

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.

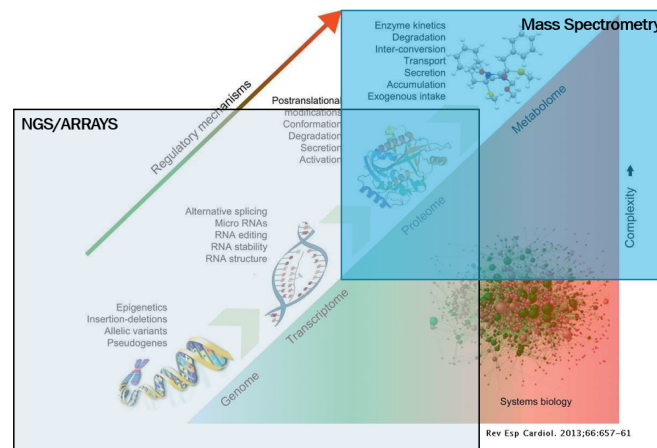
Índice general

I	Diseño experimental y principios estadísticos del análisis de datos ómicos	2
I.1	Pipeline de un experimento ómico	3
I.2	Diseño experimental	4
I.2.1	Ejemplo de diseño experimental	5
I.2.2	Réplicas vs profundidad	6
I.2.3	Pooling, batch y blocking	6
I.2.4	Diseño experimental - Ejercicios	8
I.3	Consideraciones estadísticas para datos ómicos	9

Capítulo I

Diseño experimental y principios estadísticos del análisis de datos ómicos

El transcriptoma permite estudiar cómo se expresan los ARNs, incluyendo tanto ARNs codificantes (como los ARNm) como no codificantes (como microARNs, ARN de transferencia, etc.). El proteoma se centra en el estudio de las proteínas, que son los productos funcionales de muchos ARNm. Finalmente, el metaboloma estudia los metabolitos, que son los productos finales de las reacciones bioquímicas en las células.



La genómica, transcriptómica y proteómica se pueden analizar utilizando tecnologías de secuenciación de próxima generación (NGS) y microarrays. Sin embargo, la proteómica y la metabolómica también se estudian comúnmente con espectrometría de masas, una técnica que permite identificar y cuantificar moléculas basándose en su masa y carga. Aunque la espectrometría de masas ofrece un mayor detalle en la identificación de proteínas y metabolitos, la secuenciación es más escalable y se está popularizando, especialmente en estudios a gran escala. Dos empresas comerciales que permiten la secuenciación de proteínas son Olink y Somalogic.

I.1. Pipeline de un experimento ómico

El pipeline de un experimento ómico aplica tanto a la cuantificación de la expresión génica con NGS, como a la identificación de proteínas con espectrometría de masas o de metabolitos con espectrometría. El proceso generalmente sigue los siguientes pasos:

1. **Pregunta biológica:** Todo experimento ómico comienza con una pregunta biológica clara. Esta pregunta debe ser lo suficientemente específica para guiar el diseño experimental y la elección de la plataforma tecnológica. Por ejemplo, si la pregunta es sobre la expresión génica en una cohorte grande de pacientes, la secuenciación de ARN (RNA-seq) podría ser la opción más adecuada.
2. **Elección de la plataforma tecnológica:** La elección de la tecnología (NGS, espectrometría de masas, microarrays, etc.) debe basarse en la pregunta biológica y no al revés. Por ejemplo, si se busca un alto rendimiento y escalabilidad, la secuenciación podría ser preferible sobre la espectrometría de masas.
3. **Diseño experimental:** Es crucial diseñar el experimento de manera que se minimice el sesgo y se maximice la reproducibilidad. Esto incluye la selección adecuada de controles, la replicación biológica y técnica, y la consideración de factores de confusión.
4. **Adquisición de datos:** Una vez diseñado el experimento, se procede a la recolección de datos. Esto puede implicar la secuenciación de ARN, la identificación de proteínas por espectrometría de masas, o la cuantificación de metabolitos.
5. **Preprocesamiento de datos:** Los datos crudos suelen requerir un preprocesamiento que incluye la corrección de errores, la normalización y la eliminación de ruido. Este paso es crucial para asegurar que los datos sean de alta calidad antes de proceder al análisis.
6. **Análisis de datos:** El análisis de datos en estudios ómicos generalmente incluye:
 - **Identificación de genes diferencialmente expresados:** Esto implica comparar los niveles de expresión génica entre diferentes condiciones (por ejemplo, tejido sano vs. tejido enfermo) para identificar genes que están regulados al alza o a la baja.
 - **Análisis de clusters:** Este método agrupa genes o muestras con patrones de expresión similares, lo que puede ayudar a identificar subtipos de enfermedades o vías biológicas relevantes.
 - **Ingeniería reversa de redes génicas:** Este enfoque intenta reconstruir las redes de regulación génica a partir de los datos de expresión, lo que puede proporcionar insights sobre cómo los genes interactúan entre sí.
7. **Estandarización y almacenamiento de datos:** Los datos deben ser estandarizados y almacenados en bases de datos públicas o privadas para su posterior acceso y análisis. Una base de datos importante es el *Gene Expression Omnibus (GEO)*, que alberga datos de expresión génica de diversos organismos y condiciones experimentales.

8. **Integración e interpretación biológica:** Finalmente, los datos se integran con información biológica adicional (como anotaciones funcionales, interacciones proteína-proteína, etc.) para interpretar los resultados en un contexto biológico más amplio. Esto puede llevar a la identificación de biomarcadores, dianas terapéuticas o mecanismos moleculares subyacentes a una enfermedad.

1.2. Diseño experimental

El diseño experimental es esencial en estudios ómicos debido al alto coste de los experimentos. El objetivo es minimizar el coste y maximizar la información obtenida. Para lograrlo, hay dos aspectos clave:

1. **Pregunta biológica:** Es imprescindible tener una pregunta biológica clara y específica. Esto determina si el enfoque es *data-driven* (exploratorio) o *hypothesis-driven* (basado en hipótesis).
2. **Conocimiento de la tecnología:** Es crucial entender las limitaciones y capacidades de la tecnología utilizada. Esto incluye la precisión de las mediciones, la replicación y la identificación de variables que pueden introducir sesgos o variabilidad técnica.

En un experimento, hay dos tipos de errores:

- **Errores aleatorios:** No son predecibles, pero se pueden minimizar mediante la repetición de las mediciones.
- **Errores sistemáticos:** Son predecibles y se pueden eliminar mediante la normalización o calibración de los datos.

Los principios para minimizar errores son:

1. **Replicación:** Incluye réplicas técnicas (para minimizar errores aleatorios) y réplicas biológicas (para asegurar que los resultados sean extrapolables a la población). La distinción entre réplicas biológicas y técnicas depende de qué fuentes de variación se estudien o, alternatively, 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.
2. **Randomización:** Asegura que las muestras sean representativas de la población.
3. **Blocking:** Reduce fuentes conocidas de variación que no son relevantes para la pregunta biológica.

1.2.1. Ejemplo de diseño experimental

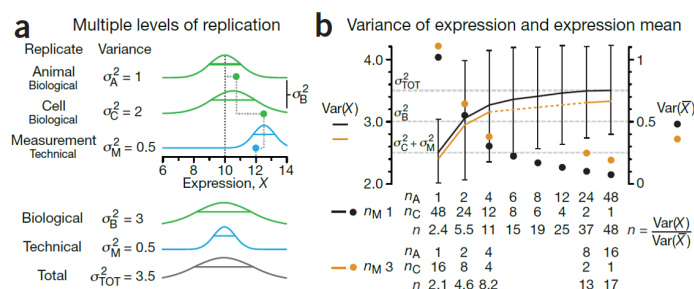
Supongamos que se mide la expresión de un gen en células de hígado de ratón, pudiendo realizar solo 48 mediciones. Se pueden considerar tres fuentes de variabilidad:

- **Replicación biológica:** se utilizan varios ratones, habiendo variabilidad entre los diferentes animales.
- **Replicación entre biológica y técnica:** se escogen varias células de cada ratón.
- **Replicación técnica:** se realizan varias mediciones de cada célula. Las distintas mediciones de una misma célula deberían ser similares.

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.

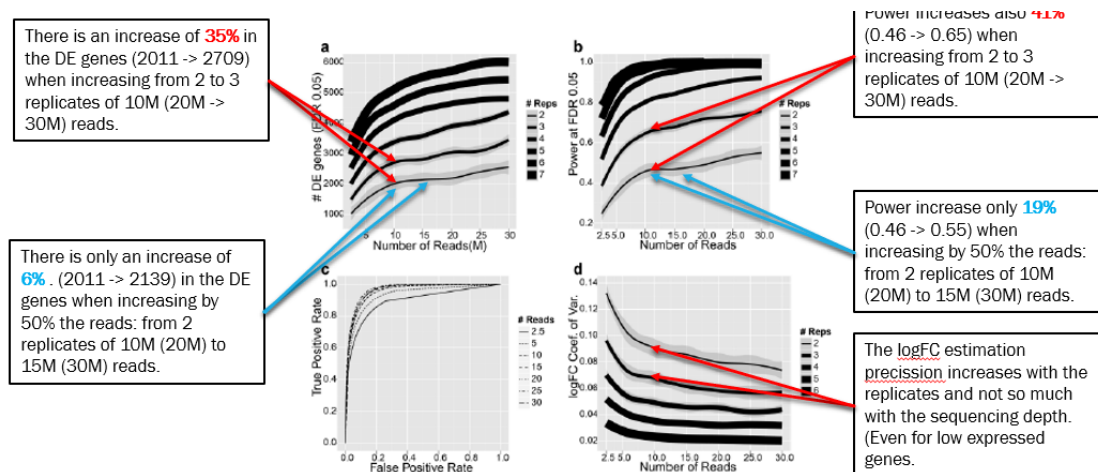
Como en este modelo propuesto hay que cuantificar la variabilidad, se pueden realizar simulaciones para determinar el número óptimo de réplicas biológicas y técnicas para minimizar la variabilidad y maximizar la precisión.

We measure the expression of a specific gene in liver cells in mice, $X=12$



I.2.2. Réplicas vs profundidad

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.

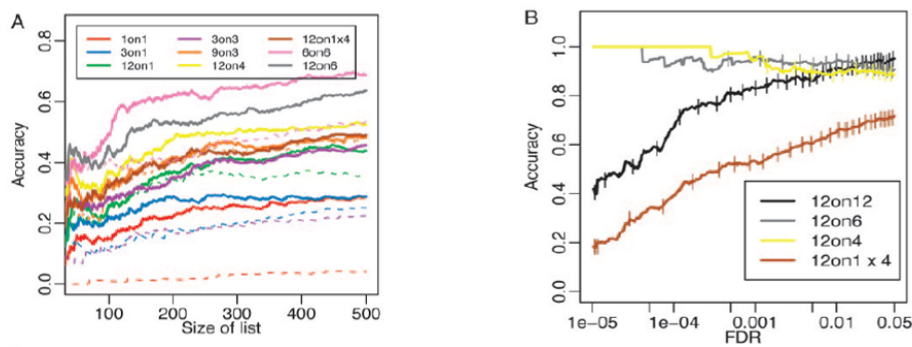


I.2.3. Pooling, batch y blocking

En el contexto de los experimentos ómicos, el **pooling** es una estrategia que consiste en combinar (o "agrupar") múltiples muestras biológicas en una sola muestra antes de realizar el análisis. Esta técnica se utiliza principalmente para reducir costes y simplificar el procesamiento de muestras, especialmente cuando se trabaja con un gran número de individuos o cuando los recursos económicos son limitados. En lugar de analizar cada muestra individualmente, se mezclan varias muestras en una sola. Por ejemplo, si tienes 12 muestras de ARN de diferentes individuos, podrías combinarlas en 3 grupos (pools) de 4 muestras cada uno. Una vez combinadas, las muestras agrupadas se procesan y analizan como una sola. Esto significa que, en lugar de obtener datos individuales para cada muestra, obtienes un resultado promedio para cada pool. En el caso de muestras humanas no se hace porque se pierde información individual fenotípica y genotípica, pero en animales sí puede ser una buena idea si las características específicas por espécimen (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:

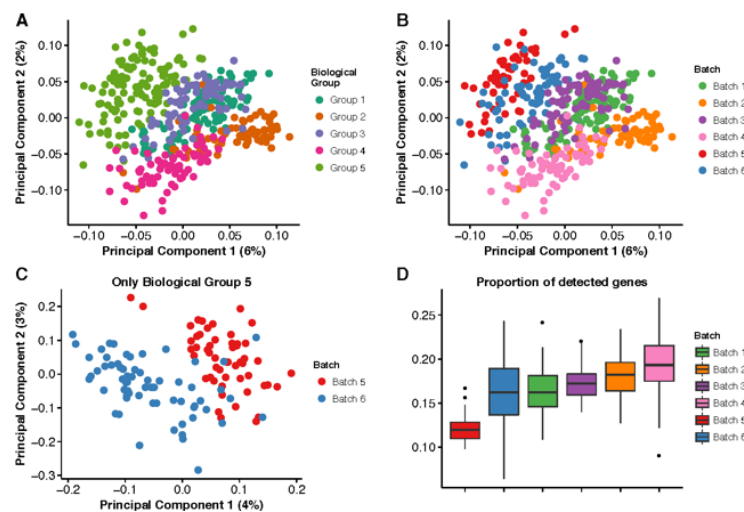
- Técnica: extracción del ARN, preparación de la librería, flow-cell, barcode, científico
- Biológica: sexo, camada/familia, edad



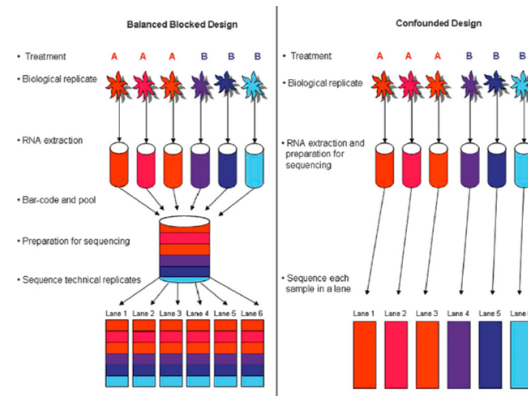
Además, hay sesgos sistemáticos y ruido por errores aleatorios.

!!!!

Cuando los experimentos se realizan en varias tandas (por ejemplo, por ser estudios muy grandes con muchas muestras), es crucial controlar el **efecto de batch** para evitar sesgos técnicos. Esto se puede lograr mediante la randomización de muestras entre batches y el uso de modelos estadísticos mixtos. 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.



I.2.4. Diseño experimental - Ejercicios

I.2.4.1. 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 secuencian todos, 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 pools de 4, o 6 y 6 y pools de 2.

I.2.4.2. 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