothesis-driven.
2. **Conocimiento de la tecnología:** medidas robustas y precisas de los datos. Replicación, tipo de variables que pueden meter sesgos en el experimento o variabilidad técnica. Estas técnicas buscan ser cuantitativas.

En un experimento, hay dos tipos de errores:

- **Errores aleatorios:** no son posibles de calibrar, pero se minimizan mediante la repetición de las mediciones.
- **Errores sistemáticos:** es posible de estimar y de eliminar de los datos. También se reduce normalizando, ya que son problemas de calibración.

A través de tres principios, se busca eliminar estos errores: replicación, randomización y blocking.

La distinción entre réplicas biológicas y técnicas depende de qué fuentes de variación se estudien o, alternativamente, se consideren fuentes de ruido. Existen las réplicas técnicas, las cuales minimizan los errores aleatorios mediante el promedio y ayudan a testar la tecnología, y réplicas biológicas, que permiten sacar conclusiones extrapolables a la población completa y no solo del individuo, además de poder controlar la variabilidad en diferentes pasos experimentales.

Supongamos que se realiza un experimento en el que se puede medir la expresión de un gen en una sola célula y se dispone de dinero para realizar 48 mediciones. Tenemos varias repeticiones: utilizamos varios ratones (replicación biológica), de cada ratón escogemos varias células (replicación técnica/biológica, está entre medias), y se realizan varias medidas dentro de cada célula (replicación técnica). Las medidas de la misma célula deberían ser muy parecidas. Si se realiza la media de los tres ratones, la medida va a ser muy variable en relación con una sola medida, pero esto sirve para el test estadístico, ya que son medidas independientes. En caso de tener medidas dependientes, no se puede utilizar la variabilidad para estudiar la significancia, ya que son medidas repetidas.

En el modelo propuesto hay que cuantificar bien la variabilidad. ¿Cómo escoger el tipo y número de réplicas? En un experimento de ómicas en el que se mide la expresión de un gen de células de hígado de ratón, se cuantifica una expresión de 12. Se realizan dos tipos de variabilidad biológica (animal y célula) y una variabilidad técnica (medición). Estas tres fuentes de variabilidad suman un 3,5 de variabilidad. Las normales están centradas en 10, ya que las medidas están saliendo en ese valor,

no en 12. Hay más variabilidad biológica que técnica, transformando las gaussianas en una parábola más aplastada. Se realizaron simulaciones cambiando el número de animales, el número de células y el número de réplicas técnicas. Se hacen 10.000 asignaciones, para que se agrupen de forma diferente las combinaciones del número total (48 animales, 1 sola célula; 24 animales, 2 células; ...; 1 animal, 48 células). Sabiendo la cantidad de animales, células y mediciones, se puede sacar el tamaño muestral real del experimento, permitiendo calcular así la diferencia entre la variabilidad experimental y la variabilidad real. En ómicas, somos poco capaces de estimar la variabilidad, ya que en general hay pocas réplicas. Si esto después de mete en un t-test, y la variabilidad es muy pequeña (o incluso 0), entonces el resultado es muy grande, teniendo un p muy pequeño, rechanzando la hipótesis nula de que no hay diferencia en la expresión.

Para tecnologías ómicas, se deben incluir al menos 3 réplicas biológicas. Todo esto es para la experimentación con animales. En caso de experimentación en humanos, la variabilidad es gigante.

Cuando comenzó la secuenciación, cuanto más se secuencie, más caro es el experimento. Por tanto, ¿es mejor más réplicas a menos profundidad, o menos réplicas a más profundidad? Hubo varios estudios con muchas simulaciones que vieron que lo importante era la secuenciación con réplica biológica. El número de lecturas tiene algo de relevancia, pero llegados a un número, no compensa a hacer mucha más secuenciación porque se llega a un plateau en cuanto a genes diferencialmente expresados. Las métricas aumentan más teniendo varias réplicas biológicas que teniendo varias lecturas.

Una forma de incrementar el número de individuos manteniendo el coste bajo es mediante el **pooling** de muestras. En el caso de muestras humanas no se hace porque se pierde información fenotípica y genotípica importante del paciente, pero en animales sí puede ser una buena idea si las características específicas por especimen (sexo, camada, edad, etc.) no son relevantes para el experimento. Lo mejor es tener cuantas más réplicas independientes posibles.

Algunos pasos en los que se introduce variabilidad en NGS son:

- Técnica: extracción del ARN, preparación de la librería, flow-cell, barcode, científico
- Biológica: sexo, camada/familia, edad

Además, hay sesgos sistemáticos y ruido por errores aleatorios.

Cuando hay estudios muy grandes, hay pasos que se deben realizar en varias tandas. Al secuenciar en diferentes días, se introduce variabilidad. Esto se conoce !!!!! como el **efecto de batch**. Nunca hay que confundir el batch con el grupo biológico relevante, ya que es imposible ver si las diferencias son debidas al grupo biológico o a la variabilidad técnica. Cuando hay condiciones biológicas muy fuertes, a veces no se ven, pero si se hacen todas las muestras de una condición en un mismo batch, probablemente se estén magnificando las diferencias observadas. Por tanto, no hay que medir las distintas condiciones biológicas en batches distintos, si no mezclar en un batch muestras de distintas condiciones biológicas para poder utilizar la variable batch en el modelo estadístico mixto, normalizando por las diferencias entre los batches.

El **blocking** reduce fuentes conocidas e irrelevantes de variación entre unidades, permitiendo una precisión mayor en la estimación de las fuentes de variación estudiadas. Minimiza el efecto de variables de tipo biológico o técnico, que no son relevantes para la pregunta biológica. Una forma de hacer blocking secuenciando es metiendo adaptadores para hacer un barcoding de cada muestra, preparar la librería con todo junto y crear, de esa muestra, las distintas alícuotas a secuenciar. De esta forma se reduce el efecto de línea.

Todo esto se basa también en la **randomización**, de forma que sea representativo de la población.

I.2.1. Diseño experimental - Ejercicios

I.2.1.1. Ejercicio de animales

Tenemos un ratón knock-out en la proteína Bmi1. Para cada camada tenemos varios ratones WT y KO. Queremos encontrar metabolitos cuya expresión cambie significativamente entre condiciones Disponemos de 6 camadas con el siguiente número de animales:

Camada	KO	WT
L1	1	2
L2	2	2
L3	1	1
L4	1	3
L5	2	3
L6	3	2

- **Caso 1: No hay limitación económica**: se secuencia todo, ya que cuantas más muestras independientes, mejor.
- **Caso 2: Se pueden secuenciar un máximo de 6 muestras**: De las 6 camadas se escogen aleatoriamente 3, de las cuales escoger un ratón KO y uno WT. Otra opción es coger las camadas 2 y 3 y secuenciar todos esos individuos. En este caso, como las camadas tienen efecto, se podría elegir un individuo de cada camada y hacer pool de 2 en 2.
- **Caso 3: L5 no tiene ningún animal KO y seguimos con el máximo de 6 muestras**: L5 no se tendría en cuenta porque podría introducir sesgos (quizás el KO no ha salido, o quizás no es viable), y del resto de camadas se escogen 3 camadas al azar para seleccionar un ratón de cada condición. Esto se debe a que no se podría comparar el pool entre la misma camada con pool entre distintas camadas.
- **Caso 4: máximo de 6 muestras si no hay efecto de la camada**: se mezclan todos los ratones de las distintas camadas, separando por condición biológica, y se sacan 3 de cada uno al azar. Se podrían coger 12 y 12 y pooles de 4, o 6 y 6 y pooles de 2.

I.2.1.2. Ejercicio de humanos

Tenemos una cohorte de 100 muestras humanas con diabetes. Queremos probar en ellas un fármaco y ver sus efectos en la expresión génica. Podemos secuenciar un total de 40 muestras. Por estudios piloto sabemos que el sexo y el IMC afectan al impacto del fármaco. La composición de la cohorte es la siguiente: Además, no podemos procesar

	Hombres	Mujeres
IMC alto	40	20
IMC bajo	20	20

todas las muestras juntas, tenemos que hacerlo en dos ejecuciones.

- **Q1: ¿Cómo se asignan los pacientes a los grupos fármaco y placebo?** Se escogen 5 personas de cada condición (sexo y BMI) para fármaco y otros 5 para placebo.
- **Q2: ¿Qué pacientes se secuenciarían en cada turno?** Se cogen ordenadamente una muestra de cada grupo y condición.

I.3. Consideraciones estadísticas para datos ómicos