Módulo:
Expresión diferencial

Bioinformática y Estadística 2

Dra. Evelia Coss

Dra. Alejandra Medina

21 al 24 de Febrero, 2023

Día 3

- Fuentes de error y variación
- Normalización
- Efecto *batch*
- Importación de datos de kallisto en R

Fuentes de error

Existen dos fuentes principales de error:

- **Error humano:** mezcla de muestras (en el laboratorio o cuando se recibieron los archivos), errores en el protocolo.

- **Error técnico:** Errores inherentes a la plataforma (e.g. secuencias de mononucleotidos en pyrosecuenciacion) –
- Todas las plataformas tienen cierto de nivel de error que se debe tomar en cuenta cuando se esta diseñando el experimento.

Errores en preparación de la muestra

- Error del usuario (e.g. etiquetar equivocadamente una muestra)
- Degradación de ADN/ARN por métodos de preservación
- Contaminación con secuencias externas
- Baja cantidad de ADN de inicio

Errores en preparación de la librería

- Error del usuario (e.g. contaminar una muestra con otra, contaminar con reacciones previas, errores en el protocolo)
- Errores de amplificación por PCR
- Sesgo por (cebadores) primers (sesgo de unión, sesgo por metilación, dímeros de cebadores [primer dimers])
- Sesgo por captura (Poly-A, Ribozero)
- Errores de máquina (configuración errónea, interrupción de la reacción)
- Quimeras
- Errores de índice, adaptador (contaminación de adaptadores, falta de diversidad de índices, códigos (barcodes) incompatibles, sobrecargo)

Errores de secuenciación e imagen

- Error del usuario (e.g. sobrecarga de la celda)
- Desfase (e.g. extensión incompleta, adición de múltiples nucleótidos)
- Fluoróforos muertos, nucleótidos dañados y señales superpuestas
- Contexto de la secuencia (e.g. alto contenido de GC, secuencias homologas y de baja complejidad, homopolímeros).
- Errores de máquina (e.g. laser, disco duro, programas)
- Sesgos de cadena

El reto – diferenciar señales biológicas de ruido/errores

- Controles negativos y positivos - ¿Qué espero?
- Réplicas técnicas y biológicas - ayudan a determinar la tasa de ruido
- Conocer los tipos de errores comunes en determinada plataforma.

Normalización

¿Por qué es necesario normalizar?

Nuestros datos están sujetos a **sesgos técnicos y biológicos** que provocan variabilidad en las cuentas.

Si queremos hacer **comparaciones de niveles de expresión entre muestras** es necesario ajustar los datos tomando en cuenta estos sesgos.

- Análisis de expresión diferencial
- Visualización de datos
- En general, siempre que estemos comparando expresión

Métodos de normalización

Normalization method	Description	Accounted factors	Recommendations for use
CPM (counts per million)	counts scaled by total number of reads	sequencing depth	gene count comparisons between replicates of the same samplegroup; **NOT for within sample comparisons or DE analysis**
TPM (transcripts per kilobase million)	counts per length of transcript (kb) per million reads mapped	sequencing depth and gene length	gene count comparisons within a sample or between samples of the same sample group; **NOT for DE analysis**
RPKM/FPKM (reads/fragments per kilobase of exon per million reads/fragments mapped)	similar to TPM	sequencing depth and gene length	gene count comparisons between genes within a sample; **NOT for between sample comparisons or DE analysis**
DESeq2's **median of ratios** [1]	counts divided by sample-specific size factors determined by median ratio of gene counts relative to geometric mean per gene	sequencing depth and RNA composition	gene count comparisons between samples and for **DE analysis**; **NOT for within sample comparisons**
EdgeR's **trimmed mean of M values (TMM)** [2]	uses a weighted trimmed mean of the log expression ratios between samples	sequencing depth, RNA composition, and gene length	gene count comparisons between and within samples and for **DE analysis**

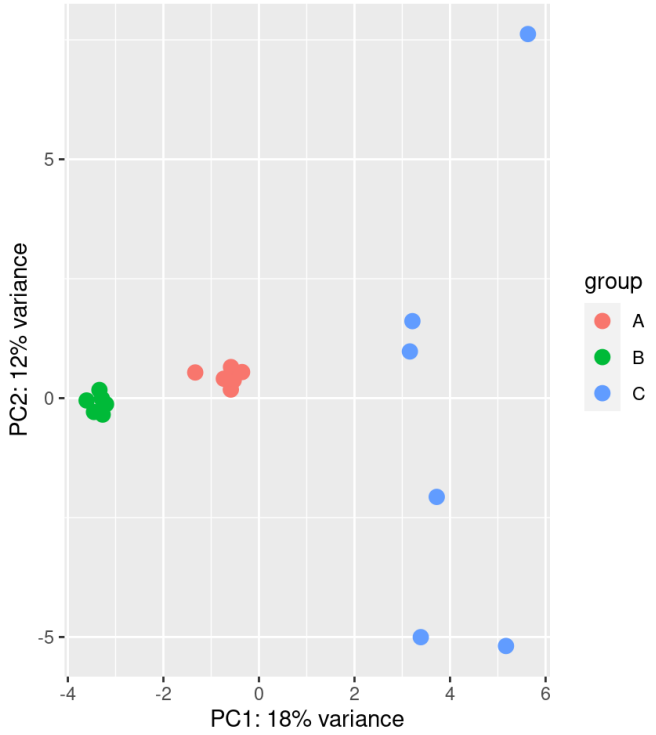
Análisis de Componentes Principales (PCA)

Es una herramienta para el **análisis exploratorio** de los datos que permite visualizar la **variación** presente de un set de datos con muchas **variables**.

En X es la mayor proporción de la varianza.
En Y la menor variabilidad.

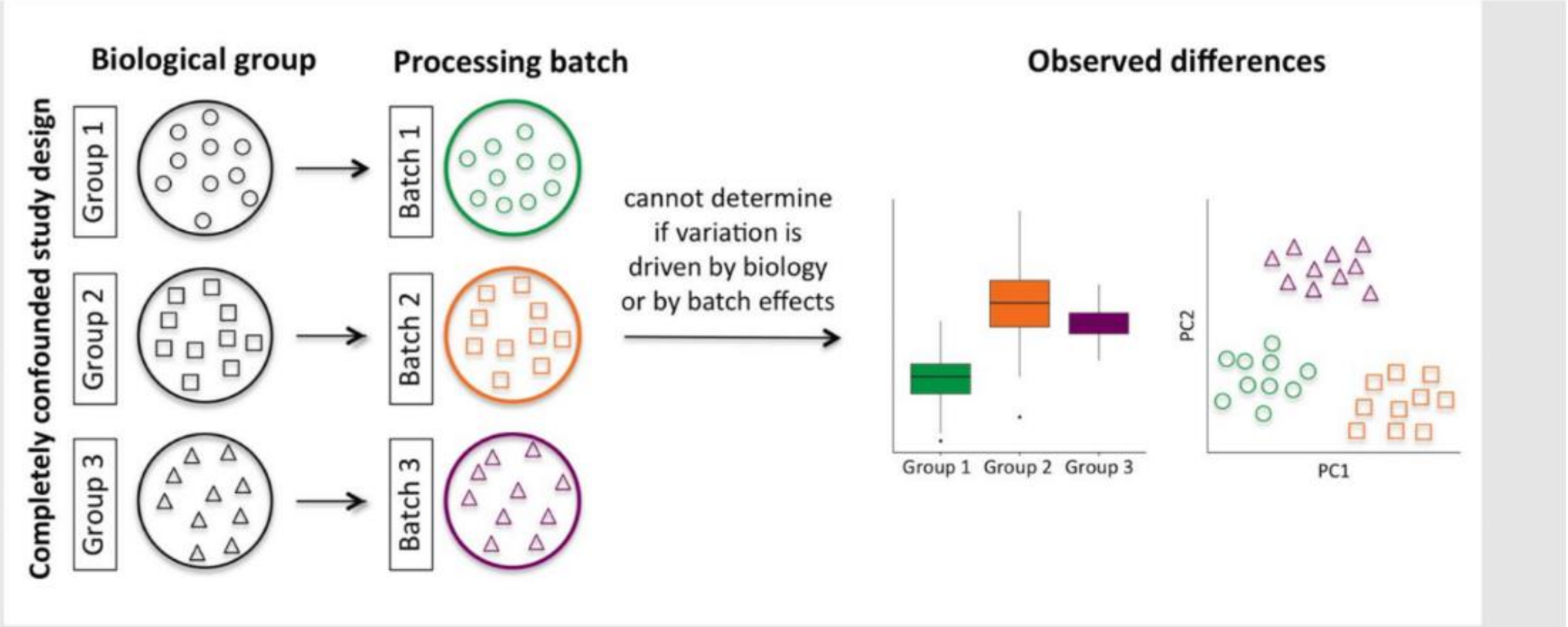
Cada dimensión o componente principal generado por PCA será una **combinación lineal de las variables originales.**

PCA reduce la dimensionalidad pero **no reduce el número de variables en los datos**.



Corrección por batch effect

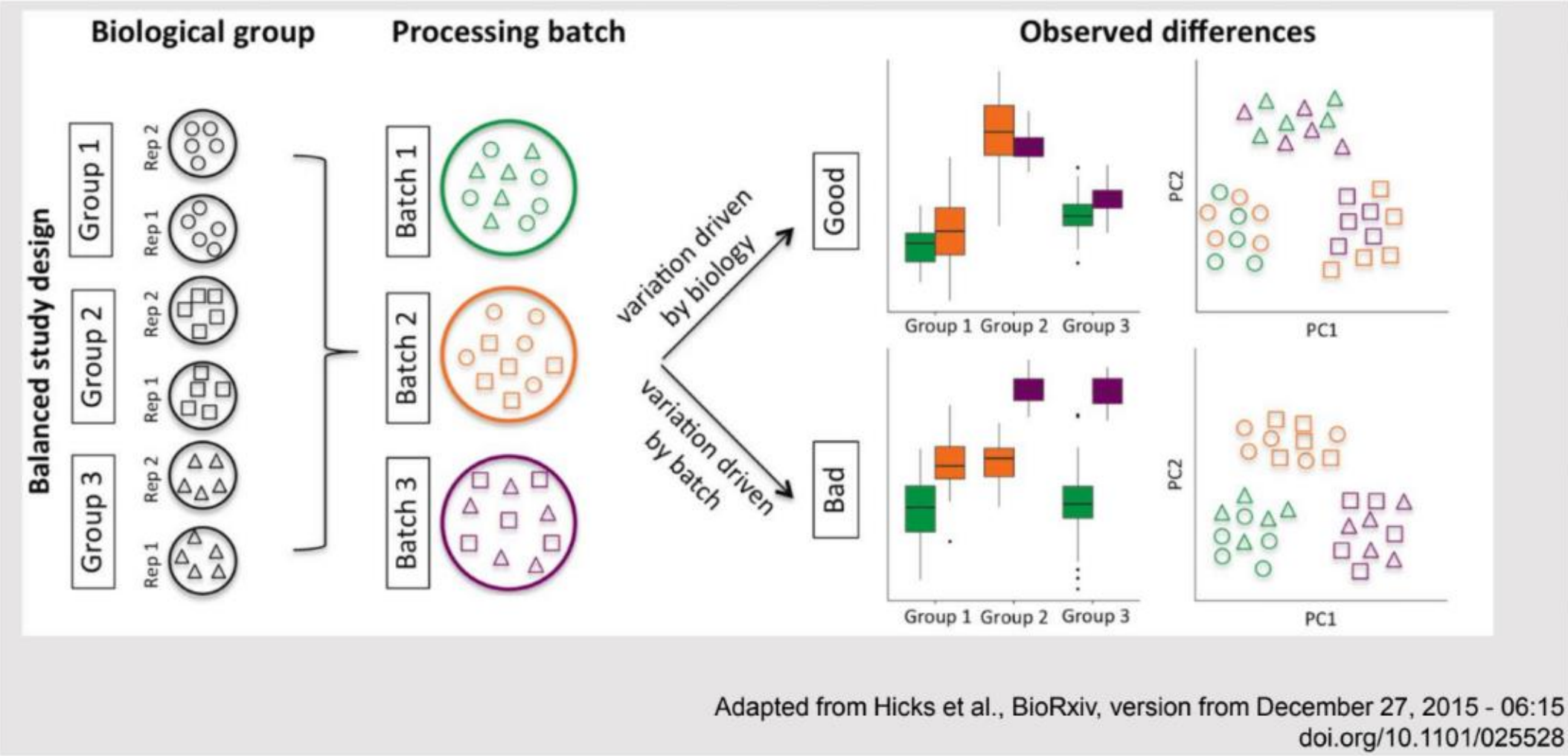
Diferentes **líneas de secuenciación** con nuestros datos resulta en una baja certeza sobre el Resultado.



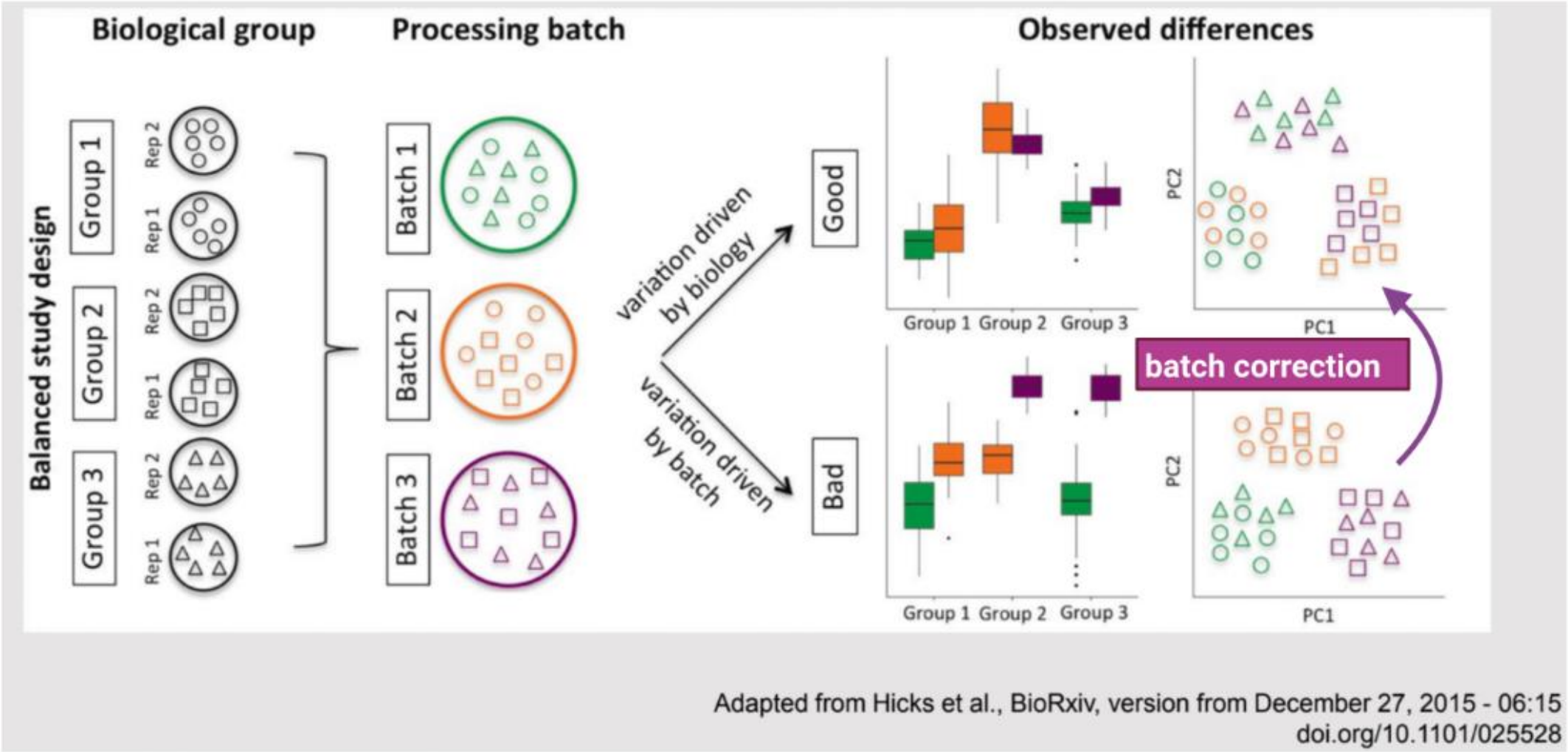
Adapted from Hicks et al., BioRxiv, version from December 27, 2015 - 06:15
doi.org/10.1101/025528

Corrección por batch effect

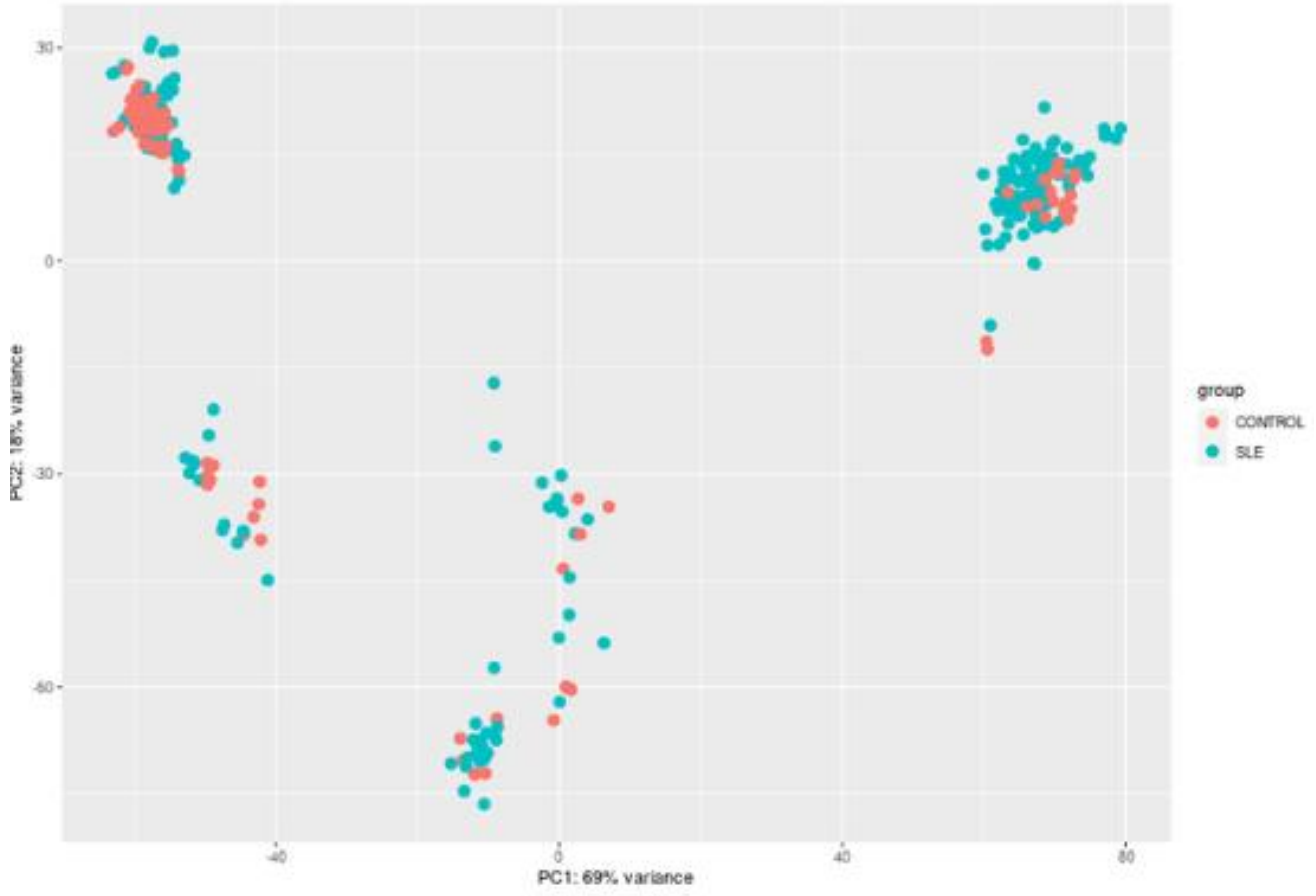
Buen diseño experimental, pero aun puede haber variación técnica.



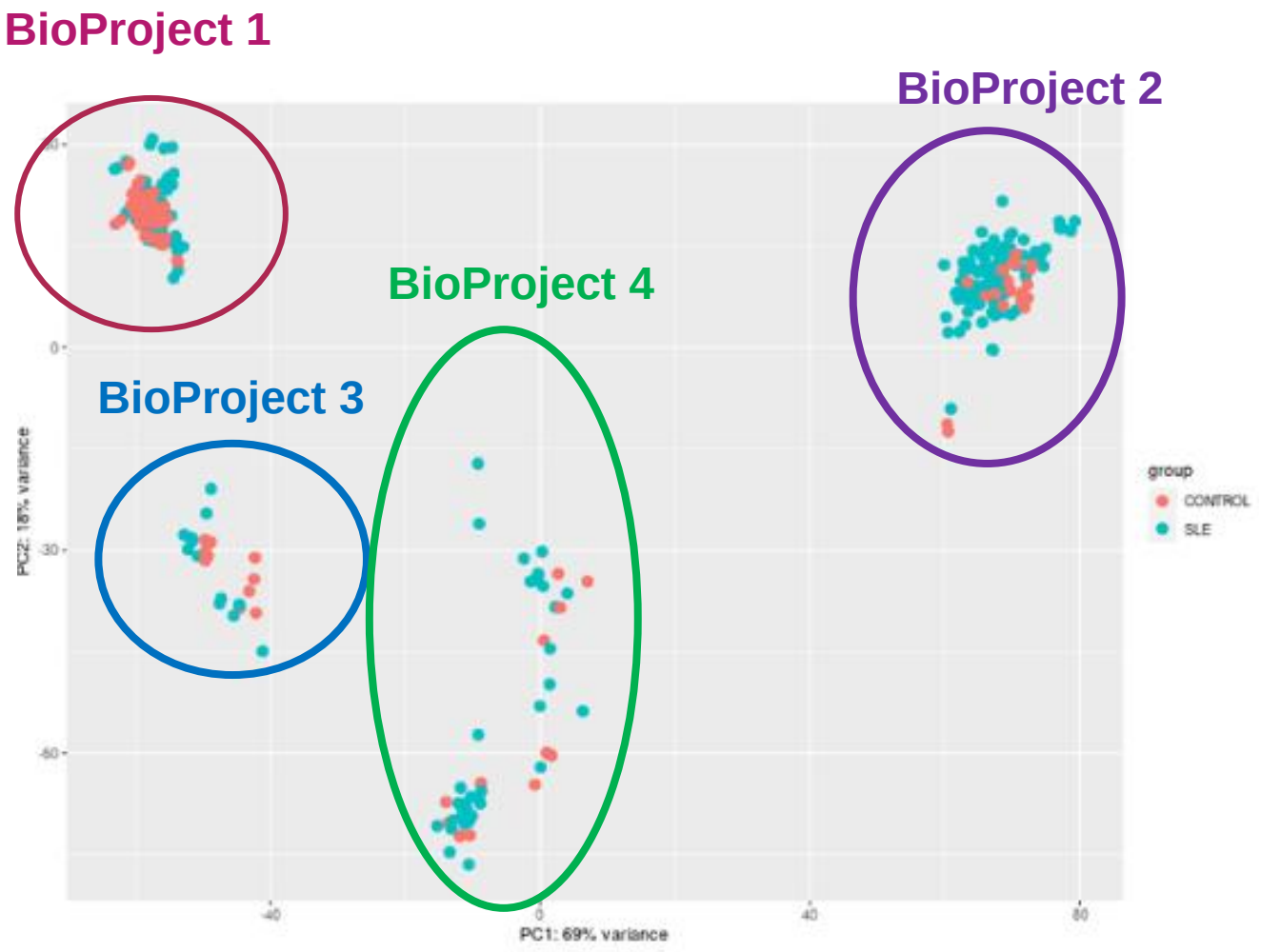
Corrección por batch effect



Efecto *batch* en análisis masivo de transcriptomas

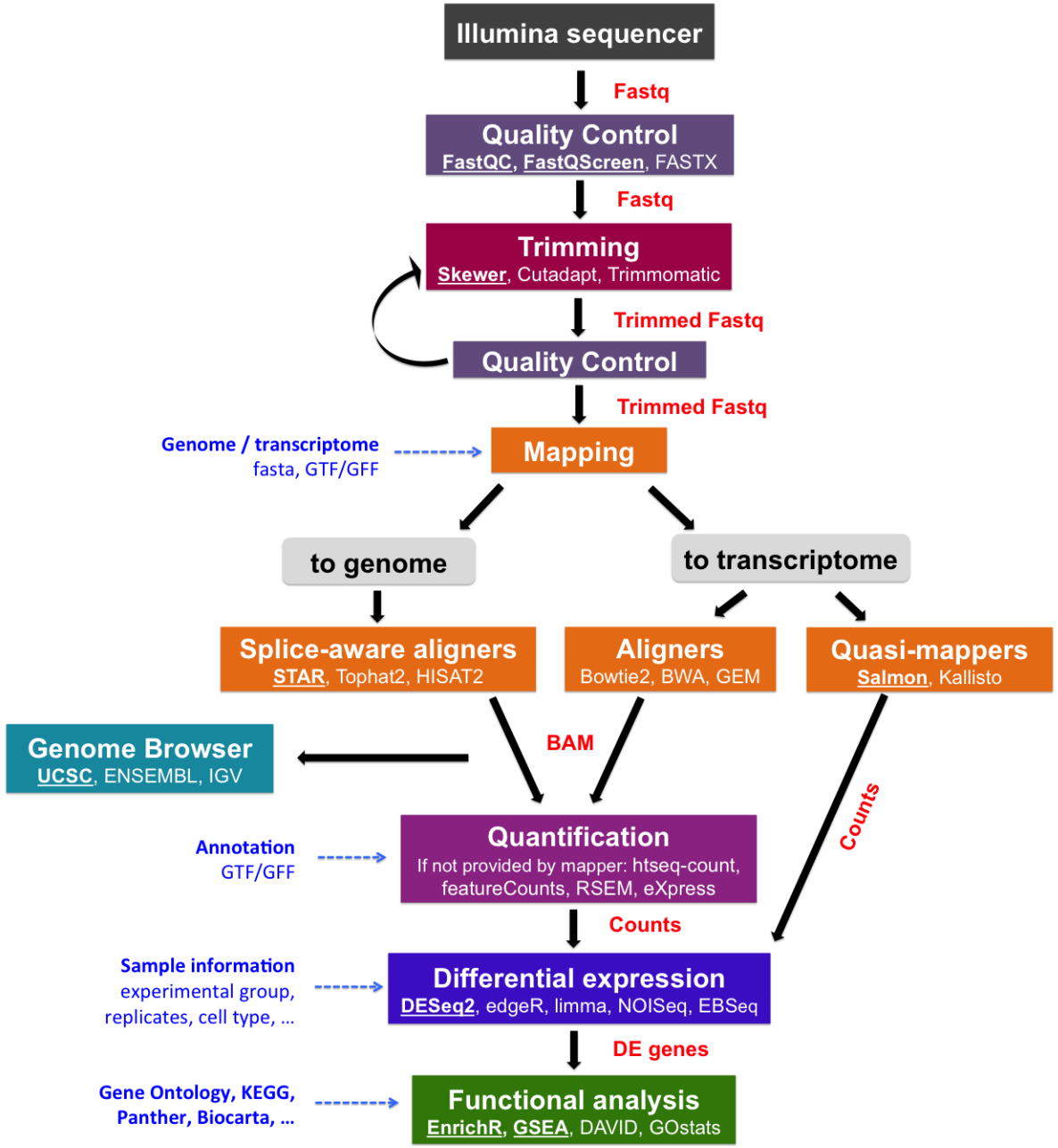


Efecto *batch* en análisis masivo de transcriptomas



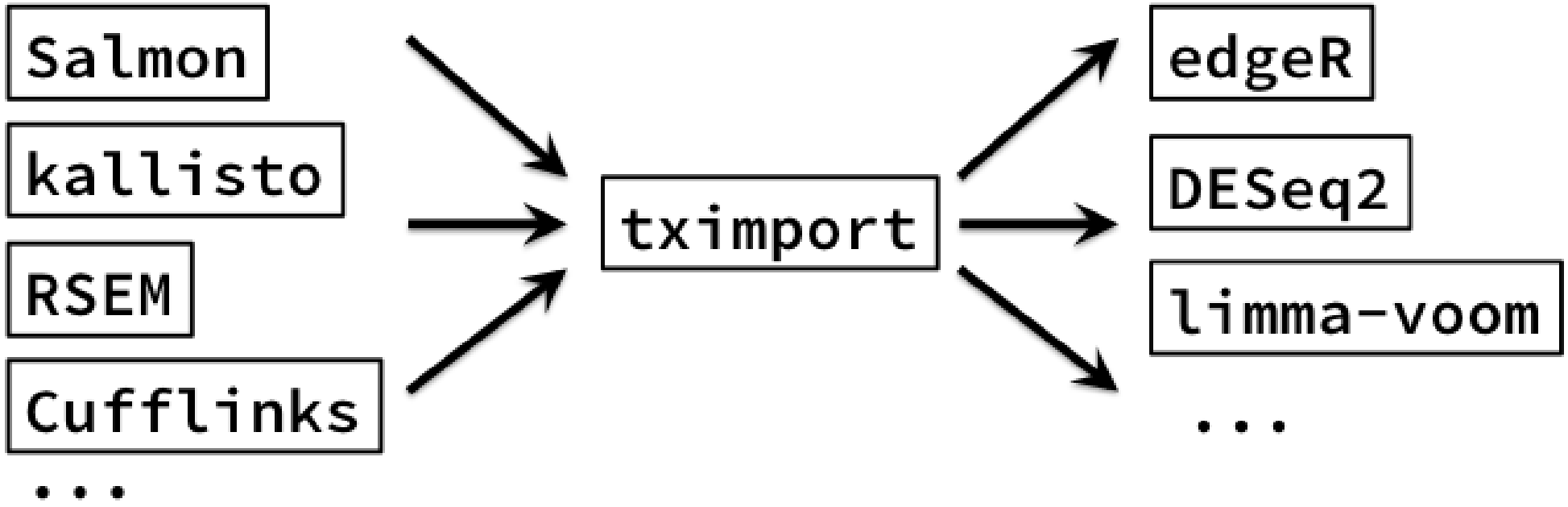
Pipeline bioinformática

Dónde estamos...

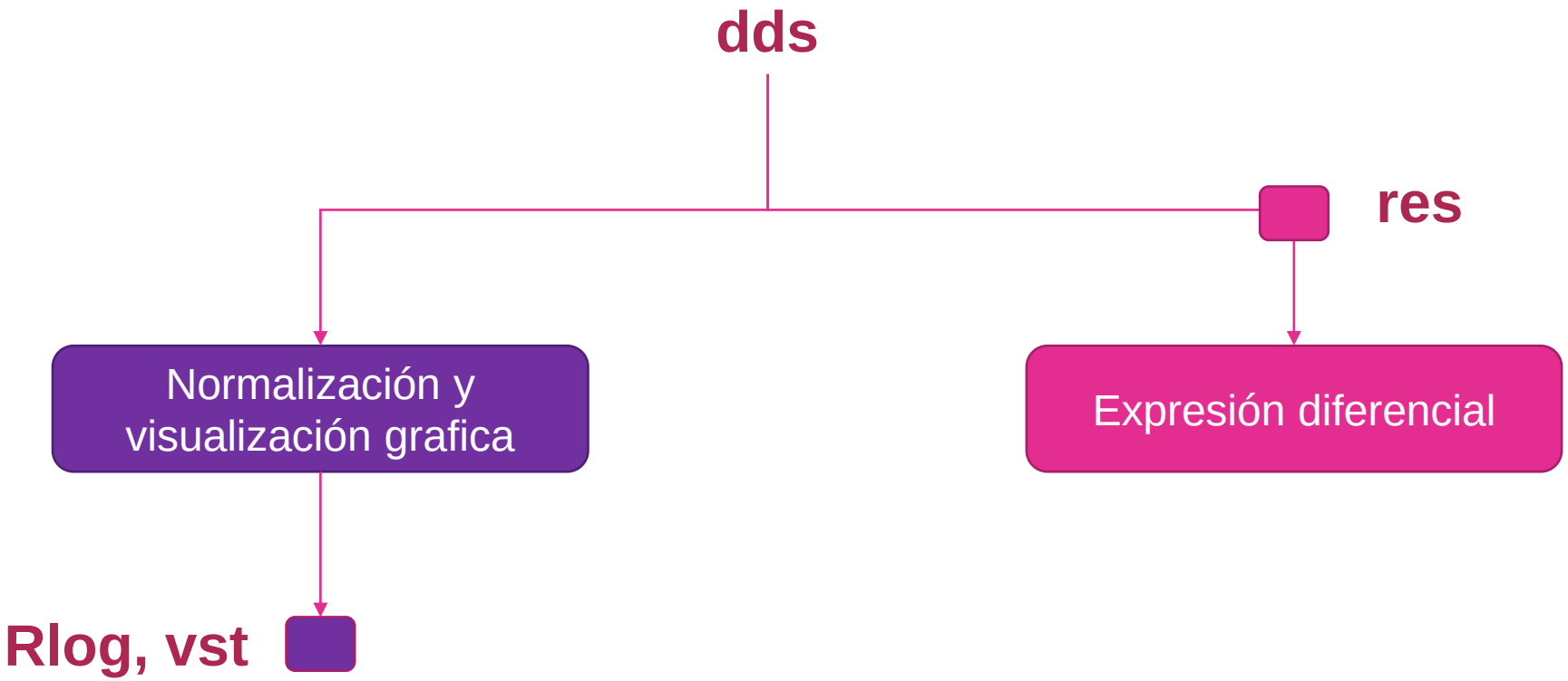


mRNA-Seq data analysis workflow "https://biocorecrg.github.io/RNAseq_course_2019/workflow.html"

Importación de datos a R



DESEq2


dds
res
Normalización y visualización grafica
Expresión diferencial
Rlog, vst

Práctica

https://github.com/EveliaCoss/RNAseq_classFEB2023/blob/main/RNA_seq/README.md#practica3