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

14 de febrero de 2025 23:26

Universidad Autónoma de Madrid
Bioinformática y Biología Computacional

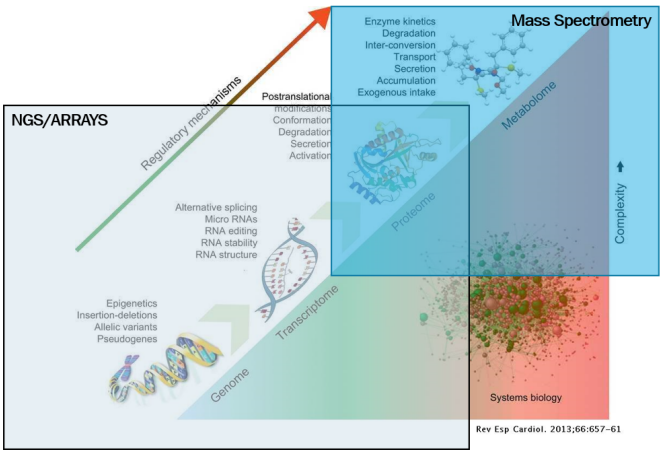
Código en Github

Índice general

Capítulo I

Diseño experimental y principios estadísticos del análisis de datos ómicos

El transcriptoma permite estudiar cómo se expresan los ARNs, 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.

I.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, 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.
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.

I.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 raón.
- **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

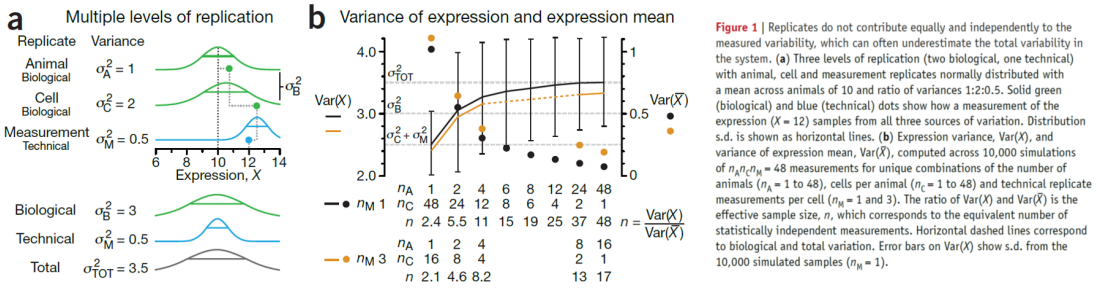


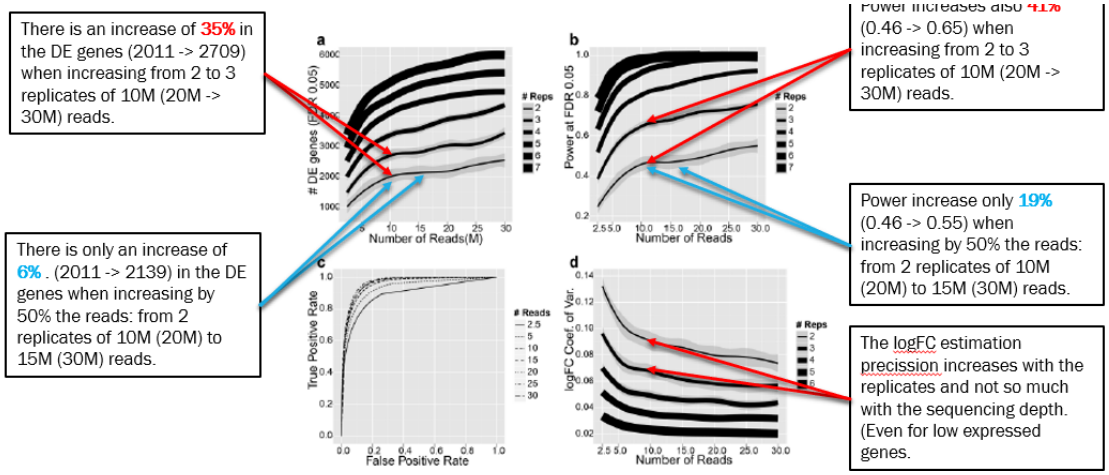
Figure 1 | Replicates do not contribute equally and independently to the measured variability, which can often underestimate the total variability in the system. (**a**) Three levels of replication (two biological, one technical) with animal, cell and measurement replicates normally distributed with a mean across animals of 10 and ratio of variances 1:2:0.5. Solid green (biological) and blue (technical) dots show how a measurement of the expression (X = 12) samples from all three sources of variation. Distribution s.d. is shown as horizontal lines. (**b**) Expression variance, Var(X), and variance of expression mean, Var($\bar{X}$), computed across 10,000 simulations of $n_A n_C n_M$ = 48 measurements for unique combinations of the number of animals (n_A = 1 to 48), cells per animal (n_C = 1 to 48) and technical replicate measurements per cell (n_M = 1 and 3). The ratio of Var(X) and Var($\bar{X}$) is the effective sample size, n, which corresponds to the equivalent number of statistically independent measurements. Horizontal dashed lines correspond to biological and total variation. Error bars on Var(X) show s.d. from the 10,000 simulated samples (n_M = 1).

En este experimento de ómicas, la expresión de un gen de células de hígado de ratón se cuantifica en 12. Las tres fuentes de variabilidad (animal, célula y medición) suman un 3,5. 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 campana 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, rechazando 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.

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.

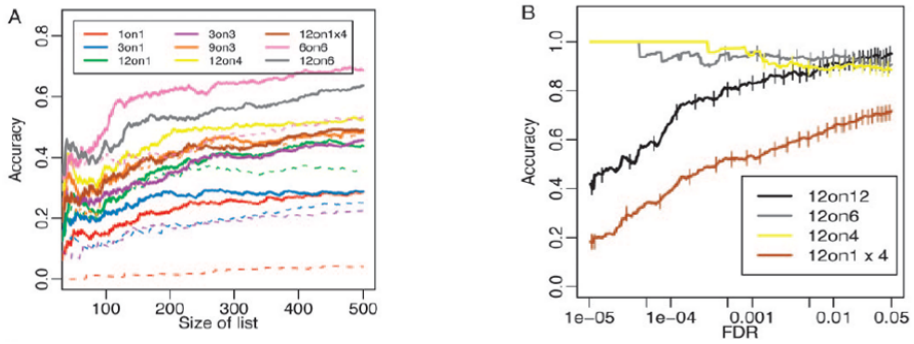


I.2.3. Pooling, batch y blocking

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 especimen (sexo, camada, edad, etc.) no son relevantes para el experimento. Lo mejor es tener cuantas más réplicas independientes posibles.

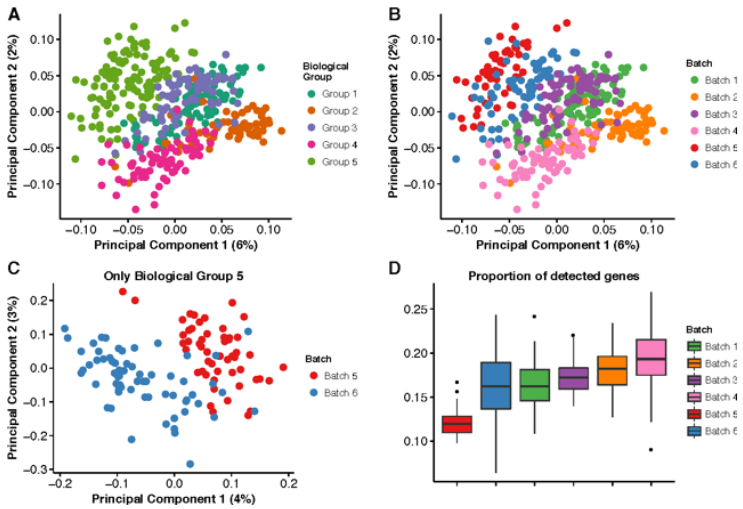
Algunos pasos en los que se introduce variabilidad en NGS son:

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

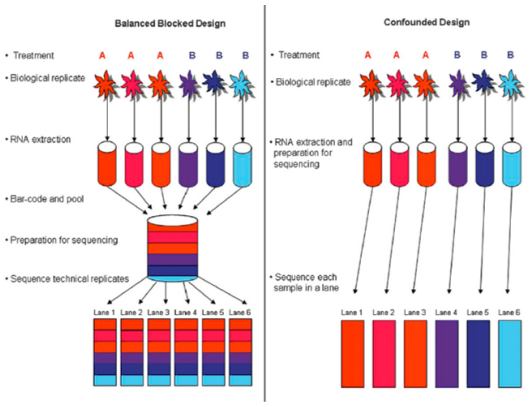


Además, hay sesgos sistemáticos y ruido por errores aleatorios.

Cuando 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 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.2.4.2. 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

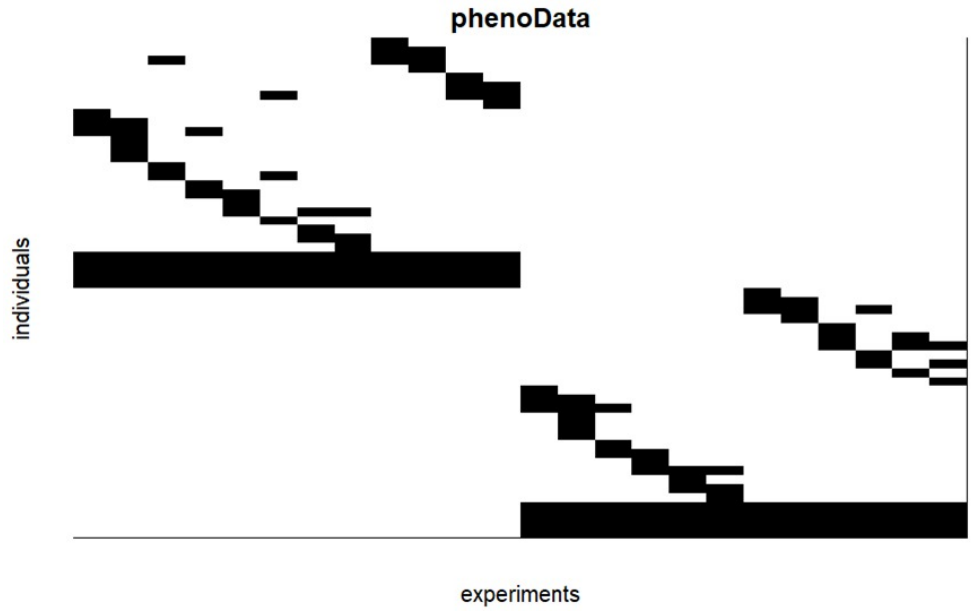
I.3.1. Ejemplo - Statistics for Omics

En este ejemplo, se utilizan datos del paquete de R maPooling, que contiene información sobre la expresión génica en dos condiciones (por ejemplo, wild-type vs. knock-out) y diferentes animales de cada condición. La matriz de diseño indica qué ratones (columna) están incluidos en cada muestra (fila), con un 1 para los ratones incluidos y un 0 para los excluidos. Por ejemplo, a3tr1 representa una réplica técnica del ratón 3 de la condición a, y aq es un pool de todos los ratones de la condición a.

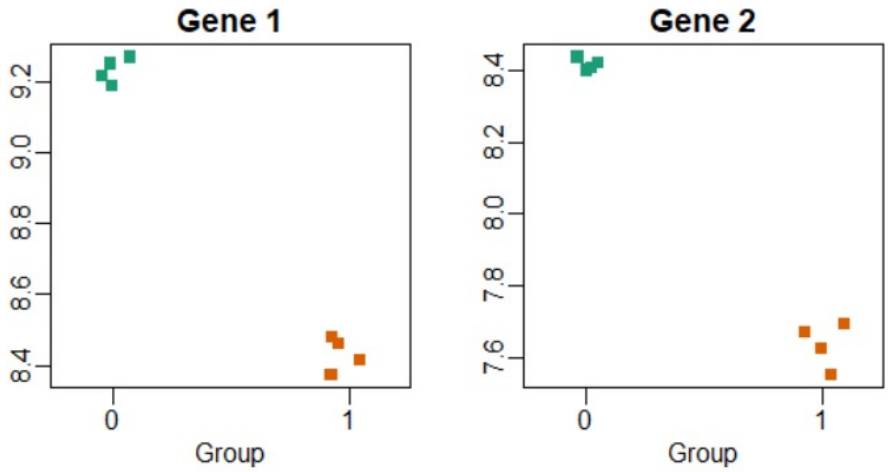
El siguiente código muestra qué ratones están incluidos en cada muestra:

```r
# r
library(rafalib)
mypar()
flipt <- function(m) t(m[nrow(m):1,])
myimage <- function(m,...) {
  image(flipt(m),xaxt="n",yaxt="n",...)
}
myimage(as.matrix(pData(maPooling)),col=c("white","black"),
        xlab="experiments",
        ylab="individuals",
        main="phenoData")
```

El objetivo es identificar diferencias en la expresión génica entre las condiciones a y b. Para ilustrar esto, se examinan dos genes específicos:




```r
# r
###look at 2 pre-selected genes for illustration
i=11425;j=11878
pooled_y=exprs(maPooling[,pooled])
pooled_g=factor(as.numeric(grepl("b",names(pooled))))
mypar(1,2)
stripchart(split(pooled_y[i,],pooled_g),vertical=TRUE,method="jitter",col=c(1,2),
          main="Gene 1",xlab="Group",pch=15)
stripchart(split(pooled_y[j,],pooled_g),vertical=TRUE,method="jitter",col=c(1,2),
          main="Gene 2",xlab="Group",pch=15)
```



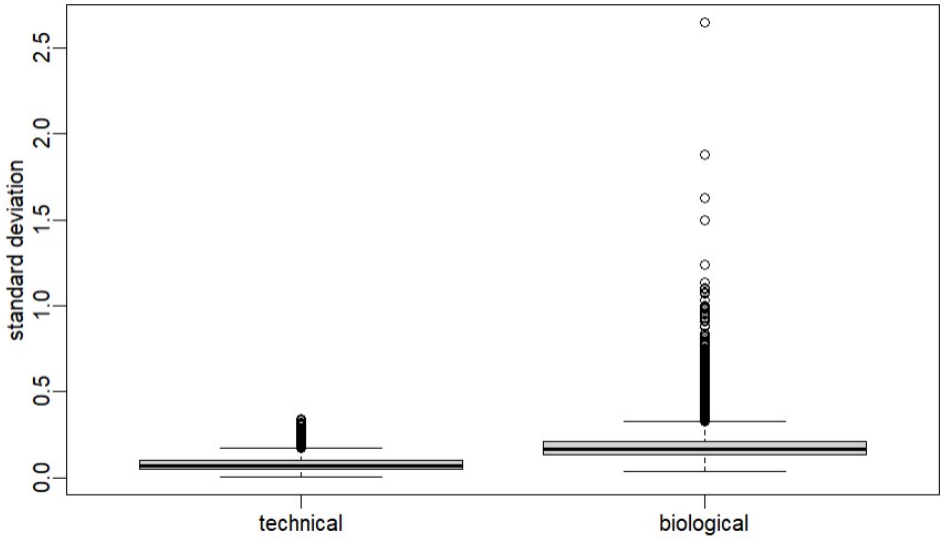
Se realiza un test estadístico (t-test) para evaluar si las diferencias en la expresión génica son significativas. Los p-valores obtenidos son muy bajos, del orden de 10^{-7}:

```r
# r
library(genefilter)
pooled_tt=rowttests(pooled_y,pooled_g)
```

```
pooled_tt$p.value[i]
pooled_tt$p.value[j]
```

Para obtener la variabilidad biológica, se eliminan las réplicas técnicas y se calcula la desviación estándar de las réplicas biológicas:

```
# r
technicalsd <- rowSds(pooled_y[,pooled_g==0])
biologicalsd <- rowSds(y[,g==0])
LIM=range(c(technicalsd,biologicalsd))
mypar(1,1)
boxplot(technicalsd,biologicalsd,names=c("technical","biological"),ylab="standard
    deviation")
```

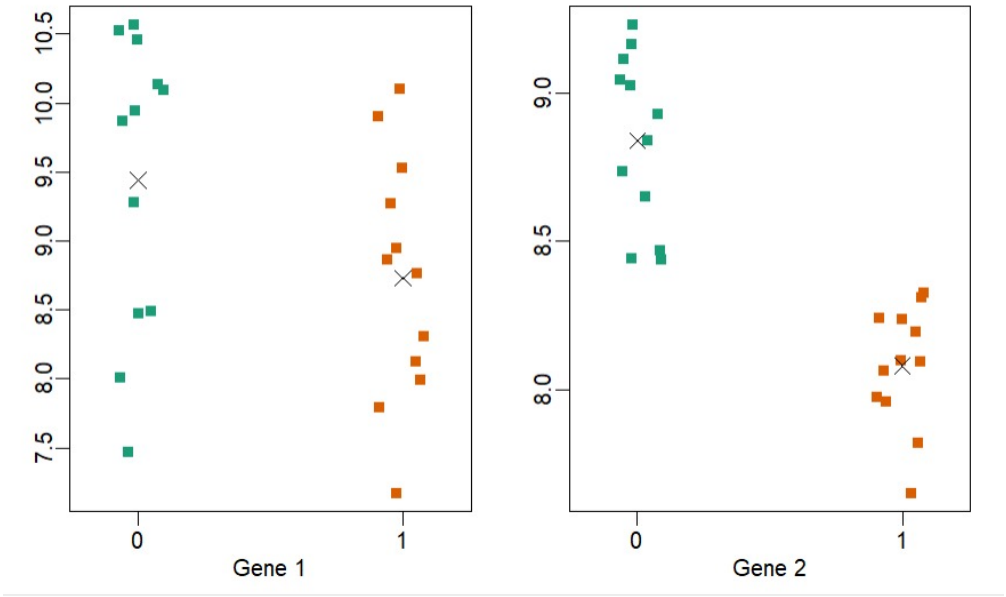


Se observa que la variabilidad biológica es mucho mayor que la variabilidad técnica. Además, la variabilidad de las varianzas también es mayor para la variabilidad biológica. A continuación, se muestran los valores de expresión de los dos genes en cada ratón individual:

```
# r
mypar(1,2)
stripchart(split(y[i,],g),vertical=TRUE,method="jitter",col=c(1,2),xlab="Gene
    1",pch=15)
points(c(1,2),tapply(y[i,],g,mean),pch=4,cex=1.5)
stripchart(split(y[j,],g),vertical=TRUE,method="jitter",col=c(1,2),xlab="Gene
    2",pch=15)
points(c(1,2),tapply(y[j,],g,mean),pch=4,cex=1.5)
```

Al volver a calcular el t-test, se observa que el gen 1 no es significativo, mientras que el gen 2 sí lo es:

```
# r
library(genefilter)
```




```
tt=rowttests(y,g)
tt$p.value[i]
tt$p.value[j]
```

I.3.2. Estadística en datos ómicos

En experimentos con animales, el número mínimo recomendado de réplicas biológicas independientes es 3, que pueden incluir pooling (combinación de muestras). Sin embargo, en humanos, la variabilidad es mucho mayor, y el pooling no es recomendable debido a la pérdida de información individual.

Los experimentos ómicos se caracterizan por tener una **matriz "skinny"**, es decir, muchas filas (genes, proteínas, metabolitos) y pocas columnas (muestras). Esto se conoce como la **maldición de la dimensionalidad** ("curse of dimensionality"). Aunque no es "Big Data" en el sentido estricto (ya que no hay muchas muestras), tener tantas características (filas) presenta desafíos estadísticos.

I.3.2.1. Problemas con la estimación de la variabilidad

Con pocas réplicas, la estimación de la variabilidad es imprecisa. En un t-test, se compara la diferencia de medias entre dos grupos en relación con la variabilidad dentro de cada grupo. Si la variabilidad se subestima, el t-test puede dar resultados falsamente significativos.

Para abordar este problema, se utiliza información de todo el experimento para estimar mejor la variabilidad. Esto se conoce como **moderación de la variabilidad o shrinkage**. La idea es "regularizar" la variabilidad utilizando la distribución de las desviaciones estándar de todos los genes. Por ejemplo:

- **Cálculo de la mediana:** Se calcula la mediana de todas las desviaciones estándar y se añade un offset (s_0) a la estimación de la variabilidad de cada gen. Esto evita que valores extremadamente bajos de variabilidad (debidos al azar) dominen los resultados.
- **Uso del percentil 90:** En lugar de la mediana, se puede usar el percentil 90 de las desviaciones estándar para obtener estimaciones más conservadoras. Esto es útil porque los genes con variabilidades muy altas probablemente no sean reales. Y en ómicas, lo más importante es que lo que se estudie sea real.

Este enfoque se implementa en métodos como el **t-test moderado** (por ejemplo, en el paquete limma de R), que ajusta las estimaciones de variabilidad para hacerlas más robustas.

El shrinkage es una técnica estadística que "contrae" las estimaciones de variabilidad hacia un valor central (como la mediana o el percentil 90). Esto ayuda a evitar sobreajustes y a obtener resultados más confiables en experimentos con pocas réplicas.

I.3.2.2. Problema de la no normalidad de los datos ómicos

Los datos ómicos (como los de expresión génica) a menudo no siguen una distribución normal, lo que puede complicar su análisis estadístico. Para abordar este problema, se utilizan técnicas de **transformación** y **normalización** de los datos, que permiten que estos se ajusten mejor a los supuestos de los tests estadísticos. Estas transformaciones pueden ser de dos tipos:

- **Normalizaciones interpretables:** Aquellas que buscan que los datos tengan un significado biológico claro.
- **Normalizaciones estadísticas:** Aquellas que buscan estabilizar la varianza o cumplir con los supuestos de los tests estadísticos.

Estabilización de la varianza Uno de los principales problemas en los datos ómicos es que la **variabilidad** no es constante en todos los niveles de expresión. En general, los genes poco expresados tienden a tener una variabilidad mucho mayor que los genes muy expresados. Esto puede sesgar los resultados de los tests estadísticos, ya que estos asumen que la varianza es homogénea.

Para solucionar este problema, se utiliza la estabilización de la varianza. Una herramienta común para esto es voom, que es parte del paquete limma en R. Voom transforma los datos de conteo de secuenciación (como los de RNA-seq) para estabilizar la varianza y hacer que los datos sean más adecuados para análisis estadísticos basados en modelos lineales.

Normalización por longitud del gen y tamaño de la librería En los datos de secuenciación (por ejemplo, RNA-seq), hay dos factores importantes que deben tenerse en cuenta al comparar la expresión génica:

- **Longitud del gen:** Los genes más largos tienden a tener más lecturas (reads) simplemente porque hay más regiones donde las lecturas pueden alinearse. Para comparar la expresión entre genes de diferentes longitudes, es necesario normalizar por la longitud del gen. Esto se hace típicamente mediante métricas como FPKM (Fragments Per Kilobase of transcript per Million mapped reads) o TPM (Transcripts Per Million).
- **Tamaño de la librería:** El número total de lecturas en una muestra (tamaño de la librería) puede variar entre condiciones experimentales. Si no se normaliza por el tamaño de la librería, las diferencias en la expresión génica podrían deberse simplemente a que una muestra tuvo más lecturas en general, en lugar de a cambios biológicos reales. Métricas como CPM (Counts Per Million) ayudan a corregir esto.

Cuando se comparan las mismas condiciones entre sí, la longitud del gen no es un problema (ya que es constante), pero el tamaño de la librería sí puede afectar los resultados y debe ser normalizado.

I.3.2.3. Problema de hipótesis múltiples

En los experimentos ómicos, se realizan miles de tests estadísticos simultáneamente (uno por cada gen, proteína o metabolito). Esto da lugar al problema de las comparaciones múltiples, que aumenta la probabilidad de obtener falsos positivos (es decir, rechazar la hipótesis nula cuando en realidad es verdadera).

Hipótesis nula La hipótesis nula en este contexto es que no hay diferencia en la expresión de un gen entre las condiciones comparadas. Sin embargo, debido al gran número de tests realizados, es probable que algunos genes aparezcan como diferencialmente expresados simplemente por azar.

Para controlar el número de falsos positivos, se utilizan métodos de corrección de múltiples comparaciones. Algunos de los más comunes son:

- **Corrección de Bonferroni:** Ajusta el nivel de significancia (α) dividiéndolo por el número total de tests realizados. Es muy conservadora y puede reducir demasiado la potencia estadística.
- **Control de la tasa de descubrimiento falso (FDR):** Controla la proporción esperada de falsos positivos entre los resultados significativos. Es menos conservadora que Bonferroni y se utiliza ampliamente en análisis ómicos. Solo se permite un 5 % de errores entre los genes que se seleccionan como diferencialmente expresados.
- **Valores q:** Son una versión ajustada de los p-valores que tienen en cuenta la corrección por múltiples comparaciones. Un valor q < 0.05 indica que se espera que menos del 5 % de los resultados significativos sean falsos positivos.

I.3.2.4. Tamaño del efecto vs significancia

En el análisis de datos ómicos, es crucial diferenciar entre la significancia estadística y la importancia biológica de los cambios observados. Mientras que la significancia estadística nos dice si un resultado es probablemente real (es decir, no debido al azar), el tamaño del efecto nos indica cuán grande o relevante es ese cambio biológicamente.

Tamaño del efecto: $log_2(FC)$ El tamaño del efecto se mide comúnmente como el logaritmo en base 2 del cambio en la expresión ($log_2(FC)$, donde FC es el "fold change" o cambio en la expresión). Este valor indica cuánto ha aumentado o disminuido la expresión de un gen entre dos condiciones. Por ejemplo:

- Un $log_2(FC) = 1$ significa que la expresión del gen se ha duplicado.
- Un $log_2(FC) = -1$ significa que la expresión del gen se ha reducido a la mitad.

Los datos suelen transformarse a log_2 por dos razones principales:

- **Normalización:** La transformación logarítmica ayuda a estabilizar la varianza y hace que los datos se ajusten mejor a los supuestos de los tests estadísticos.
- **Interpretabilidad:** Los valores en log_2 son más fáciles de interpretar, ya que los cambios se expresan en términos de duplicaciones o reducciones a la mitad.

Importancia del $log_2(FC)$ En cualquier experimento, es esencial examinar tanto el p-valor ajustado (que indica significancia estadística) como el $log_2(FC)$ (que indica el tamaño del efecto). Sin embargo, el umbral para considerar un $log_2(FC)$ como biológicamente relevante puede variar dependiendo del contexto experimental. Por ejemplo:

- En algunos casos, un $log_2(FC)$ pequeño (por ejemplo, 0.5) puede ser biológicamente relevante si afecta a genes clave en una vía importante.
- En otros casos, solo cambios grandes (por ejemplo, $log_2(FC) > 1$) pueden considerarse relevantes.

Dependiendo del experimento, puede ser o no interesante filtrar los resultados por el $log_2(FC)$:

- **Interpretación individual de genes:** Si el objetivo es identificar genes individuales con cambios grandes en la expresión, es útil filtrar por $log_2(FC)$. Por ejemplo, se podría considerar solo aquellos genes con $|log_2(FC)| > 1$.
- **Interpretación de grupos de genes:** Si el objetivo es analizar vías o grupos de genes (por ejemplo, mediante análisis de enriquecimiento funcional), puede no ser necesario filtrar por $log_2(FC)$. En este caso, incluso cambios pequeños en múltiples genes de una misma vía pueden ser biológicamente relevantes.

I.3.3. Continuación ejemplo - Statistics for Omics

Generamos nuestra propia distribución nula. Se asigna aleatoriamente 0 y 1 a cada una de las columnas. Para esta nueva expresión barajada, se realiza el t-test para ver cuántos salen diferencialmente expresados.

```r
# r
set.seed(0)
shuffledIndex <- factor(sample(c(0,1),sum(g==0),replace=TRUE ))
nulltt <- rowttests(y[,g==0],shuffledIndex)
NfalselySigAt01 = sum(nulltt$p.value<0.01)
NfalselySigAt01 #11
NfalselySigAt05 = sum(nulltt$p.value<0.05)
NfalselySigAt05 #201 falsamente significativos
```

En lugar del p-valor, vamos a calcular el q-valor, y comprobamos que efectivamente no sale ningún gen significativo.

```r
# r
library(qvalue)
nullqvals = qvalue(nulltt$p.value)$qvalue
sum(nullqvals<0.05) #0
sum(nullqvals<0.01) #0
```

I.3.4. Ejemplo inferencia

Este ejemplo se basa en el paper "A Model for Studying Mechanisms and Treatment of Impaired Glucose Tolerance and Type 2 Diabetes". Script en carpeta de prácticas de la asignatura. En el CSV mice_pheno se muestra el sexo, dieta y peso de los ratones. El experimento está bastante equilibrado: 225 hembras con dieta control, 200 hembras con dieta grasa, 224 machos con dieta control y 197 machos con dieta grasa.

Hay varias representaciones. Entre ellas, el barplot es una figura muy poco explicativa. El boxplot es algo más indicativo, mostrando que hay poca diferencia entre los dos grupos de dietas al cruzarse las barras de error.

Como los ratones machos suelen ser más grandes (y por tanto, más pesados) que las ratonas hembra, solo seleccionamos a las últimas. Tras volver a hacer el boxplot, vemos que las medias no están tan desencaminadas, pero la variabilidad es mucho más grande que con toda la población. A continuación se muestrean 12 ratones con dieta control y 12 ratones con dieta alta en grasas de la población total.

Se realiza lo mismo para 1000 muestras y se simula la distribución con 5 réplicas en lugar de 12, siendo el resultado una distribución más amplia (hay más variabilidad si se tienen menos réplicas).

En ómicas, como hay muchos datos, se puede simular cómo sería la distribución nula y comparar si hay una diferencia real entre los grupos.

Parte I

Transcriptómica

Capítulo II

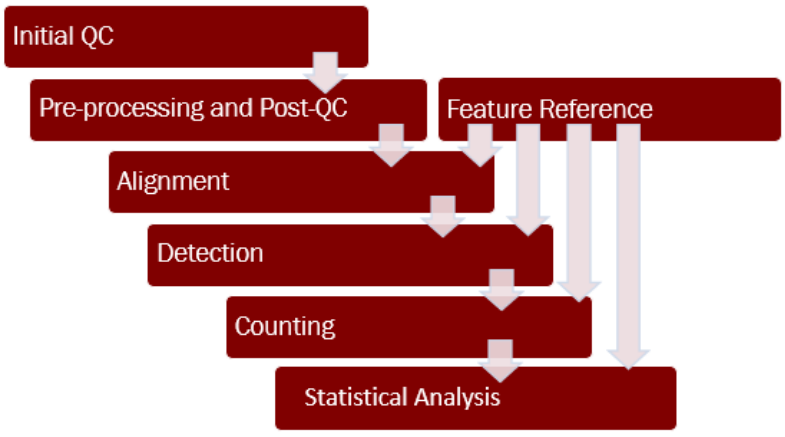
RNA-Seq

II.1. Pipeline general y alineadores

En este curso nos centraremos en los NGS de lectura corta (segunda generación). Para transcriptómica, se secuencia el cDNA generado a partir del ARNm. Este cDNA se fragmenta y se secuencia en reads cortos. Las máquinas y la forma de secuenciar es la misma que aquella vista en la asignatura "Fundamentos de Secuenciación".

Las lecturas pueden ser solo de la primera parte del fragmento (single-reads) o lecturas pareadas para tener una mayor precisión en el alineamiento. Del secuenciador sale un fichero FastQ, el cual se alinea con un genoma de referencia en FastA. Una vez con los alineamientos, se pueden mirar las regiones con reads mapeadas que estén en el transcriptoma de referencia (GTF/GFF) y cuantificar la expresión (matriz de conteo en CSV/TSV).

En general, el workflow es el siguiente: descargar datos, QC inicial, pre-procesamiento y QC posterior, alineamiento a una referencia, detección, conteo y análisis estadístico. En transcriptómica, se quiere cuantificar la expresión de las partes del genoma que se transcriben. Por ello, se necesita un genoma de referencia, pero también otro archivo GTF que relacione los exones con los tránscritos y los genes. El contaje por detección es el que se hace con CHIP-Seq, al secuenciar la parte del ADN genómico a la que se ha pegado un factor de transcripción predefinido.



II.1.1. Control de calidad inicial

Los objetivos del control de calidad de las lecturas crudas es detectar problemas de secuenciación, detectar adaptadores y comparar librerías para análisis posteriores. Distintos experimentos requieren interpretaciones distintas del análisis de control de calidad. Una herramienta muy utilizada para esto es FastQC. El análisis de la calidad por base, representa la distribución de las puntuaciones de calidad en todas las lecturas por la posición de cada lectura. En general, es normal que las últimas posiciones tengan una calidad algo peor que las demás, pero una buena muestra debe seguir teniendo una calidad alta. Si una muestra no tiene gran calidad, se puede optar por utilizar solo aquella porción de las muestras que tienen una calidad aceptable, pero hay que tener en cuenta que al acortar las lecturas, el mapeado puede darse en un mayor número de sitios. También se mide el contenido de cada base por posición (que debería ser bastante constante a lo largo de toda la lectura dependiendo de la "complejidad de la muestra", es decir, variedad de tránscritos diferentes) y la puntuación de calidad media por secuencia. La pipeline para el ARNm y los miRNA es la misma, pero hay peculiaridades. Los microARNs son ARNs de unos 20-30 nucleótidos que reprimen la expresión génica de los genes a los que se unen. Dado el bajo número de miRNAs codificados por el genoma y el más reducido número expresado en cada tejido es de esperar ver perfiles de baja complejidad en las librerías de miRNAs, dádose así un patrón irregular del contenido de bases por posición.

Las secuencias sobrerrepresentadas son listas de secuencias que están presentes más veces de lo esperado por azar. La lista sobrerrepresentada se anota con el tipo de secuencia si se proporciona una lista con la que comparar. Normalmente, las secuencias adaptadoras pueden estar sobrerrepresentadas si hay altos niveles de ligaciones de dímeros de cebadores en el paso de preparación de la biblioteca. A menudo se debe a un desequilibrio entre los niveles de adaptadores y los niveles de fragmentos de muestra. Algunos RNA-Seq de tejidos particulares pueden dar también secuencias sobrerrepresentadas. Por ejemplo, las muestras de sangre contienen grandes cantidades de transcritos de hemoglobina que siempre se reportan como lecturas sobrerrepresentadas. Las muestras de miARN siempre muestran secuencias sobrerrepresentadas. Las bibliotecas de ARN total muestran secuencias sobrerrepresentadas de ARN ribosómicos.

II.1.2. Preprocesado

El preprocesado tiene como objetivo mejorar la calidad, la mapeabilidad, quitar contaminantes y sesgos, etc. Hay diferentes herramientas, como cutadapt o trim-galore.

La calidad de las bases puede afectar al análisis de llamadas de variantes y, si es grave, también al mapeo de características. La calidad de las bases suele disminuir al final de la lectura y a veces al principio. Además, la secuenciación de baja calidad puede producir un grupo de lecturas de baja calidad a lo largo de su longitud. Es esencial eliminar las bases de baja calidad para el análisis de llamada de variantes. Para otros análisis, elimínelas sólo si afecta al rendimiento del mapeo.

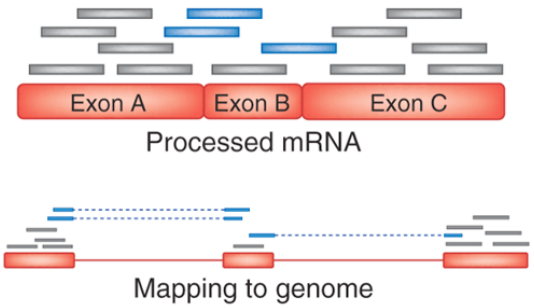
II.1.3. Alineamiento y mapeado

Hay dos tipos de alineamientos: local y global. En el caso del local, se busca que en partes específicas el alineamiento sea bueno, mientras que en el global se busca meter la lectura en la secuencia completa, metiendo gaps.

La cobertura en un segmento se mide como el número de reads que mapean a ese fragmento del genoma y la longitud de cada lectura dividido por la longitud del fragmento. Para poder hacer el mapeado se necesita la referencia en fasta, las reads en fastq y una referencia indexada. Esto es distinto de los alineadores tradicionales como BLAST. El objetivo es mapear las lecturas a las características. En transcriptómica, el alineamiento se realiza al mismo tiempo que la cuantificación.

Lo importante es la indexación del genoma de referencia para ahorrar tiempo de computación. El genoma se corta en trozos para que sea más fácil realizar las búsquedas. Esto se puede hacer por ejemplo con BWA y alineadores Bowtie que utilizan la transformación de Burrows-Wheeler al ser más rápidos.

Aunque se hable de expresión génica, los genes no se expresan, son los tránscritos. Con estas técnicas es muy complicado hilar tan fino, por lo que se cuantifican los reads al gen y contar. A la hora de alinear, si se intenta alinear reads al genoma de referencia, las reads salen del tránscrito, por lo que puede ocurrir que una parte de un read caiga en un exón y la otra parte en el otro. Esto se puede visualizar con el visor IGV. Las lecturas partidas se conocen como **exon junctions**. Por ello, se puede mapear al genoma permitiendo esa característica. Otra opción es alinear directamente al transcriptoma, pero es más grande que el genoma (puede haber 100.000 tránscritos definidos vs 20.000 - 30.000 genes) y puede que haya lecturas que no se puedan mapear a un tránscrito concreto, si no que puedan mapear a varios tránscritos con el mismo exon.



II.1.4. Galaxy

Galaxy es una plataforma con muchas herramientas y workflows ya hechos para investigación biomédica intensiva en datos. Permite generar y hacer públicas pipelines. Se puede utilizar en el servidor europeo o montar un servidor local.

Para nuestro proyecto, utilizaremos los datos del paper "Next-generation sequencing facilitates quantitative analysis of wild-type and Nrl(-/-) retinal transcriptomes". En la parte de "Related Information" se encuentran los datasets subidos a la base de datos GEO (Gene Expression Omnibus). En general, las revistas

buenas exigen poner los datos en una base de datos pública. En este caso hay 6 muestras, 3 wild-type y 3 knock-out. Cada muestra en GEO tiene un ID.

Galaxy se conecta a las bases de datos mediante API, por lo que se puede poner el link a los reads crudos y Galaxy lo lleva a nuestra sesión sin necesidad de descargarlos de las bases de datos y subirlos a Galaxy de forma manual.

Nos vamos a descargar la información de los nombres de las muestras (los metadatos). Desde GEO, hay un acceso a SRA Run Selector donde tenemos disponibles esos datos. Hay dos tablas disponibles: metadata y lista de las accesiones con los IDs de las muestras. Los metadatos se necesita posteriormente para saber qué muestras son WT y cuáles KO, pero por ahora solo necesitamos los IDs para subir a Galaxy. El siguiente paso es decirle a Galaxy que, utilizando esos identificadores, se descarguen los FastQ. Para ello, en Get Data hay una opción de Faster Download and Extract Reads in FastQ format from NCBI SRA. Para esa herramienta se selecciona la opción de "List of SRA accessions, one per line" y se ejecuta.

Una vez con los datos, vemos que en Pair-end tenemos 0 datos y en Single-end 6, indicando que las muestras son single-end (aunque esto ya lo sabíamos porque venía en SRA Run Selector). El siguiente paso es ir a FastQC con los datos de single-end. La salida es un fichero txt con los números y un html con las imágenes. Tras analizarlo brevemente, vemos que no hay ningún problema con las muestras, pudiendo continuar con el análisis.

En Ensembl nos vamos a la página de FTP Downloads donde se encuentran todas las referencias de la última versión del genoma. Para reproducir unos resultados, hay que utilizar la referencia de la fecha de publicación de los datos que se estén utilizando, pero en nuestro caso podemos utilizar la última versión generada. Como no queremos descargar el Fasta a nuestro ordenador, vamos al FastA y buscamos el fichero de primary assembly. Con click derecho, podemos copiar el enlace y en Galaxy, en Upload, se puede utilizar la función Paste/Fetch data y pegar ahí la dirección. También subimos el GTF de los cromosomas.

El siguiente paso es buscar la herramienta Trim-Galore con los datos single-end dejando todas las opciones como las predeterminadas. Con esta herramienta queremos quitar los adaptadores, y en caso de tener muestras dañadas, podríamos también eliminar esa parte.

II.2. Expresión diferencial

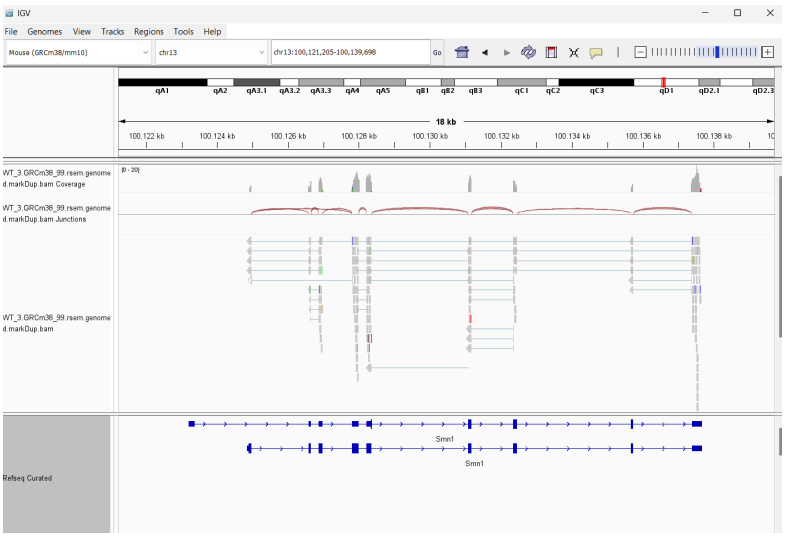
II.2.1. Visualización con IGV

Para subir un fichero a IGV, nos vamos a File y Load from File. Hay que tener en la misma carpeta el fichero BAM con el fichero BAI, es decir, el fichero indexado. Hay que cargar el BAM. Se muestran tres tracks, siendo uno la cobertura, otro los junctions y el último los duplicados (click derecho y expandir). Previamente hay que seleccionar el genoma correcto; en este caso, el de ratón.

Podemos irnos al cromosoma 13 y buscar el gen Smn1, con eso saltamos a esa región del genoma. Podemos ver las distintas isoformas del gen y dónde han mapeado

las lecturas. En Junctions, vemos que hay lecturas con un arco grande, indicando que la lectura ha mapeado a esos exones que estaban juntos en el tránscrito, pese a que en el genoma estén separados por intrones y otras regiones no codificantes. La cobertura coincide con los exones al tratarse de un RNA-Seq. Además, tiene una cobertura 0-20, indicando que en ese rango hay como máximo 20 reads y como mínimo 0.

La isoforma superior tiene un exón al principio de la proteína que no tiene ninguna lectura. Esto puede darse por la cobertura baja, indicando que nos estamos perdiendo esa isoforma.



II.2.2. Redundancia de mapeo

Pueden darse redundancias de mapeo, ya que la secuencia del genoma es larga y contiene muchas secuencias repetitivas. Por ello, hay que mirar la calidad de mapeado, ya que una lectura puede mapear en un sitio con 1 mismatch y en otra región con 2. Para reducir la ambigüedad en el mapeado (hay genes pareados, isoformas), se puede utilizar pair-end o secuenciación de reads más largas.

En NGS, se pueden detectar regiones por enriquecimiento viendo, en base a la cobertura del experimento, las regiones donde hay señal y que indicarán genes o factores de transcripción. El proceso es crosslinking, sonicación, inmunoprecipitación y secuenciación. El resultado de este tipo de experimento es un archivo tipo BED o WIG en el que se obtiene la posición donde se encuentra la señal.

En el conteo por ocurrencias, se cuentan cuántas reads caen en las distintas regiones. Para ello se requiere el GTF que asocia los disitntos exones con los tránscritos o isoformas.

En Galaxy, se puede utilizar un solo programa para alinear las lecturas a la referencia y la cuantificación. Al final, no importa si una read pertenece a una isoforma o a otra si queremos abstraer la cuantificación de una proteína, es decir, obtener la cuantificación absoluta. Para la expresión diferencial, se necesita más cobertura y los métodos son un poco diferentes para poder diferenciar las isoformas. Los tránscritos tienen una estructura de dependencia muy complicada, por lo que la expresión diferencial se suele hacer a nivel de gen.

II.2.3. Cálculo de expresión

Una vez con las lecturas mapeando a un gen, si queremos tener una medida robusta de la expresión, hay que tener en cuenta la longitud del gen y el tamaño de la librería. Una librería con una secuenciación mayor, la expresión va a parecer mayor que en una secuenciación con un tamaño menor de librería. Además, un tránscrito más corto va a tener menos reads que caigan en él por mera probabilidad, por lo que hay que normalizar por el tamaño del gen. Para ello, primero se obtienen los counts (la cobertura) y se utilizan los RPKMs:

$$RPKM : 10^9 \cdot \frac{\text{Reads mapped to the transcript}}{\text{Total reads} \cdot \text{Transcript length}}$$

Esta fórmula se modificó a la siguiente para normalizar todo a la vez:

$$TPM = 10^6 \cdot \frac{\text{reads mapped to transcript / transcript length}}{\sum \text{reads mapped to transcript/transcript length}}$$

Para las reads que mapean a varias isoformas o a varios genes del genoma, se utilizan los programas RSEM, Salmon o Sailfish. Estos métodos son procesos iterativos. Se aprovechan de la información de todos los reads para mejorar la probabilidad de qué read pertenece a qué sitio. Teniendo tres isoformas que coinciden en el primer exón, se ven las reads que caen en las partes distintas de los tránscritos para inferir las reads de la parte común de los tránscritos. La probabilidad se va cambiando y ajustando según cambian las probabilidades. Estos métodos probabilísticos hacen una estimación de los counts. Estos algoritmos cuantifican la expresión por isoforma, ya que están hechos para lidiar con el multimapping. La columna IsoPct indica el porcentaje de expresión de esa isoforma sobre toda la expresión del gen completo. Cuando se ve un proceso de splicing alternativo, la isoforma mayoritaria es la primera, y en otra condición puede darse que todas las isoformas estén igual de expresadas o que se convierta en la menos expresada.

El gen SMN1 de ratón es esencial, no puede haber ninguna mutación al ser inviable. De hecho, en humanos, una mutación en este gen causa SMA (spinal muscular atrophy). Esto se debe a que tenemos otro gen, SMN2, que solo se diferencia del 1 en una base y puede ayudar a compensar. Aunque las reads sean prácticamente idénticas, RSEM es capaz de asignarlas a una forma u otra.

II.2.4. Galaxy

Volviendo a la práctica, nos vamos a Ensembl y BioMart. Seleccionamos el genoma de ratón, seleccionamos como atributos solo Gene stable ID y transcript stable ID (quitamos los version) y exportamos los resultados como tsv. En Galaxy subimos ese fichero (Gene2Transcript) y utilizamos la herramienta "Sort Column Order by heading" para poner la columna 2 como identificador.

El siguiente paso es construir la referencia del tránscrito (el fasta del transcriptoma) a partir del GTF que habíamos subido previamente con la herramienta gffread. Debería coger automática el fichero gtf, y en caso contrario lo seleccionamos manualmente. En el apartado de Reference Genome, debemos poner "From your history" para poder

indicar el fasta. Además, en "select fasta outputs", seleccionamos la opción de "fasta file with spliced exons for each GFF transcript". También hay que activar "full GFF attribute preservation", y en Feature File Output poner GTF.

A continuación utilizamos salmon_qual utilizando el fichero exons.fa que se acaba de generar. Los alineadores/mapeadores alinean los reads base a base y cuantifican por exón, tránscrito y gen el número de reads que caen en cada una de esas regiones. Por ello, debe recibir la referencia exons.fa creado a partir del GTF y del Fasta. También recibe la tabla que mapea los tránscritos a los genes (la que hemos construido con el sort column) para conseguir la cuantificación por gen. Aunque nosotros hayamos utilizado salmon, existen otros algoritmos como sailfish y RSEM. En RNA-Seq no quitamos los duplicados al poder significar una mayor expresión. Además, los tres métodos permiten el multimapping para poder ver si las reads van a una u otra isoforma mediante métodos estadísticos de expectation-maximization.

El resultado contiene el gen (no tránscrito, eso sería otro análisis), la longitud del gen, la longitud efectiva (la que se puede mapear), los TPMs y el número de reads. Esto último es el dato crudo de cuántas reads del experimento caen en ese gen, mientras que los TPMs son los datos normalizados. Para el análisis de expresión diferencial, vamos a utilizar las reads crudas.

II.3. Análisis de expresión diferencial

II.3.1. Galaxy

Para el análisis de la expresión diferencial, necesitamos una tabla con el ID del gen y las columnas con los reads. Esto lo vamos a hacer dentro de Galaxy con la función cut con el output de salmon y escogiendo las columnas 1 y 5. El siguiente paso es utilizar columnjoin sobre este resultado, siendo la columna 1 el identificador y con 1 línea de encabezado en cada fichero de input.

Con los counts normalizados por el tamaño de gen y de librería, si hiciéramos un plot de expresión sobre lgFC, habría más variabilidad a baja expresión. Limma-voom y trend permite estabilizar la varianza para poder realizar posteriormente el

test estadístico. El ejemplo en el que estamos trabajando es bastante sencillo, pero para cuando nos veamos en situaciones más complejas (varias condiciones, medidas repeticas, etc) hay un tutorial de limma escrito por su autor, Gordon Smyth. voom permite meter información sobre la calidad de las réplicas, pero en este caso no vamos a usarlo. Seleccionamos single count matrix. El factor es la variable que determina las condiciones. Por ello, nosotros vamos a añadir un factor llamado genotype y damos para cada muestra el grupo al que pertenece. Esta información está en los metadatos; las primeras tres muestras son WT y las otras tres KO, por lo que debemos proporcionar lo siguiente: WT, WT, WT, KO, KO, KO.

En Ensembl BioMart, seleccionamos Ensembl Genes y Mouse genes. En Attributes, debemos marcar Gene stable ID y Gene Name y obtener los resultados únicos. Con los resultados descargados, en Galaxy permitimos Gene Annotations y subimos este fichero. Este paso es opcional, es simplemente para que podamos ver mejor los genes. En contrast, debemos poner KO-WT.

A la hora de hacer el análisis de expresión diferencial, cuantas más hipótesis (genes) testemos, más habrá que corregir el p-valor. Por ello, genes no expresados, no aportan información y aumentan el error de tipo 1. Por ello, se pueden quitar los filtrados. En este caso, no queremos quitar aquellos genes que entre las dos condiciones esté downregulado (y en KO no tenga expresión, por ejemplo), pero si un gen no está expresado en más de 4 muestras (de 3 que tenemos por cada condición), sí podemos quitarlos, por lo que sí permitimos el filtrado de genes poco expresados. El filtrado se hace en base de las CPM, poniendo que haya al menos 1 count por million en al menos 3 muestras. Dentro de opciones de salida, seleccionamos todos los posibles plots. Del output, siempre hay que usar el p-valor ajustado.

Por detrás, se ha ejecutado un script de R, que lo veremos más en detalle a continuación. Utilizamos el script `limma_example_rma.r` localizado en la carpeta de prácticas.

De Galaxy, nos descargamos la tabla resultante (localizada en carpeta de prácticas): Galaxy64-[limma-voom_KO-WT].tabular. Ahora tenemos una colección de genes diferencialmente expresados y tenemos dos opciones: ir mirando uno a uno los genes (por ejemplo, Nlr tiene un p-valor ajustado de 10^{-12} y un logFC de -8, es un gen utilizado para el Knock-Out y verificamos que ha funcionado bien. A partir de ahora, buscaríamos rutas de genes que se verían afectados con el KO de este gen.

Hay dos tipos de análisis funcional:

- **ORA (overrepresentation analysis):** Tenemos 6407 genes diferencialmente expresados de 18387 (obtenidos mediante un filtro de p-valor ajustado < 0.5). Cada uno de los genes se puede clasificar en función del proceso biológico en el que está involucrado (biological process; BP), la función molecular (molecular function; MF) o el compartimento celular (cellular compartment; CC). Para un gen, podemos buscar esto en Gene Ontology Overview. Dentro de cada categoría, hay muchas clases que permiten clasificar los genes de forma cada vez más específica, y un gen puede estar en varias clases. Podemos escoger el número de procesos y la profundidad que deseamos. Asignamos así cada gen a un proceso, y contamos para cada proceso cuántos DEGs hay. Este número se compara con los valores de Whole genome. Aquellas proporciones muy enriquecidas son las que se marcan como cambio significativo. Así, se puede decir que el experimento

afecta sobre todo al proceso x (por ejemplo, regulación génica). En caso de no obtener ningún DEG, se podrían utilizar filtros más laxos, o utilizar GSEA.

- **GSEA (gene set enrichment analysis):** para este método, se cogen la lista de todos los genes y se ordenan de menor a mayor en función de su logFC (aunque se puede elegir en función de qué hacer la ordenación). Seguimos teniendo la lista con los procesos a los que pertenecen los genes. Por ello, podemos ver dónde caen los genes de cada proceso. Si para un proceso todos los genes se encuentran cerca, esto se puede reportar (tienen una magnitud de cambio similar; el experimento perturba ese proceso).

En el NIH está la herramienta DAVID que permite subir una lista o un background (pestaña functional annotation). En el caso de los arrays, no se miran todos los genes, por lo que no tiene sentido comparar en el ORA con whole genome, si no que se utiliza otro background (por ejemplo, solo el cromosoma 1, lo que se utilice como total). En nuestro caso copiamos los primeros 150 gene ID, pero también se podría subir en forma de fichero. Se selecciona que los identificadores sean de genes de Ensembl y que se trata de una lista, no background. Hay 14 genes que no se pueden detectar con esta herramienta, pero es algo asumible. Aparece una pestaña de Gene Ontology que especifica los distintos niveles de las tres partes (BP, MF y CC). Por ejemplo, para BP, hay 5 niveles, cada uno a un mayor nivel. En direct, se mezclan los distintos niveles para obtener una clasificación que sea detallada, pero sin ser exhaustiva. Como se están testando muchos genes, hay que volver a hacer el ajuste del p-valor, que en este caso aparece en una columna con la corrección de Benjamini. Esto sería el equivalente al ORA.

Capítulo III

ChIP-Seq

III.1. Procedimiento experimental

ChIP-Seq (Chromatin Immunoprecipitation Sequencing) es una técnica diseñada para identificar las regiones del ADN donde se unen los factores de transcripción. Utiliza secuenciación de nueva generación (NGS) de lecturas cortas (short reads) y se basa en la inmunoprecipitación de factores de transcripción. Si no se dispone de un anticuerpo específico para el factor de transcripción de interés, la técnica no puede llevarse a cabo.

El procedimiento experimental es el siguiente:

1. **Fijación y fragmentación del ADN:** Se fijan las células de un tejido con un agente químico para mantener el factor de transcripción unido al ADN. Se extrae el ADN junto con las proteínas unidas y se fragmenta mediante sonicación, obteniendo fragmentos de aproximadamente 200 pares de bases, ideal para la secuenciación con tecnología Illumina. Este tamaño permite además acotar la región donde buscar posteriormente los enriquecimientos.

2. **Inmunoprecipitación:** Se introduce un anticuerpo específico para el factor de transcripción en la solución de fragmentos de ADN. Como control, se puede preparar una muestra sin anticuerpo para evaluar la distribución del coverage del genoma. Los fragmentos de ADN unidos al factor de transcripción se capturan utilizando una columna con bolitas que se unen a la región constante del anticuerpo.

3. **Preparación de la librería y secuenciación:** Se deshace la fijación para eliminar las proteínas y se prepara la librería de ADN para secuenciar. Se realiza la secuenciación, idealmente obteniendo entre 20 y 40 millones de lecturas (reads).

III.2. Análisis de Datos

Alineamiento de lecturas Las lecturas se alinean al genoma de referencia. En general, se utiliza secuenciación single-end, donde las lecturas no están pareadas. Se

observa un patrón donde las lecturas se alinean en la región 5' en un sentido y en la región 3' en sentido inverso, lo que indica la señal de enriquecimiento.

Identificación de regiones enriquecidas Se utilizan herramientas como MACS (Model-based Analysis of ChIP-Seq) para identificar regiones enriquecidas. MACS modela una distribución nula, fragmenta el genoma en bins y calcula el número de lecturas en cada bin. La herramienta desplaza y junta los picos de las distribuciones (una por cada sentido) para aumentar la señal y evaluar las regiones enriquecidas en comparación con el control.

Análisis de motivos de unión Se identifican las secuencias de ADN donde se une el factor de transcripción y los genes asociados. Se utilizan herramientas para buscar motivos enriquecidos de 10-15 nucleótidos mediante el algoritmo de expectation-maximization. Esto permite estimar la secuencia consenso de unión y predecir otros sitios potenciales en el genoma mediante el logo generado.

III.3. Aplicación práctica: Pipeline de análisis con Galaxy

III.3.1. Obtención de datos

Vamos a construir una pipeline para analizar datos de ChIP-Seq del artículo "Analysis of the DNA-binding profile and function of tale homeoproteins reveals their specialization and specific interactions with hox genes/proteins". Este estudio evalúa los sitios de unión de tres factores de transcripción: Prep1, Meis1 y Pbx1, relacionados con los genes Hox del desarrollo embrionario.

En este estudio se realizaron tres ChIP-Seqs con tres anticuerpos, uno para cada factor. Meis1 salió muy ruidoso, Prep1 tenía unos picos muy marcados, y Pbx1 mostró algo intermedio. Los picos de Meis tenían dos distribuciones, y donde estaba Prep, Meis también tenía un pico marcado.

Para esta pipeline utilizamos el servidor americano de Galaxy. Ahí podemos buscar en historias públicas la historia "Pbx1 ChIPSeq Raw Data" del usuario "cartof", que es Carlos Torroja (el profesor). Los datos del artículo están disponibles en GEO, y podemos acceder desde la sección de "Additional Information" de PubMed. Desde ahí podemos navegar a SRA Run Selector, donde aparecen todas las muestras del artículo. Hay dos tipos de assays, RNA-Seq y ChIP-Seq, por lo que seleccionamos solo aquellas pertenecientes a ChIP-Seq. Hay un botón de "Computing Galaxy" que importa los datos directamente al servidor americano; para utilizar el servidor europeo, habría que descargar las tablas de metadatos e importarlas manualmente como se hizo en la pipeline de RNA-Seq.

III.3.2. Preprocesamiento de datos

Los datos ya están subidos a Galaxy, pero hay que preprocesar el contenido. Las anotaciones de los campos son libres, no hay estándar, por lo que la pipeline se debe adaptar a los datos concretos. En este caso, solo nos queremos quedar con tres columnas: la columna 1 run con los identificadores SRR, la columna 17 con el nombre de la librería y la columna 32 con el nombre del anticuerpo utilizado. Para esto, utilizamos la herramienta `cut columns`. Tras ver el resultado, vemos que algunas celdas contienen "signos prohibidos" en bioinformática, como espacios o paréntesis. Por ello, el siguiente paso es utilizar `column regex find and replace`. Esto se debe ejecutar dos veces:

1. **Columna 2**: debemos realizar dos comprobaciones. Primero, se sustituye .*WT\s([\w\s]+) por \1. El segundo check debe buscar \s y reemplazar por _.
2. **Columna 3**: buscamos [\/\s].+ y lo queremos reemplazar por nada, por lo que se deja el campo en blanco.

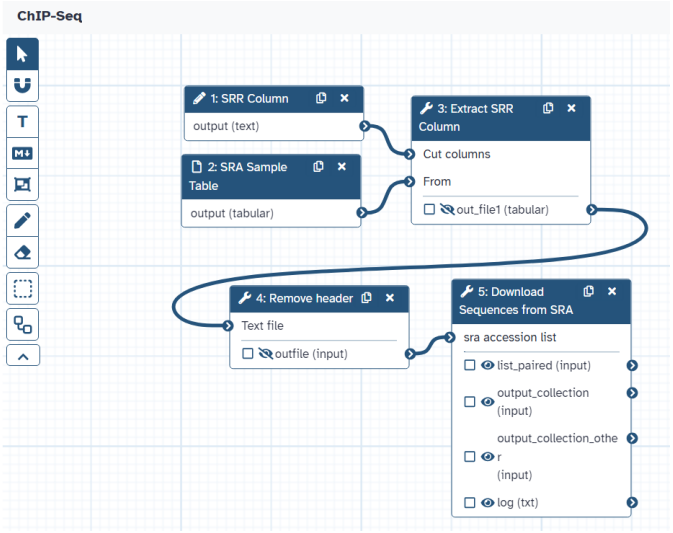
III.3.3. Descarga y extracción de lecturas

Hasta ahora, hemos realizado los pasos anteriores directamente en la historia. Los siguientes pasos los vamos a crear como parte de un flujo de trabajo, creando un nuevo workflow desde la pestaña con ese mismo nombre. Los flujos de trabajo se construyen con casillas que representan distintas entradas o herramientas, pudiendo ordenar todo.

Primero incluimos una casilla de "input database" que llamamos SRA Table con descripción "Sample table with SRR column and sample name column" y formato tabular. A continuación añadimos la herramienta "cut" con nombre "Extract SRR Column", desactivando los parámetros de la herramienta. De esta forma, a la hora de lanzar el flujo de trabajo, se pide al usuario la introducción de las columnas que se desean mantener. La alternativa sería, en lugar de desactivar el parámetro, generar una entrada que se utilice en la herramienta. Esto se consigue pulsando el icono de las dos flechas, permitiendo así enlazar una casilla de "input" con la herramienta.

La primera fila del fichero contiene la cabecera de las columnas, y no queremos que forme parte del análisis bioinformático. Para ello, incluimos una casilla con la herramienta "select last lines from a database (tail)". Entre los parámetros, debemos especificar "keep everything from this line on" y "2", de manera que siempre se vaya a eliminar la cabecera.

El flujo de trabajo contiene por ahora solo pasos intermedios. Para evitar su aparición en la historia cuando se lance el flujo, generando ruido visual, se pueden ocultar pulsando sobre el símbolo del ojo. Ahora ya podemos incluir la herramienta "Faster Download and Extract Reads in FastQ format". Por defecto, no permite introducirle una entrada, primero hay que seleccionar la opción de "list of SRA accession, one per line". La salida de esta herramienta se divide en cuatro. Para una secuenciación single-end, debemos trabajar con la salida "output_collection", mientras que para una secuenciación pair-end, "list_paired".



III.3.4. Análisis posterior

Los siguientes pasos del análisis serían realizar un control de calidad con FastQC y utilizar una herramienta de trimmeado (como Trim-Galore) para eliminar los adaptadores de la secuencia, aumentando así la capacidad de alinear. A continuación se podría pasar al alineado para saber la parte del genoma de donde provienen las lecturas mediante herramientas como BWA. En el FastQC se podría ver el tamaño de las regiones, que en este caso son de unos 35 nucleótidos, por lo que se puede utilizar la herramienta BWA estándar al no necesitar incluir gaps u otros eventos complejos más propios de secuencias más largas.

El resultado es un fichero BAM que se utiliza a continuación para realizar un control de calidad con la herramienta `plotFingerprint`. Del genoma se obtienen fragmentos pequeños que se utilizan para contar las lecturas que caen en cada fragmento. Esto se calcula para todas las muestras, y posteriormente se genera un gráfico que muestre la información: fragmentos por counts y un gráfico cumulativo.

En un análisis de ChIP-Seq, no hay un genoma de referencia, por lo que es necesario incluir un paso de detección de los picos de lectura. Esta detección se realiza en base a un patrón o en base al enriquecimiento. Esto se realiza con la herramienta MACS, la cual se encarga de la detección, del conteo y del análisis estadístico. Se identifican las regiones enriquecidas a través de la distribución de las lecturas. Se deben encontrar dos distribuciones, una para cada sentido de lectura. MACS puede detectar una distribución y emplearla para buscar una distribución opuesta en el rango de la sonicación (es decir, unos 200 pares de bases que se especifican como parámetro). A continuación junta ambas distribuciones en un pico central para aumentar la detección. El siguiente paso es la búsqueda en distintos tamaños de fragmento, quedándose con el que mejor se adapte y simulando la distribución nula con los datos. Además, se pueden modificar el tamaño efectivo del genoma de referencia e incluir en las salidas adicionales la de "peak summits".

El resultado es un fichero bed que se debe ordenar según el score. Este fichero se emplea para la búsqueda de los motivos de unión del factor de transcripción. En el fichero bed se incluyen las posiciones del nucleótido donde empieza y donde termina, por lo que debemos aumentar el número en 50 nucleótidos en ambas direcciones (restar

50 a la posición de inicio y sumar 50 a la de fin). De la colección de regiones, se pueden filtrar las 500 entradas más altas y utilizar la herramienta "Extract Genomic DNA" para obtener la secuencia de esas posiciones. Con esto se puede utilizar la herramienta "MEME", obteniendo así los logos. Pulsando la flecha en Submit/Download, podemos buscar ese logo en la base de datos para encontrar otras secuencias iguales recogidas en ella.