scRNA-Seq y OSCA

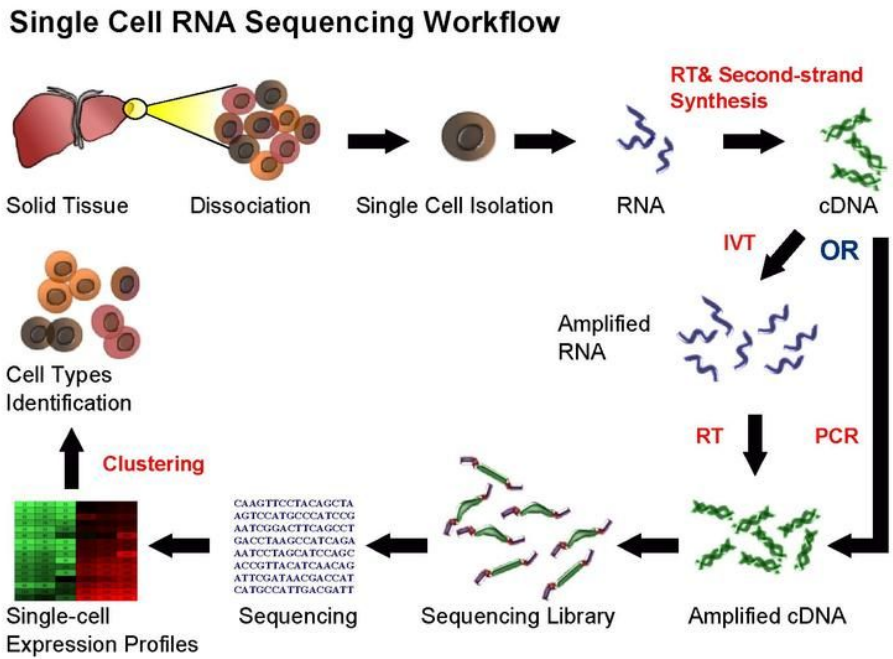
Citlali Gil Aguillon

Elisa Marquez Zavala

Diapositivas basadas de:

- **Libro Orchestrating Single Cell Analysis with Bioconductor de Aaron Lun, Robert Amezquita, Stephanie Hicks y Raphael Gottardo.**
  - **DOI: 10.1038/s41592-019-0654-x**
- **Curso de scRNA-seq para WEHI creado por Peter Hickey.**

single cell RNA-Seq



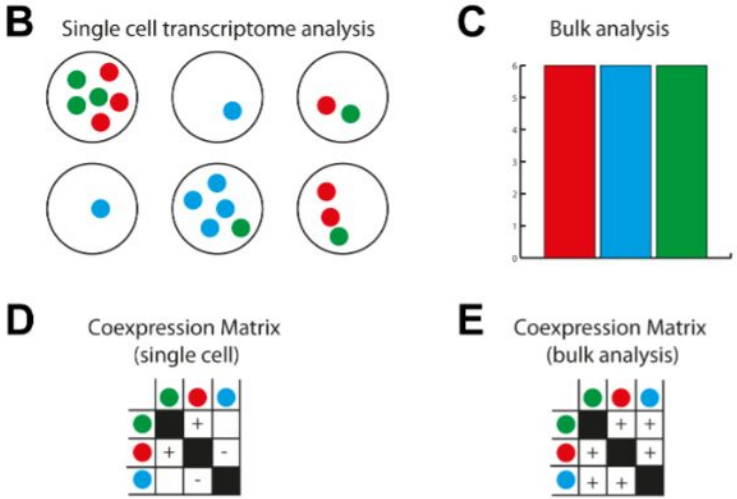
https://scrnaseq-course.cog.sanger.ac.uk/website/introduction-to-single-cell-rna-seq.html

qué cosas se pueden responder?

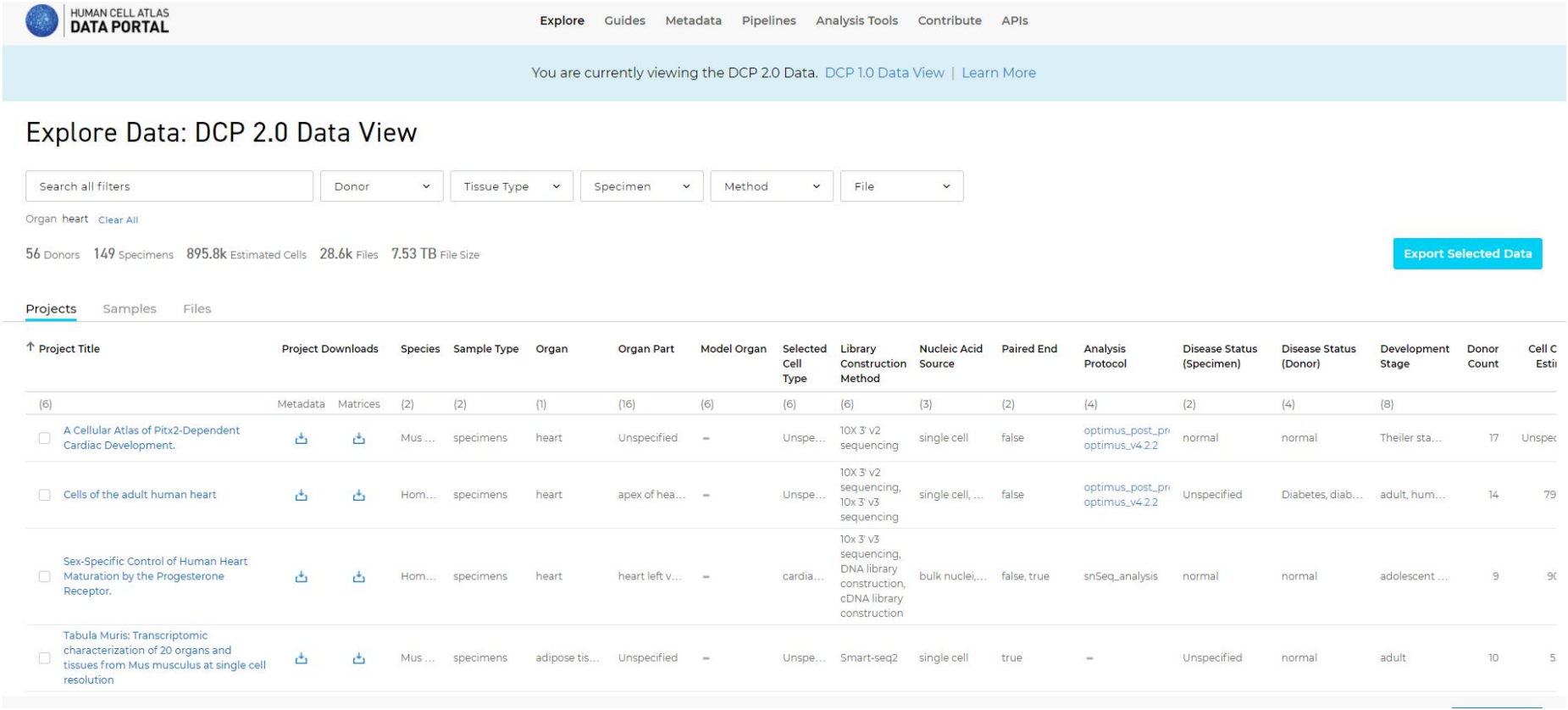
HETEROGENEIDAD

DESARROLLO

- descubrir estadios celulares
- investigar desarrollo temprano (por bajo número celular)

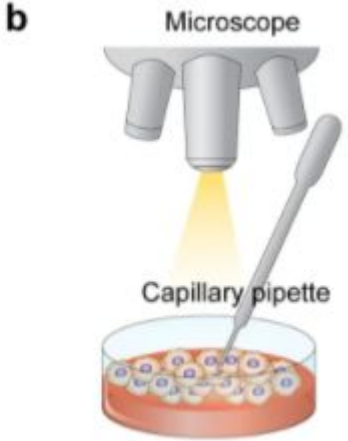


EJEMPLO

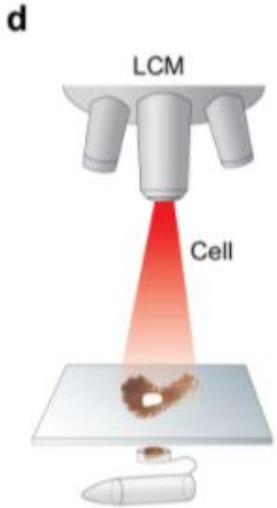


https://data.humancellatlas.org/explore/projects

por dilución



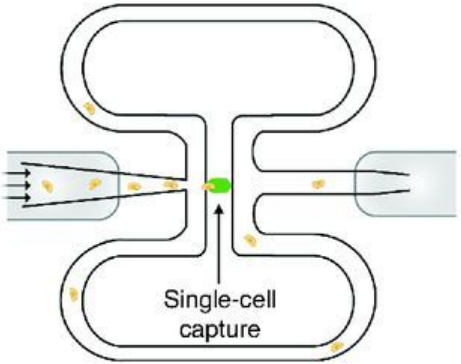
por microscopia



microdisección por captura láser LCM

n reduce

Circuit microfluidics

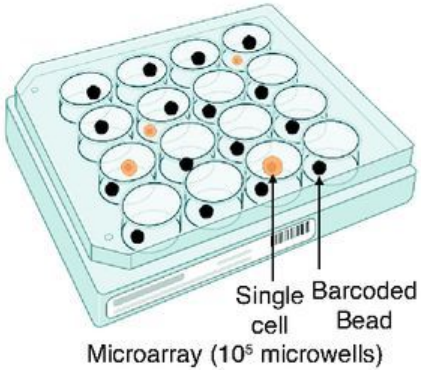


requiere menos células

sistemas integrados para captura y preparación de librerías

poca contaminacion!

Microwell(nanowell)

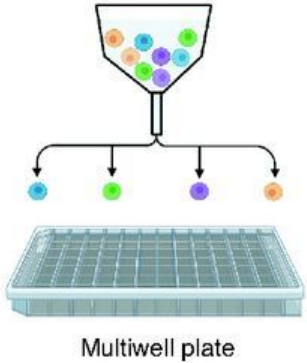


reduce tasa de cell doublets

mejora la captura de células

(checar morfologia y doubletes por microscopio, permite eliminarles)

Flow cytometry



separar con marcadores en superficie celular

https://kidney360.asnjournals.org/content/2/7/1196

¿En qué varían las tecnologías entre sí?

- aislamiento celular
- lisis celular
- transcripción reversa
- amplificación
- cobertura de transcritos
- transcrito completo o 3’ o 5’
- UMI (identificado único molecular)

Isolation method	Throughput (cells/run)	Applied sequencing technology	Commercial Platform	Ref
Limiting dilution	Low (10–200)	None	None	[7]
Micromanipulation	Low (10–200)	Smart-seq2	None	[8]
Laser capture microdissection (LCM)	Low (10–200)	Smart-seq2	None	[9]
Flow-activated cell sorting (FACS)	Medium (100–1000)	Smart-seq2 CEL-seq2 STRT-seq MARS-seq	None	[10]
Circuit microfluidics	Medium (100–1000)	Smart-seq2 CEL-seq2 STRT-seq	Fluidigm C1 system (Fluidigm)	[16]
Microdroplet microfluidics	High (1000–9000)	Drop-Seq InDrop-seq	Chromium system (10x Genomics) InDrop system (1cellBio) Nadia (Dolomite Bio)	[134]
Microwell platform	High (1000–9000)	Cyto-seq SEQ-well	None	[12], [13]
In-situ barcoding	Very high (>10000)	SPLiT-seq Sci-RNA-seq	None	[14], [15]

https://www.sciencedirect.com/science/article/pii/S1368837519303422

TABLE 2 Comparison between plate-based Smart-seq2 and droplet-based scRNA-seq platforms

	Smart-seq2	inDrop	Drop-seq	10× genomics
cDNA coverage	Full length	3′ end	3′ end	3′ or 5′ end
Plate or droplet	96- or 384-well plate	Droplet	Droplet	Droplet
UMI	None	Yes	Yes	Yes
Throughput (number of cells)	96 or 384	1k–10k	1k–10k	1k–10k
Sequencing depth (read per cell)	10^6	10^4–10^5	10^4–10^5	10^4–10^5
Feature	FACS sorting, isoform analysis	Emulsion, low cost	Emulsion, low cost	Emulsion, low cost
Long-term storage	Yes, cells sorted into lysis buffer	No, must process immediately	No, must process immediately	No, must process immediately

Transcrito completo vs 3’ 5’

Transcritos completos (tecnología: smart-seq2) útil para:

- baja expresión
- análisis de isoformas (exon skipping, intron retention)
- expresión alélica

Transcritos solo 3 o 5’ (tecnología: Chromium, Dropseq, STRT-seq):

- uso de UMIs (identificadores únicos moleculars) ayuda a checar problemas técnicos. (secuencias conocidas que se añaden antes de amplificación PCR, notaremos si hay duplicados)
- expresión diferencial de genes

UMIs y spike-in

Para checar errores durante el proceso técnico, disminuir ruido al poder tener más control en los procesos de calidad.

spike-in (smart-seq2, SUPeR-seq):

- ácidos nucleicos sintéticos que se añaden a la muestra. como control, si sale alterado posible problema técnico (dependiendo podría solo ajustarse o hacer inservible los resultados)

UMI (drop-seq, indrop-seq, mars-seq)

- Identificadores únicos moleculares para control técnico, identificar PCR duplicados

transcritos con polyA

No sirve para **non coding RNA**, en cambio si solo te interesa expresión regular sí (smart-seq2)

MATQ-seq!

https://www.nature.com/articles/nmeth.4145

Desafíos

- cantidad de muestra
  - cada tecnología te pide distintas cantidades… ¿nuestras muestras lo permiten?
- problemas al asegurar la separación de las células
  - funciona lo droplets de emulsión y lo del citómetro
- problemas en la transcripción reversa
  - no todas las moléculas lo logran (10 a 20 por ciento)
- ciertas regiones se amplifican más del DNA
  - está más accesible o oligos se pegan mejor (problema más técnico que biológico)

Desafíos

- costo / presupuesto
  - obtener transcritos completos cuesta más que solo los 3’ o 5’
- gene dropouts
  - genes que son observados con expresión moderada pero no son detectados en otra
- ruido en datos
  - cobertura de transcritos sesgada
  - baja eficiencia de captura
  - cobertura de secuenciación
- batch effect
  - diferencias producidas en la preparación de librerías, secuenciación (día, lugar, persona encargada, reagents,etc)

Single-cell combinatorial indexing RNA (sciRNA)

No requieres aislar físicamente todas las células, una combinatoria de barcodings lo harán por ti!!

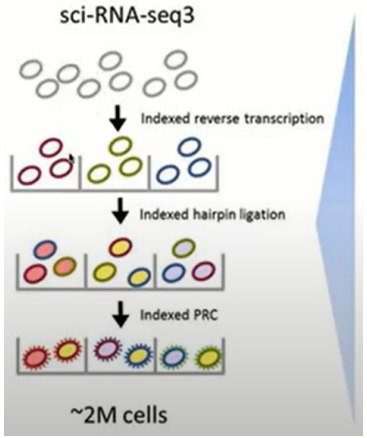
consta de varios ciclos en los que:

- N células entran a un pozo donde se les añadirá un barcode
- se vuelven a juntar todas las células y volver a cambiar de pozo
- se inserta un barcode distinto

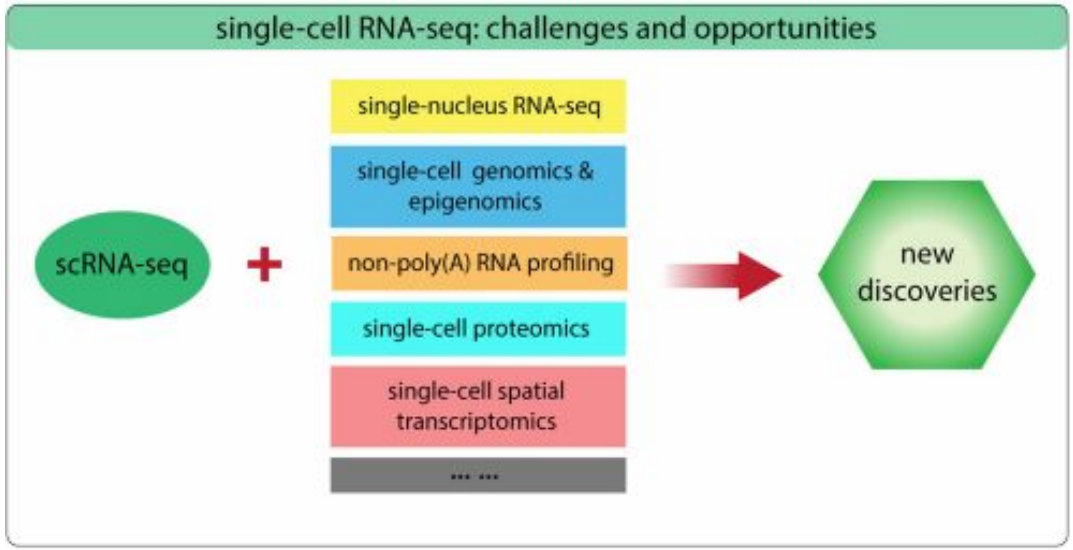
Terminamos con células únicas por combinación de barcodes!!

entre más barcodes mayor número de células a procesar

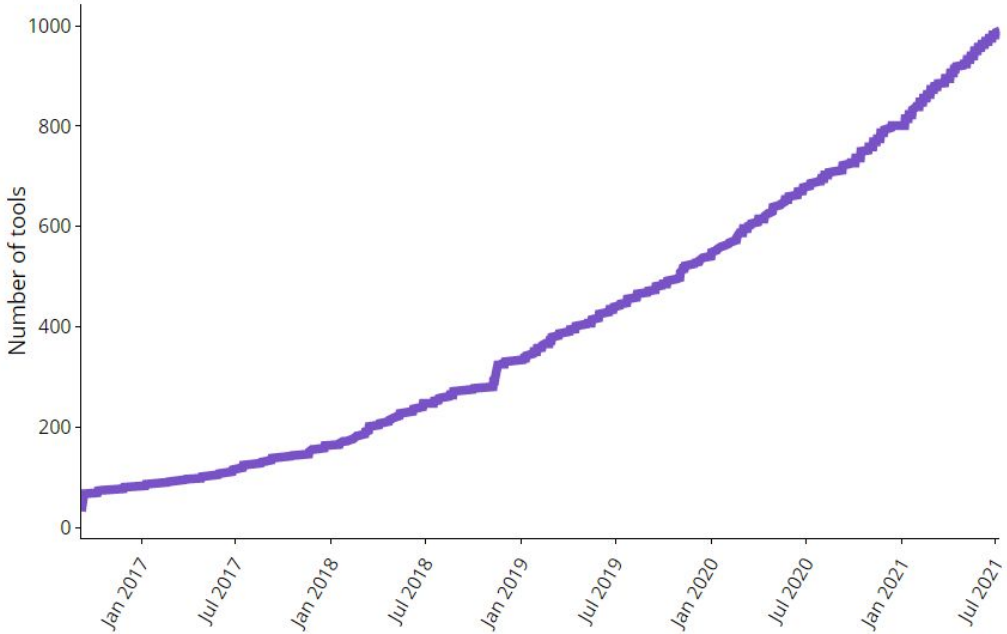
Puedes tener múltiples muestras (dif. poblaciones celulares, tejidos, individuos, réplicas) en el mismo experimento (separando los subsets en la primera ronda del indexing)



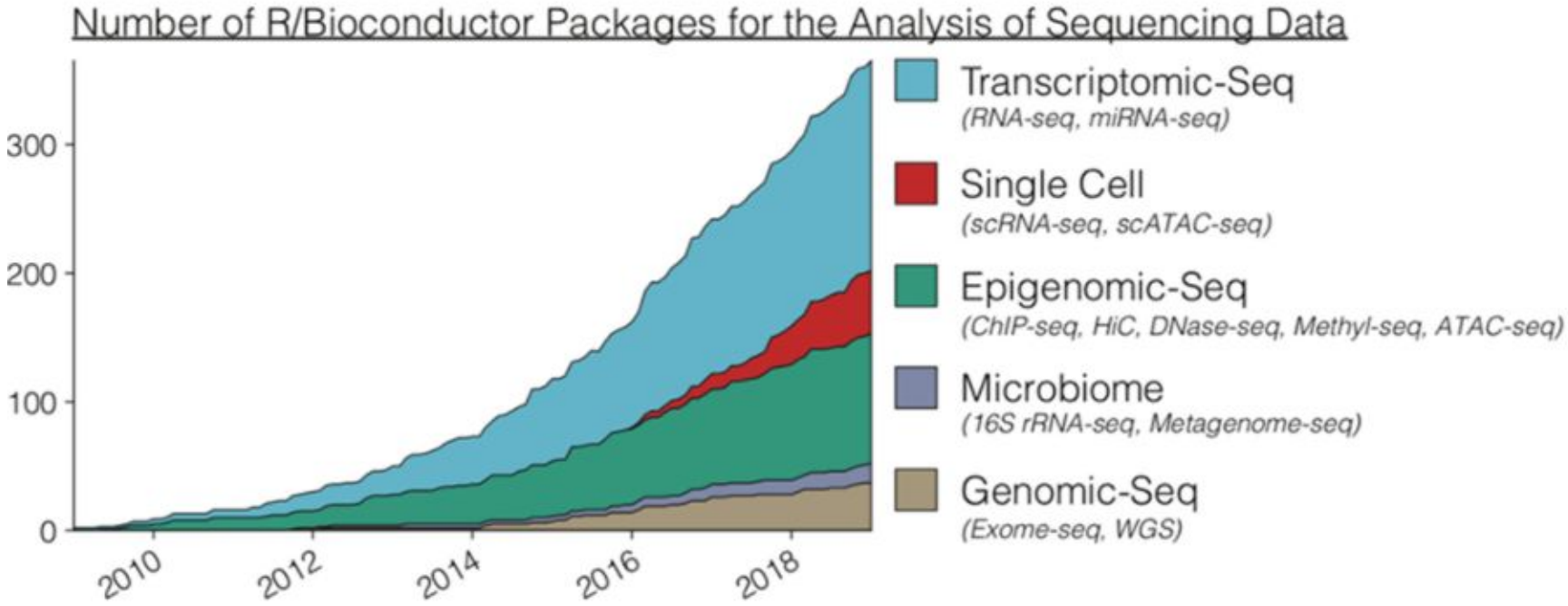
artículo de interés
https://www.rna-seqblog.com/comprehensive-single-cell-transcriptional-profiling-of-a-multicellular-organism/



https://wires.onlinelibrary.wiley.com/doi/full/10.1002/wdev.396



https://www.scrna-tools.org/analysis



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7358058

do not forget

- Include positive and negative control cells/samples
- Include multiple biological replicates
- Experimental groups should not be confounded with batch
- **Individual cells are not replicates**

what about bulk RNA-seq?

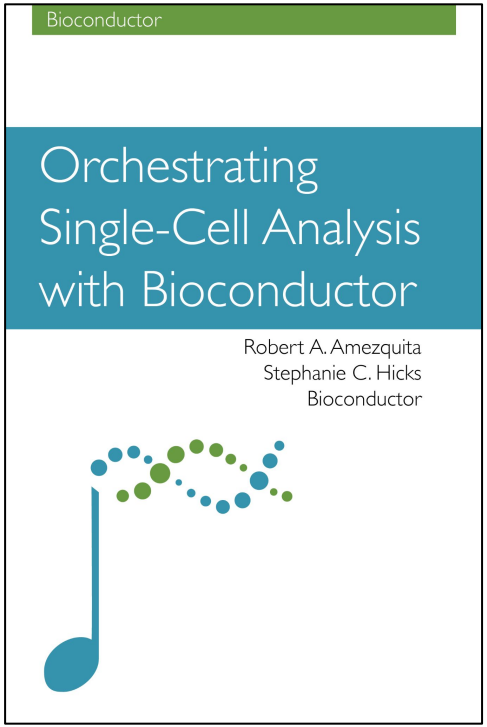
diapositiva de Peter Hickey

OSCA

Esfuerzo colaborativo de comunidad de Bioconductor

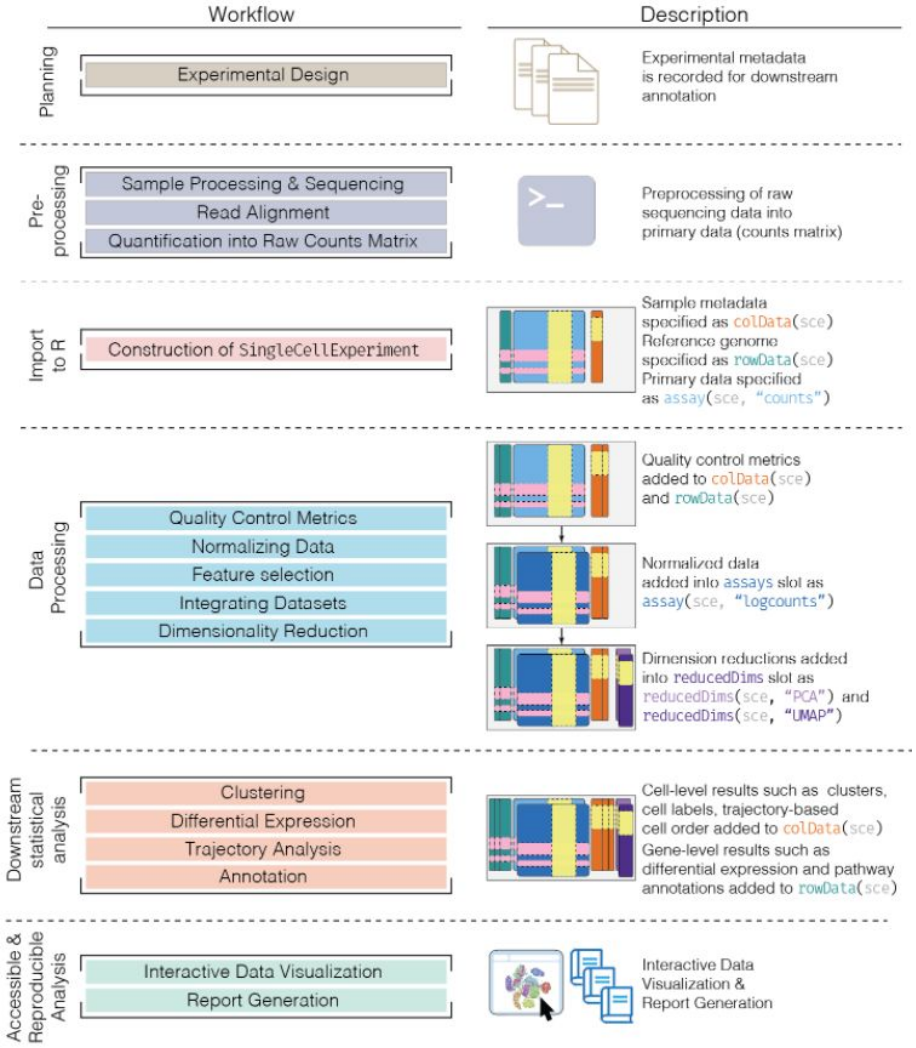
Pueden checar el libro aquí!

https://bioconductor.org/books/release/OSCA/

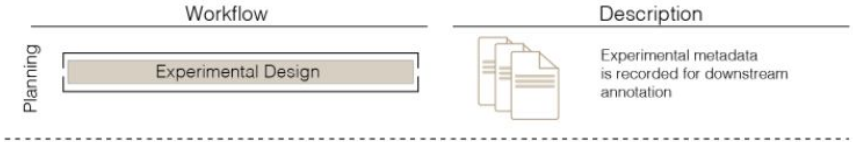


todo lo que se va a ver aquí son templates:

- lo que significa que **no siempre será igual**
    - necesitas explorar y conocer tus datos para entender mejor qué tipo de análisis conviene o son útiles para tu proyecto

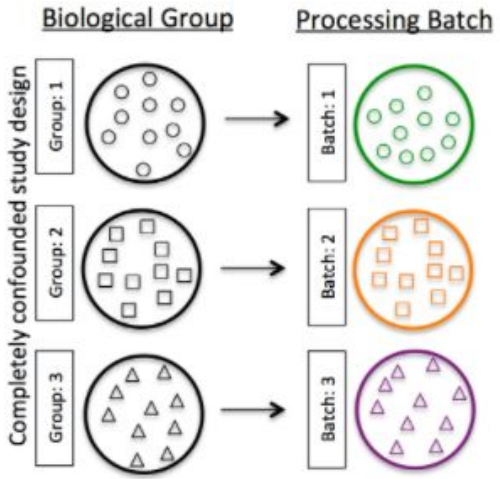

Workflow
Description
Planning
Experimental Design
Experimental metadata is recorded for downstream annotation
Pre-processing
Sample Processing & Sequencing
Read Alignment
Quantification into Raw Counts Matrix
>_
Preprocessing of raw sequencing data into primary data (counts matrix)
Import to R
Construction of SingleCellExperiment
Sample metadata specified as colData(sce)
Reference genome specified as rowData(sce)
Primary data specified as assay(sce, "counts")
Data Processing
Quality Control Metrics
Normalizing Data
Feature selection
Integrating Datasets
Dimensionality Reduction
Quality control metrics added to colData(sce) and rowData(sce)
Normalized data added into assays slot as assay(sce, "logcounts")
Dimension reductions added into reducedDims slot as reducedDims(sce, "PCA") and reducedDims(sce, "UMAP")
Downstream statistical analysis
Clustering
Differential Expression
Trajectory Analysis
Annotation
Cell-level results such as clusters, cell labels, trajectory-based cell order added to colData(sce)
Gene-level results such as differential expression and pathway annotations added to rowData(sce)
Accessible & Reproducible Analysis
Interactive Data Visualization
Report Generation
Interactive Data Visualization & Report Generation

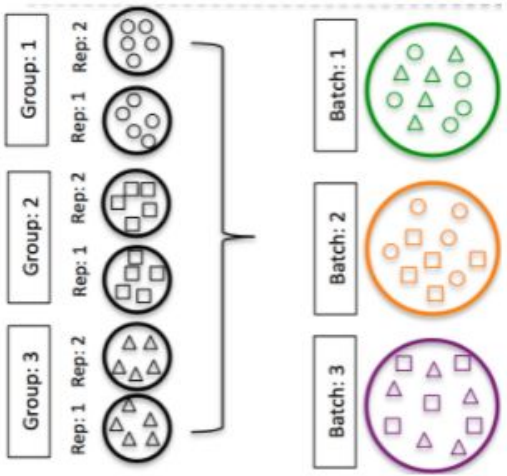
Diseño experimental



- **¿qué tecnología emplear?**
  - Droplet-based: 10X Genomics, inDrop, Drop-seq
  - Plate-based with unique molecular identifiers (UMIs): CEL-seq, MARS-seq
  - Plate-based with reads: Smart-seq2
  - Other: sci-RNA-seq, Seq-Well

- **Controles**!!!!!!!
  - **necesarios** + y - (de cell y muestra)
  - réplicas biológicas
  - réplicas técnicas
  - contemplar -y tratar de evitar- posibles efectos de batches (diferencias técnicas)
    - dividimos las muestras de tal manera que en cada corrida vayan de distintos grupos (nunca al revés!!)
  - células individuales no pueden ser tratadas como réplicas
  - **es mejor gastar un poco más que obtener información poco útil**





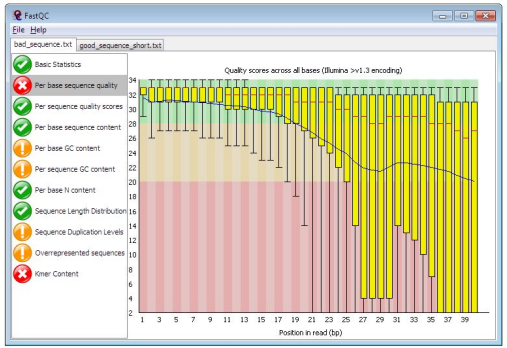
https://www.biorxiv.org/content/10.1101/025528v1

Pre procesamiento

(Parte antes de R, se hace con plataformas ya estandarizadas)

Fastqc
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

- calidad de reads, contenido de GC, oligos, bases N, etc.



Alineamiento a transcriptoma de referencia

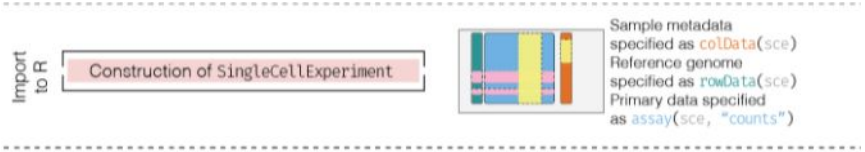
Generación matriz de cuentas de expresión

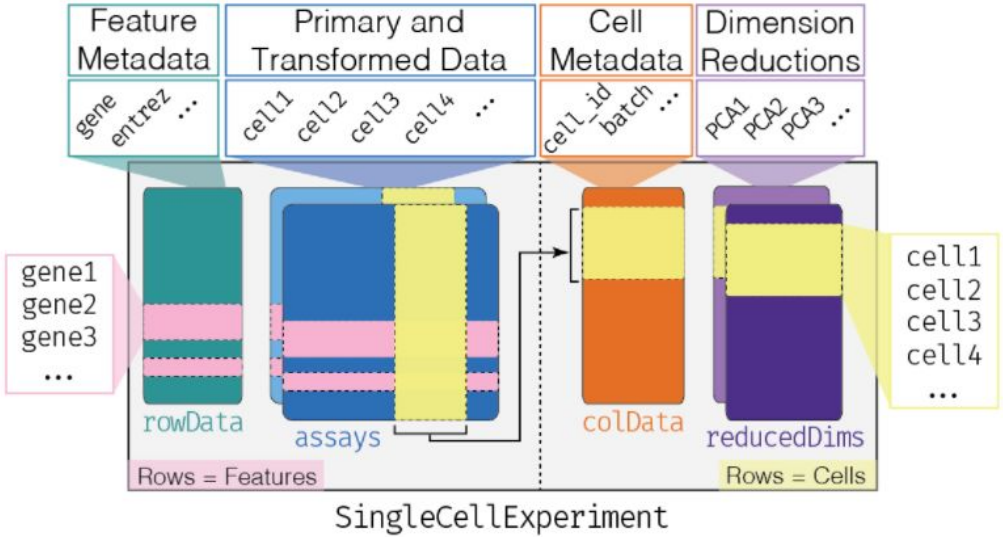
cuantificación por célula y genes

https://www.nature.com/articles/s12276-018-0071-8/figures/2

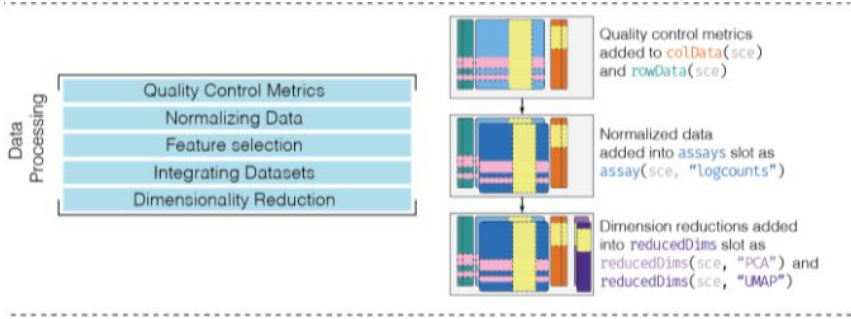
Construcción SCE

- Datos primarios
  - count matrix
  - datos transformados
- Metadata
  - cell
  - feature
  - experiment
- Reducción de dimensiones
  - PCA, tSNE, etc
- Alternative experiments





Control de calidad



Eliminar datos que traigan ruido o poca info

por ejemplo células que sufrieron daños, con poca cobertura, ausencia de célula y doublets

checar:

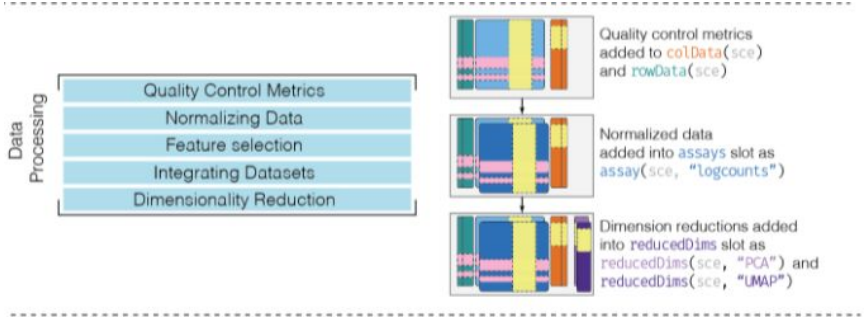
- tamaños pequeños de librerías (total de counts) (o que sepas que eso puede suceder por tu tipo celular)
- células pocos genes expresados (buenos counts pero **muy** pocos genes)
- alta proporción de lecturas mitocondria/cloroplasto
- células con alta proporción de spike-in

“¿Qué es lo peor que podría pasar?”

puedes crear interpretaciones biológicas a ruido UnU

- diferencias entre distintos grupos que NOO
- encontrar genes con alta expresión tampoco puede ser real
- **(batch effect : porque tu secuenciador funcionó diferente dos días distintos)**

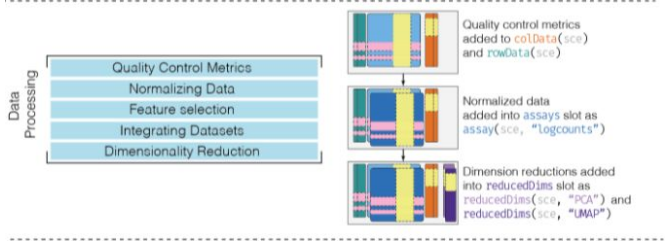
Normalización



Pueden haber problemas o diferencias en la cobertura de librerías **para poder compararlos entre sí!!!**

- por tamaño de librería (tengo X reads por célula y cuantos por genes)
- por deconvolución (conoces composición de expr. de genes o de células)
- por spike-in/UMI, identificadores, (cuantos se esperaba de eso vs realidad. Nivelar reads a lo esperado)

Selección de atributos (features)



Descartar y seleccionar elementos que enriquezcan la información, posibles genes de interés

(dependiendo de nuestra pregunta -- qué atributos queremos, cuales quitamos y por qué?)

Lo usual es buscar genes con alta variabilidad. dif enfoques a aplicar:

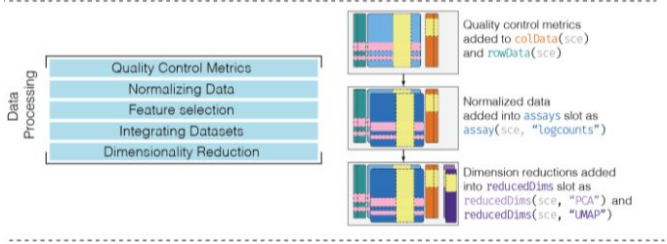
- transformaciones logarítmicas (problema con varianza estable, valores altos y bajos y los 0s)
- deviance: varianza con corrección , basada en las **UMIs**.
  - cuantifica qué tan bien un gen dado se ajusta a un modelo nulo de expresión constante en todas las células. (con UMIs puede decir que la varianza que ves no está tan alta, más bien se acota y esto puede reducir ciertas amplitudes)

O **con lo que conocemos de nuestro modelo** nos vamos guiando a los genes relevantes

También métodos basados en spike ins y dropouts

reducción si notamos cambios sistemáticos en pérdidas

Reducción de dimensiones



Crear representaciones de datos en dimensiones pequeñas pero que preservan una estructura significativa

si identificamos grupos de células nos quedarían grupos 1,2,...,12, en lugar de tener 100 cells vs x cells

Un caso puede ser genes correlacionados (proceso biológico en común, pathways)

Análisis de Componentes Principales (PCA)

checar si los grupos van bien biológicamente

pa visualizar:

t-SNE (t-distributed stochastic neighbor embedding)

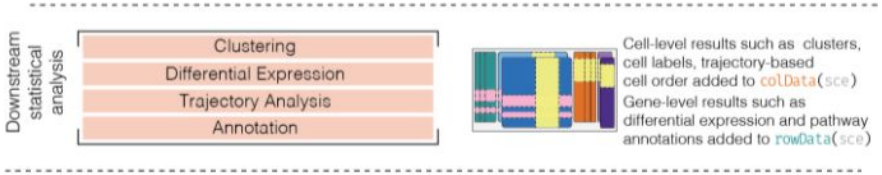
UMAP (uniform manifold approximation and projection)

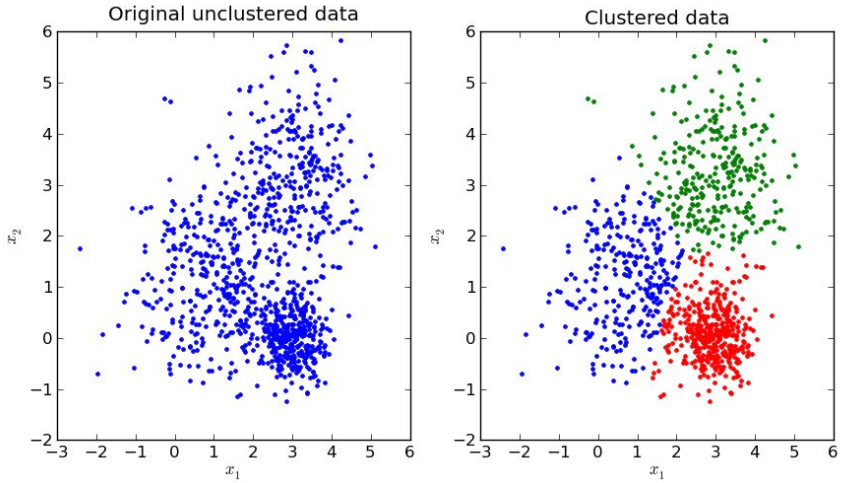
Clustering

Agrupamiento de acuerdo a perfiles similares de expresión . ¡para explorar datos!

Problema: al jugar con los parámetros o métodos podrías conseguir múltiples agrupaciones.. ¿cómo elegir?

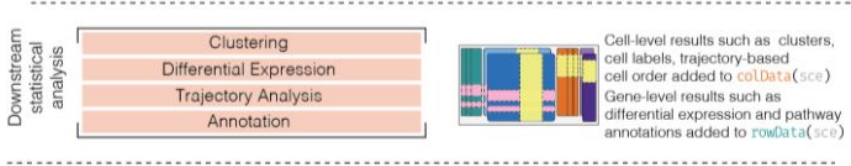
- grafos
- k-means (centroides)
- jerárquico (dendograma)





https://towardsdatascience.com/k-means-data-clustering-bce3335d2203

Análisis de expresión diferencial



Encontrar genes que se expresan diferencialmente!!

- genes de interés (siempre podemos checar lo biológico)

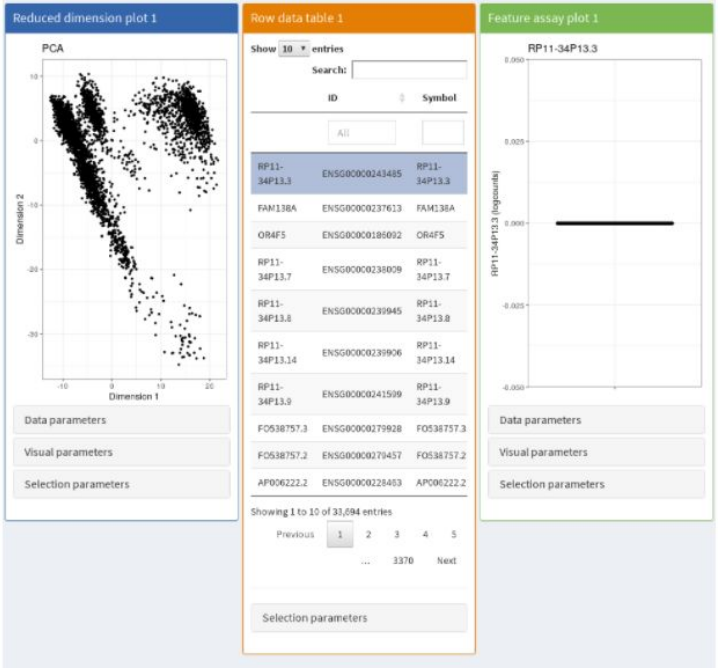
checar si encontramos genes con exp. dif. comparando entre los grupos (con las varianzas) o al comparar exp. gen. entre clusters

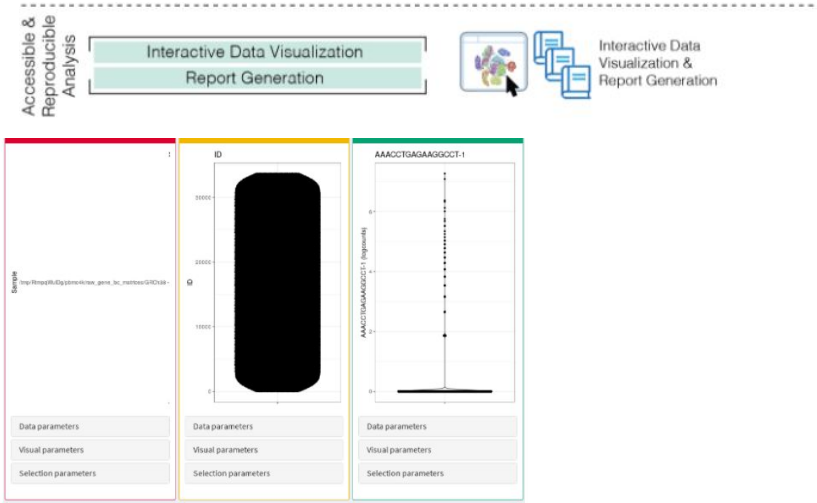
O tests no paramétricos (no asume distribución particular de valores de expresión)

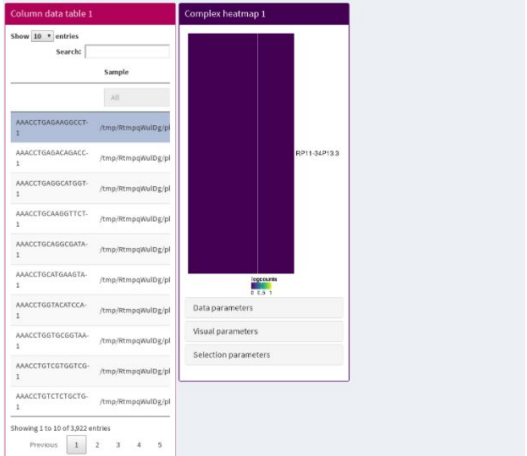
- Kolmogorov-Smirnov test (KS-test)

Exploratory data analysis

paquete **iSEE**







¡Gracias!