Single cell RNA-seq

Mónica Padilla

Alejandra Eugenia Medina Rivera, PhD.

Basado en:
Orchestrating single-cell analysis with Bioconductor (2020) Robert A. Amezquita… Stephanie C. Hicks.
Material de Peter Hickey

Overview

- ¿Qué es el scRNA-seq?
- Captura de una sola célula
- Tecnologías de scRNA-seq y desafíos
- Procesamiento general de datos de scRNA-seq
  - Preprocesamiento de datos
  - Control de Calidad (QC)
  - Normalización
  - Imputación
  - Selección de atributos (feature selection)
  - Reducción de dimensiones
  - Integración de conjuntos de datos

¿Qué es el scRNA-seq?

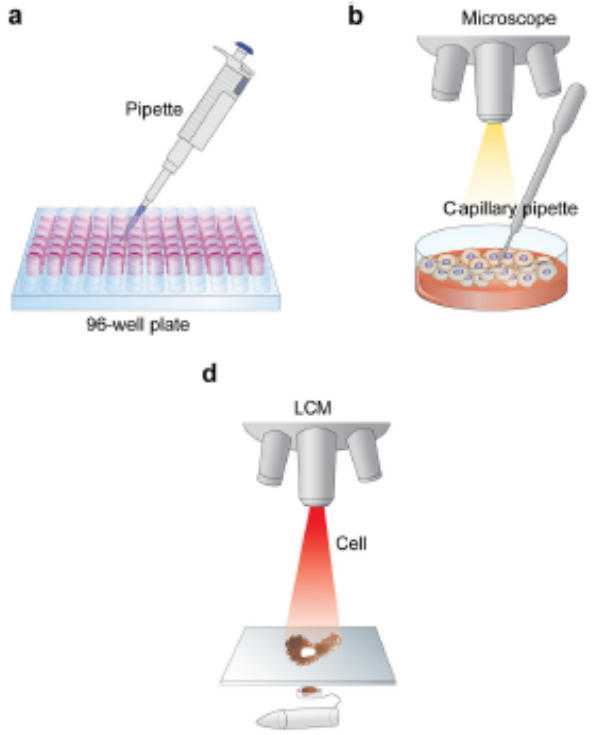
Tecnología que permite la disección de la expresión génica a la resolución de una sola célula.

Chen, G., Ning, B., & Shi, T. (2019). Single-cell RNA-seq technologies and related computational data analysis. Frontiers in genetics, 10, 317.

Captura de una sola célula

Técnicas low-throughput

- **Dilución limitante**
  - Ineficiente
- **Capilar guiado por microscopio:**
  - Para muestras en suspensión o en tejido
  - Muestras de pocas o frágiles células
  - Consume tiempo, técnicamente retador
- **Microdisección por captura láser (LCM)**
  - Aislamiento de células específicas en tejido
  - Conserva la relación especial
  - Automático



Hwang, B., Lee, J. H., & Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & molecular medicine*, *50*(8), 1-14. (editado)

Captura de una sola célula

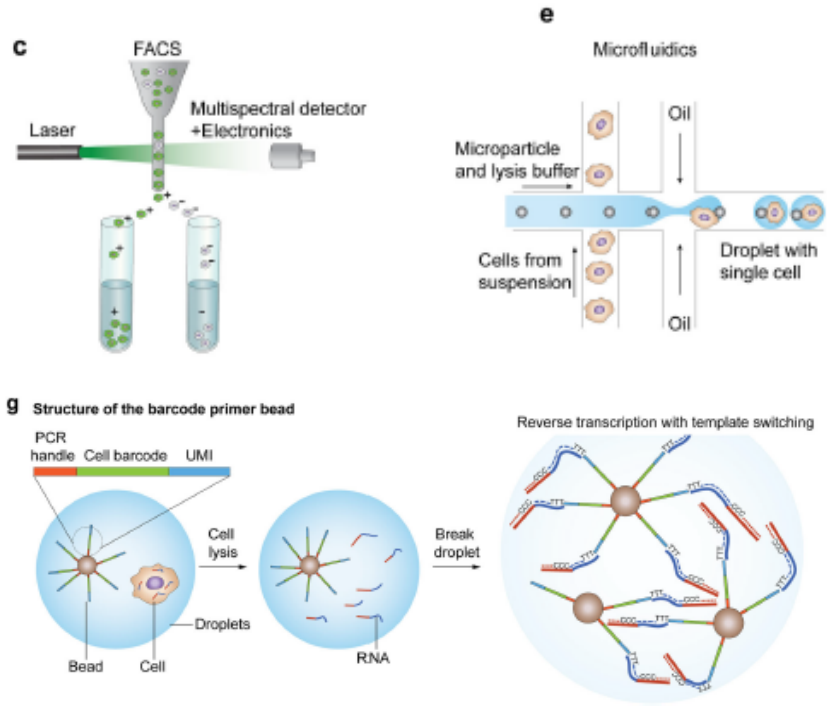
Técnicas high-throughput

Ambas técnicas toman células de una suspensión, son automáticas y son *cost-effective*

- **Flow-activated cell sorting (FACS):**
  - Se requieren >= 10,000 células y anticuerpos (muy usando en inmunología).
  - Puede dar lugar a pozos vacíos o con *doublets* (dos células).
  - Sujeto a contaminación.

- **Microfluidic technology**
  - Se requieren >= 1000 células
  - Menor riesgo de contaminación
  - Control de fluido y costo bajo de análisis

  - **Microdroplet-based microfluidics**
    - Uso de beads con *barcodes* únicos
    - Encapsulamiento en *droplets*
    - Etiqueta UMI



Hwang, B., Lee, J. H., & Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & molecular medicine*, *50*(8), 1-14. (editado)

Tecnologías y desafíos

Pasos en común:

- Aislamiento de células
- Lisis celular
- Transcripción en reversa para la primera hebra del cDNA
  - Se estima que solo del 10 al 20% de los transcritos de retrotranscriben
  - Uso de una transcriptasa reversa modificada del virus de la leucemia murina
- Síntesis de la segunda hebra
  - Uso de *poly(A) tailing* o del mecanismo *template-switching*
- Amplificación de cDNA

Tecnologías y desafíos

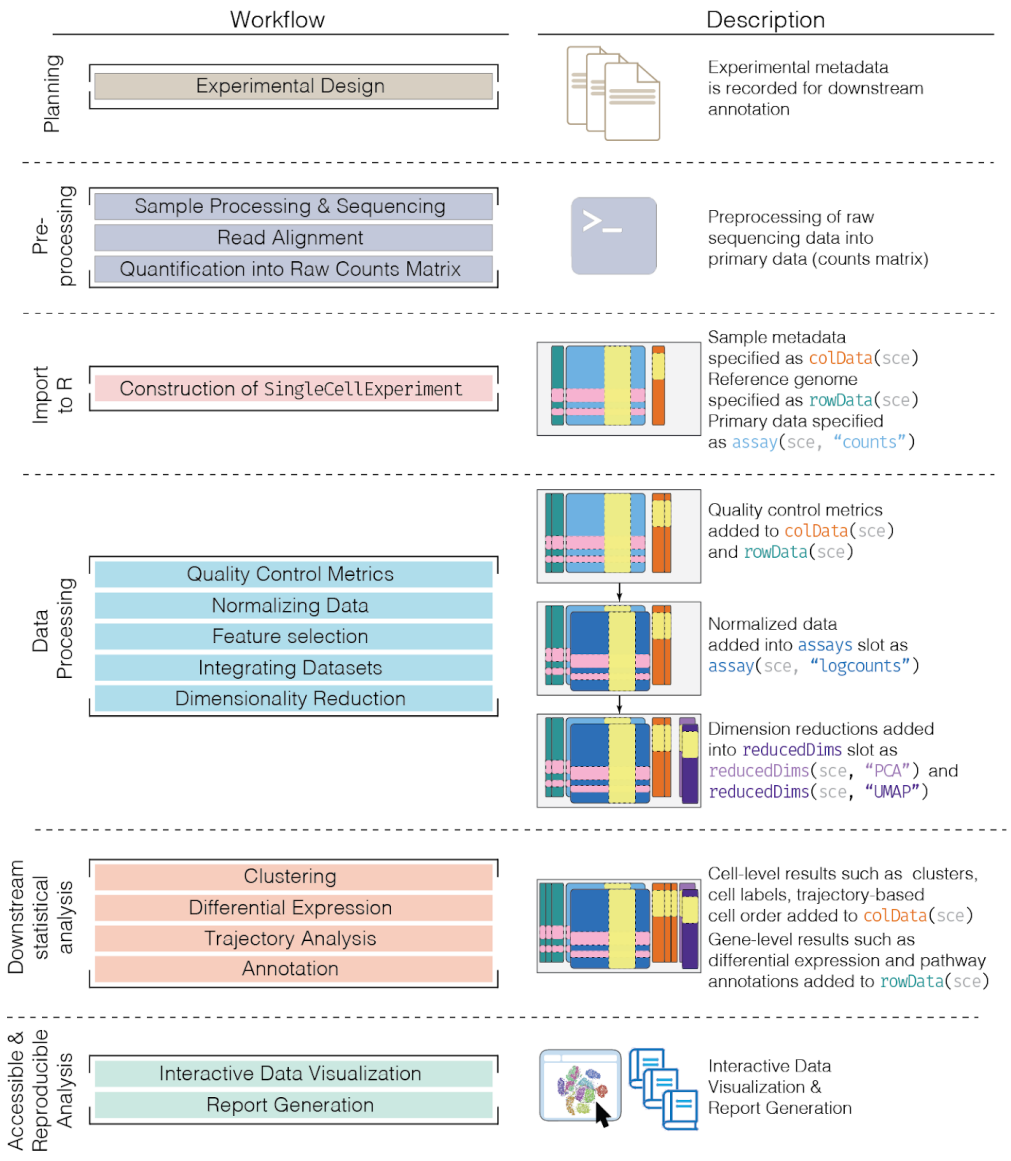
- La captura de transcritos puede ser tanto del tamaño completo o de los extremos 3’ o 5’
- Algunos métodos pueden capturar RNAs poliadenilados y no polineadenilados
- Uso de *spike-ins* y Unique molecular identifiers (identificador molecular único, UMI) para estimar la variación técnnica entre células

- La cobertura de transcritos sesgada, la baja eficiencia de captura y la cobertura de secuenciación resultan en datos con más ruido que los de *bulk* RNA-seq.

TABLE 1 | Summary of widely used scRNA-seq technologies.

Methods	Transcript coverage	UMI possibility	Strand specific	References
Tang method	Nearly full-length	No	No	Tang et al., 2009
Quartz-Seq	Full-length	No	No	Sasagawa et al., 2013
SUPeR-seq	Full-length	No	No	Fan X. et al., 2015
Smart-seq	Full-length	No	No	Ramskold et al., 2012
Smart-seq2	Full-length	No	No	Picelli et al., 2013
MATQ-seq	Full-length	Yes	Yes	Sheng et al., 2017
STRT-seq and STRT/C1	5′-only	Yes	Yes	Islam et al., 2011, 2012
CEL-seq	3′-only	Yes	Yes	Hashimshony et al., 2012
CEL-seq2	3′-only	Yes	Yes	Hashimshony et al., 2016
MARS-seq	3′-only	Yes	Yes	Jaitin et al., 2014
CytoSeq	3′-only	Yes	Yes	Fan H.C. et al., 2015
Drop-seq	3′-only	Yes	Yes	Macosko et al., 2015
InDrop	3′-only	Yes	Yes	Klein et al., 2015
Chromium	3′-only	Yes	Yes	Zheng et al., 2017
SPLiT-seq	3′-only	Yes	Yes	Rosenberg et al., 2018
sci-RNA-seq	3′-only	Yes	Yes	Cao et al., 2017
Seq-Well	3′-only	Yes	Yes	Gierahn et al., 2017
DroNC-seq	3′-only	Yes	Yes	Habib et al., 2017
Quartz-Seq2	3′-only	Yes	Yes	Sasagawa et al., 2018

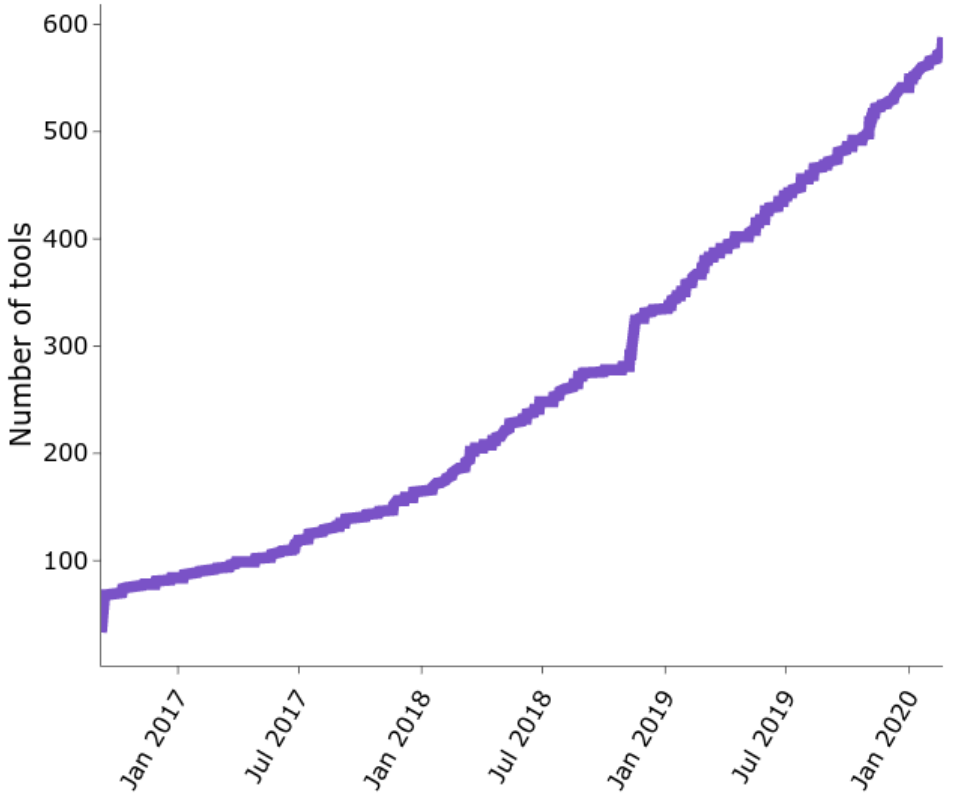
Chen, G., Ning, B., & Shi, T. (2019). Single-cell RNA-seq technologies and related computational data analysis. Frontiers in genetics, 10, 317.


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

Procesamiento de datos

1. Preprocesamiento
2. Control de Calidad
3. Normalización
4. Imputación
5. Selección de atributos
6. Reducción de dimensiones
7. Integración de conjuntos de datos

Number of scRNA-seq tools over time

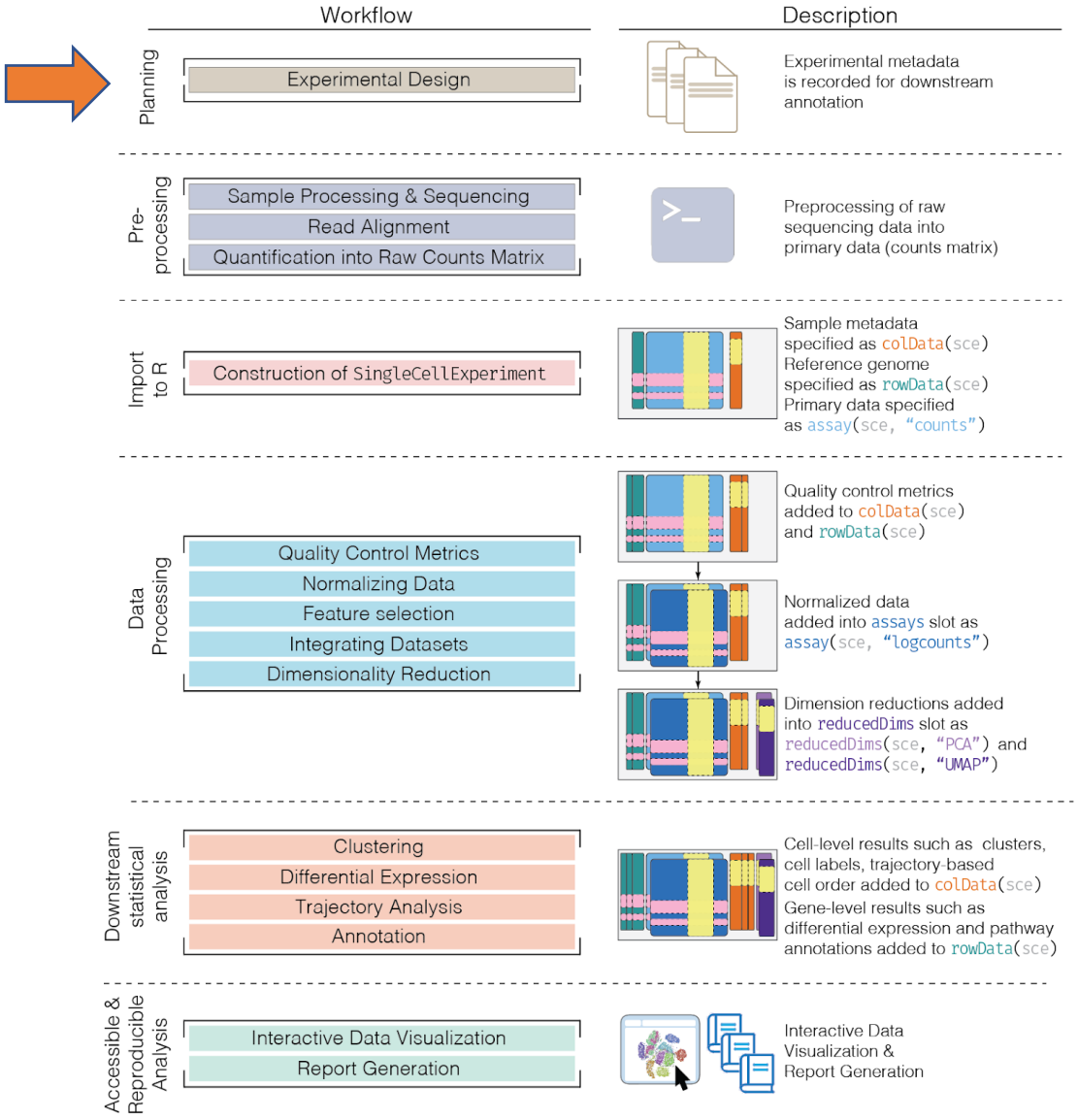


Modificado de Peter Hickey

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

No hay recetas!!! Hay flujos de trabajo !

Modificado de [Peter Hickey]

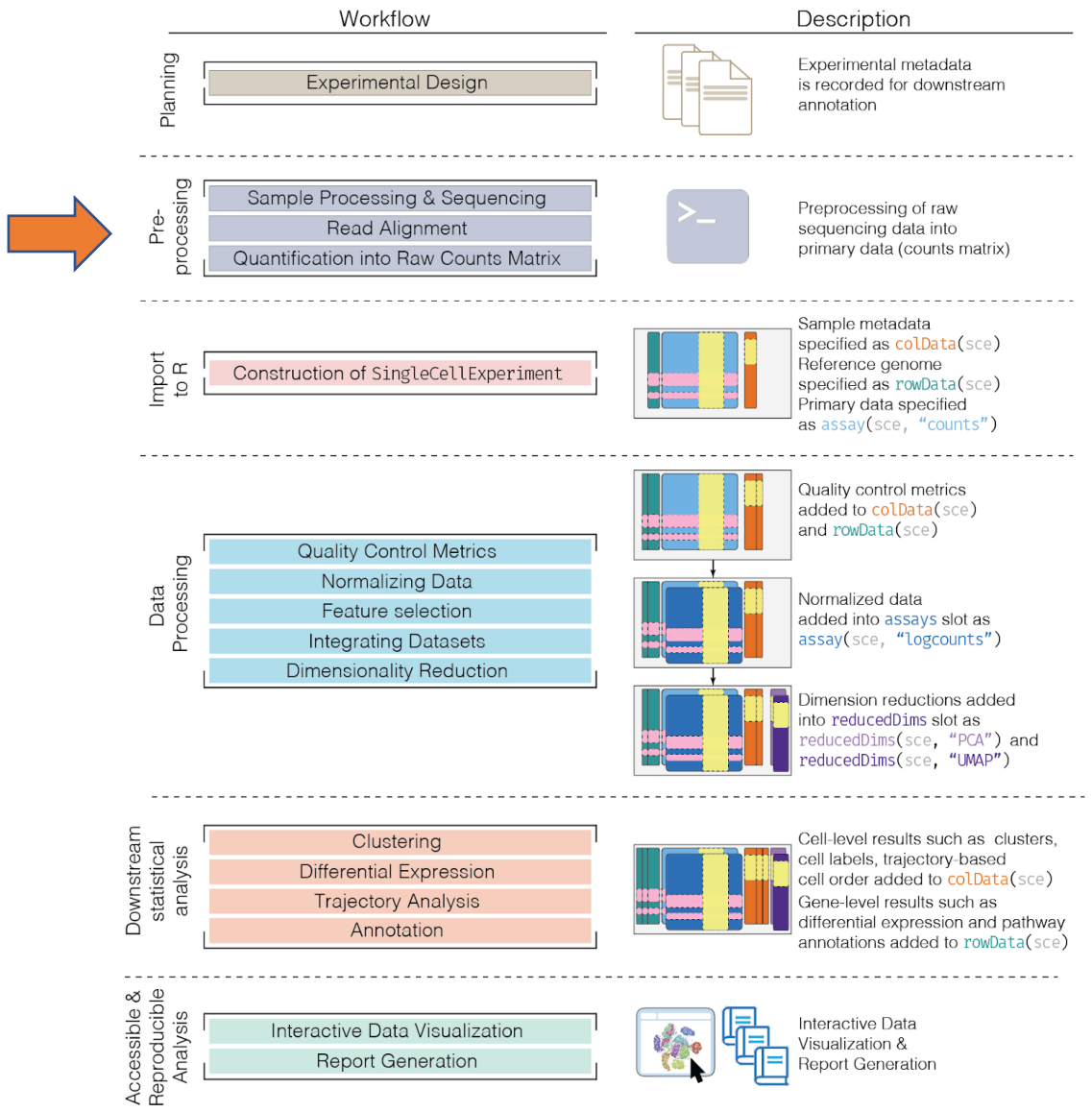


Modificado de [Peter Hickey]

Controles! Controles! Controles!

- Necesitamos controles positivos y negativos de las células y de las muestras.
- Necesitamos varias replicas bilógicas.
- Los grupos experimentales no son batches. Los batches son técnicos.
- Las células individuales no pueden ser tratadas como replicas.
- Aquí los ahorros nos pueden salir muy caros después.

Modificado de Peter Hickey



Modificado de [Peter Hickey]

Procesamiento de datos: Preprocesamiento

- QC de secuenciación (FASTQC) https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
- Alineamiento a transcriptoma de referencia
- Cuantificación por célula y por gen para la generación de la matriz de cuentas de expresión

Procesamiento de datos: Preprocesamiento

- Herramientas para hacer la matriz de cuentas:
  - CellRanger para datos 10X
  - scPipe para CEL-Seq2 y otros tipos de datos
  - Kallisto y alevin para datos de droplets

Cómo se ven los datos?

Cuentas de Genes

	Cell 1	**Cell 2**	**...**	**Cell N**
Gene 1	0	1	...	0
Gene 2	1	3	...	0
...	...	...	...	...
Gene M	2	2		4

Modificado de Peter Hickey

Cómo se ven los datos?

Información de las células

	Barcode	**Donor**	**...**	**Treatment**
Cell 1	ACTGTA	D1	...	Drug
Cell 2	TGCATA	D1	...	Control
...	...	...	...	...
Cell N	CCTATA	D6		Drug

Modificado de Peter Hickey

Cómo se ven los datos?

Información de los Genes

	ID	**Symbol**	**...**	**Chromosome**
Gene 1	ENSG00000155816	FMN2	...	1
Gene 2	ENSG00000229807	XIST	...	X
...	...	...	...	...
Gene M	ENSG00000139618	BRCA2		13

Modificado de Peter Hickey

Cómo queremos que se vean los datos normalizados?

Normalización log de expression de genes

	Cell 1	**Cell 2**	**...**	**Cell N**
Gene 1	0	0.6	...	0
Gene 2	0.3	0.8	...	0
...	...	...	...	...
Gene M	0.35	0.67		2.1

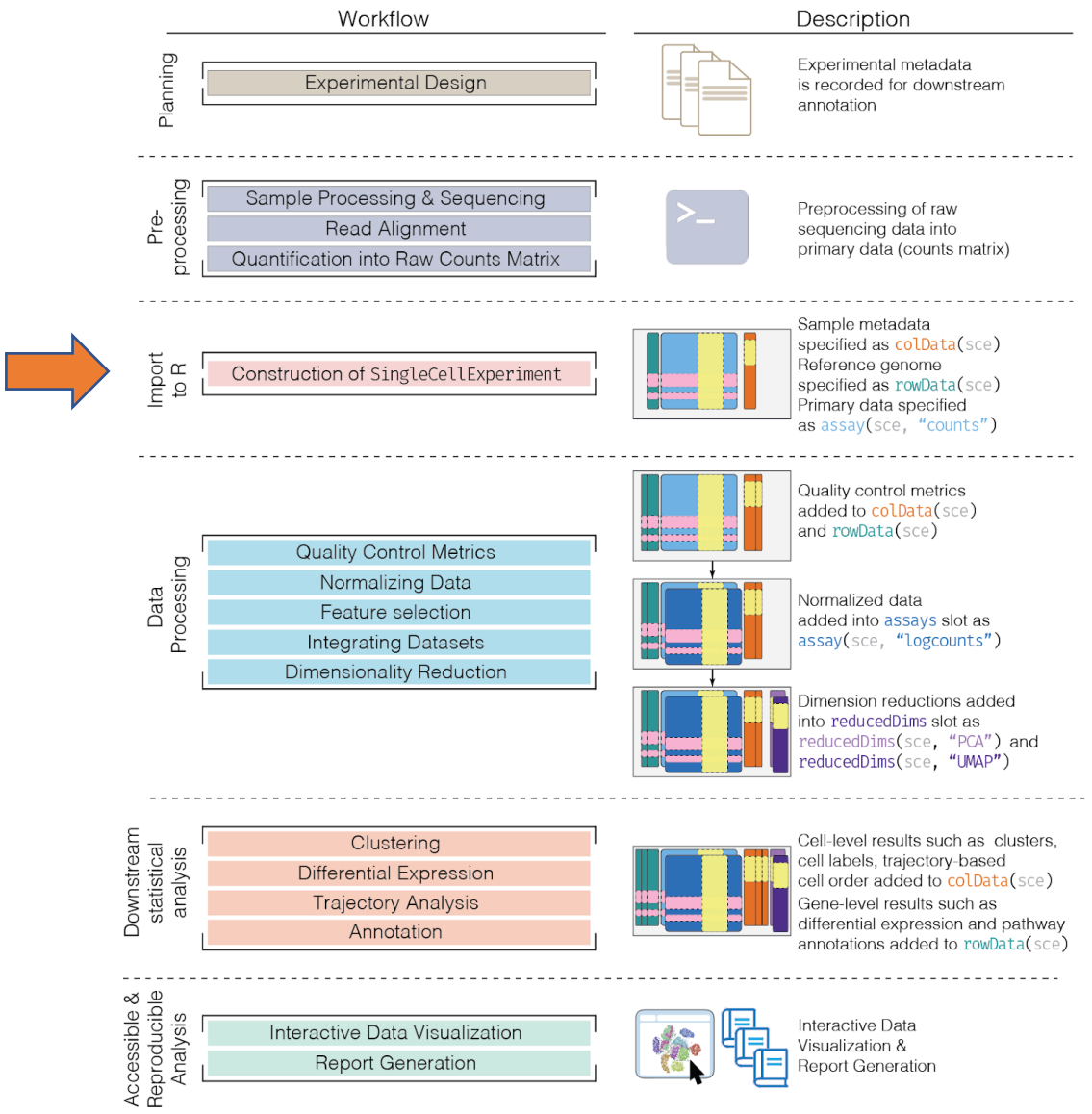
Modificado de Peter Hickey

Cómo se verán nuestros datos cuando hagamos reducción de dimensiones?

	PCA 1	**PCA 2**	**...**	**PCA K**
Cell 1	0.93	1.28	...	0.03
Cell 2	0.32	1.22	...	0.09
...	...	...	...	...
Cell N	-0.66	1.00		0.15

	t-SNE 1	**t-SNE 2**
Cell 1	1.24	8.93
Cell 2	-0.33	7.85
...	...	...
Cell N	0.46	3.41

Modificado de Peter Hickey



Modificado de [Peter Hickey]

Hay que cargar nuestro experimento en R

Necesitamos objetos que nos permitan organizar toda la información

Información de genes

	Cell 1	**Cell 2**	**...**	**Cell N**
Gene 1	0	1	...	0
Gene 2	1	3	...	0
...	...	...	...	...
Gene M	2	2		4

Información de células

	Barcode	**Donor**	**...**	**Treatment**
Cell 1	ACTGTA	D1	...	Drug
Cell 2	TGCATA	D1	...	Control
...	...	...	...	...
Cell N	CCTATA	D6		Drug

Información de los genes

	ID	**Symbol**	**...**	**Chromosome**
Gene 1	ENSG00000155816	FMN2	...	1
Gene 2	ENSG00000229807	XIST	...	X
...	...	...	...	...
Gene M	ENSG00000139618	BRCA2		13

Modificado de [Peter Hickey]

Hay que cargar nuestro experimento en R

Necesitamos objetos que nos permitan organizar toda la información

Información de genes

	Cell 1	**Cell 2**	**...**	**Cell N**
Gene 1	0	1	...	0
Gene 2	1	3	...	0
...	...	...	...	...
Gene M	2	2		4

Información de células

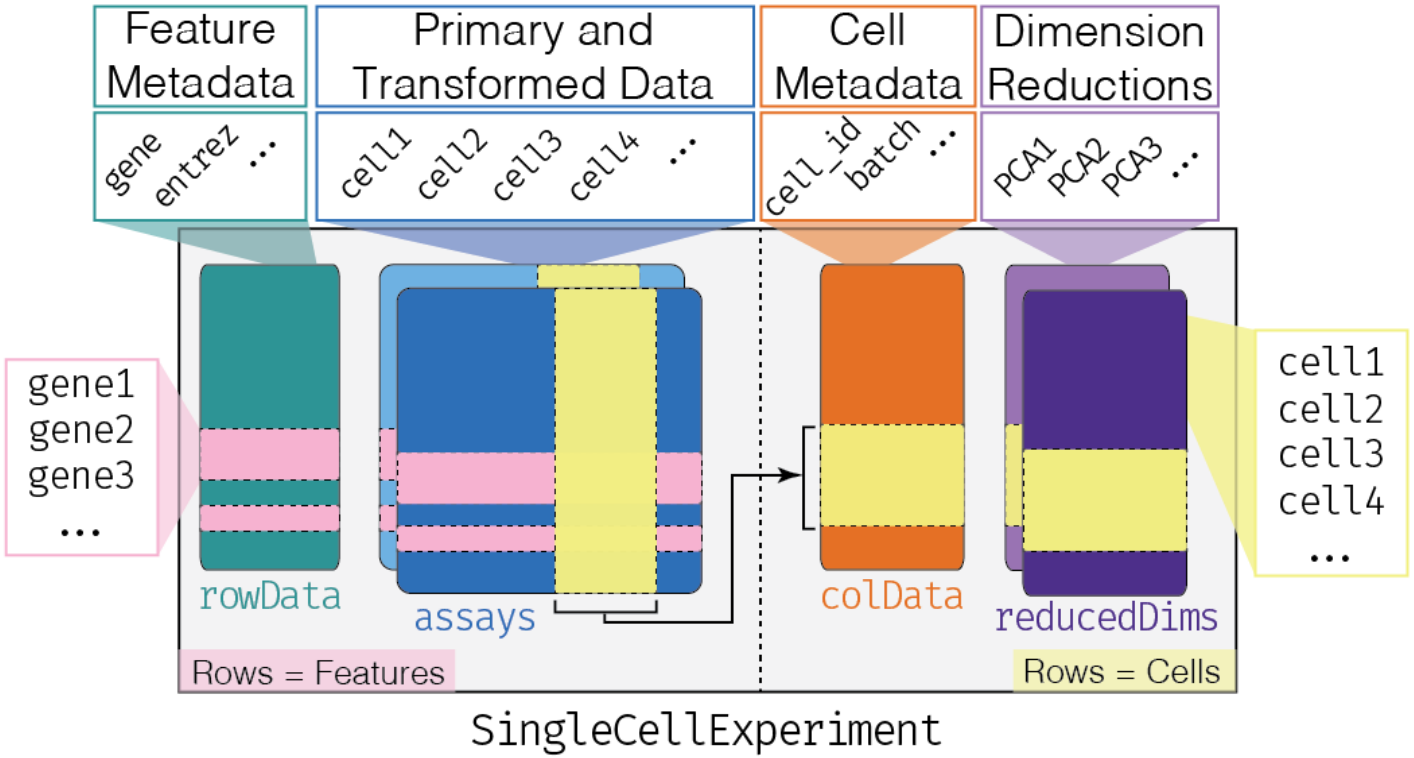
	Barcode	**Donor**	**...**	**Treatment**
Cell 1	ACTGTA	D1	...	Drug
Cell 2	TGCATA	D1	...	Control
...	...	...	...	...
Cell N	CCTATA	D6		Drug

Información de los genes

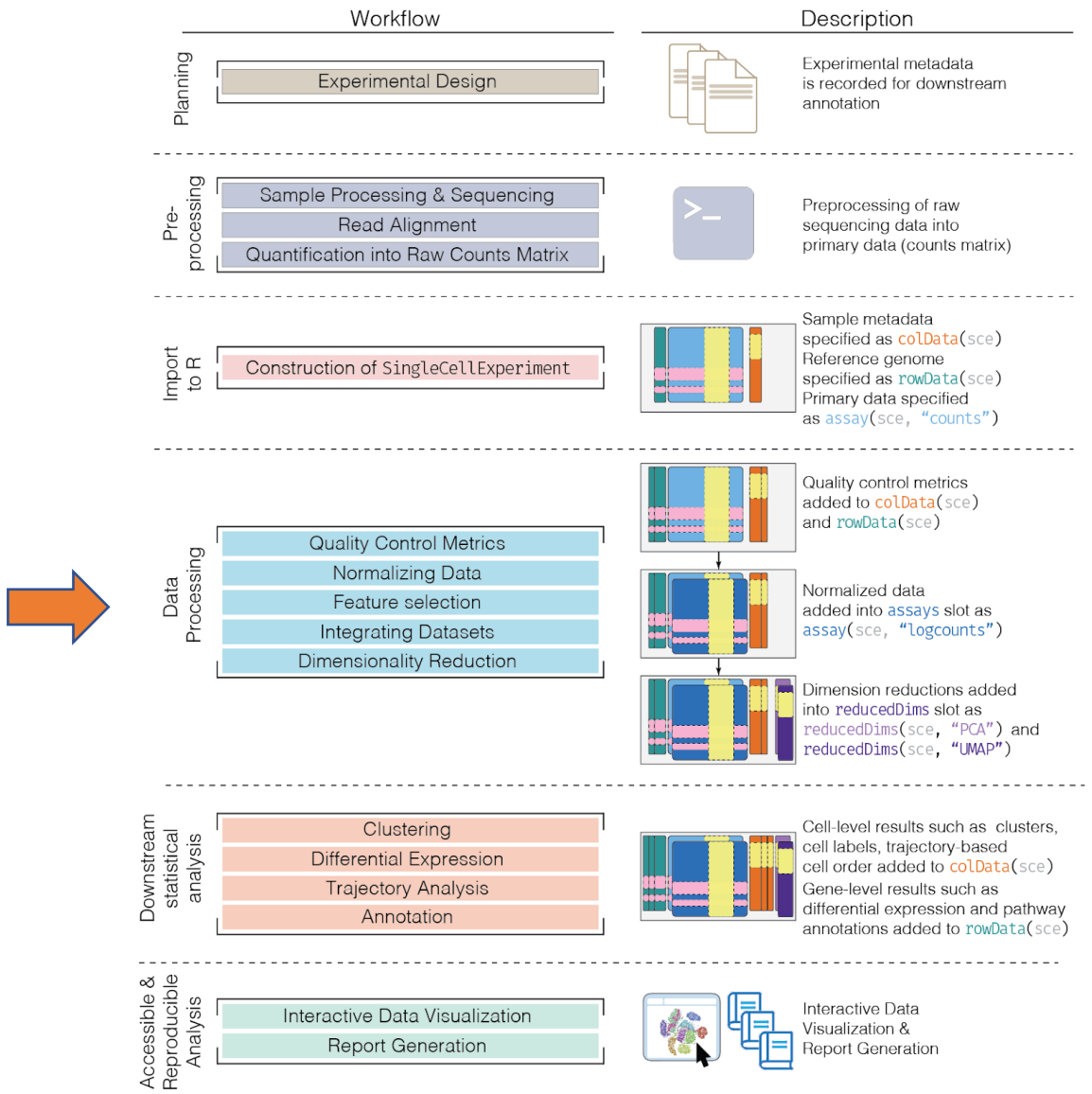
	ID	**Symbol**	**...**	**Chromosome**
Gene 1	ENSG00000155816	FMN2	...	1
Gene 2	ENSG00000229807	XIST	...	X
...	...	...	...	...
Gene M	ENSG00000139618	BRCA2		13

Modificado de [Peter Hickey]

Hay que cargar nuestro experimento en R

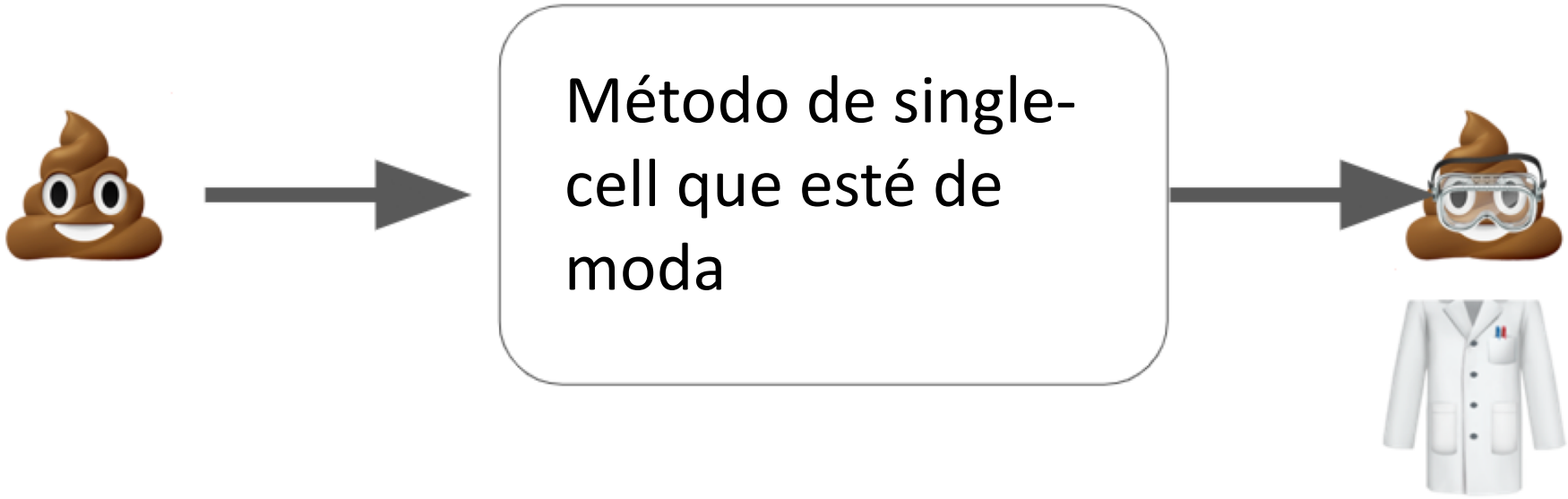


Modificado de [Peter Hickey]



Modificado de [Peter Hickey]

Procesamiento de datos: QC



Modificado de [Peter Hickey]

Procesamiento de datos: QC

Queremos eliminar

- células que sufrieron daño durante la disociación
- Células con poca cobertura (el RNA pudo haberse perdido en algún paso anterior)
- La ausencia de células y los *doublets*

Modificado de Peter Hickey

Procesamiento de datos: QC

Queremos eliminar

- células que sufrieron daño durante la disociación
- Células con poca cobertura (el RNA pudo haberse perdido en algún paso anterior)
- La ausencia de células y los *doublets*

Esto se suele manifestar como

- Células con un tamaño pequeño de librería (cuentas totales)
- Células con pocos genes expresados
- Una alta proporción de lecturas mitocondriales
- Células con alta proporción de *spike-ins*

Modificado de Peter Hickey

No hay recetas!! Cada grupo de datos será distinto

Los umbrales de calidad de un experiment no necesariamente sirven en otro.

Modificado de Peter Hickey

Procesamiento de datos: QC

Qué pasa si no hacemos bien nuestro QC?

- Vamos a ver grupos de ”datos” que no serán nada biológico, pero podemos pensar que si.
- Veremos poblaciones heterogéneas que en verdad solo son ruido.
- Algunos genes se verán super expresados pero no significa nada.

Modificado de Peter Hickey

Procesamiento de datos: Normalización

Se observan diferencias de cobertura entre librerías. Esto parte de diferencias en la captura de cDNA o de la eficiencia de la amplificación. También hay un efecto por el número de distintos tipos celulares.

- Normalización por tamaño de librería
- Normalización por deconvolución
- Normalización refinada por *spike-ins/UMIs*, cuando están presentes

Procesamiento de datos: Imputación

Necesario para vencer el reto de escasez de datos (los ceros).

Por qué hay tantos ceros?

	Cell 1	**Cell 2**	**...**	**Cell N**
Gene 1	0	0.6	...	0
Gene 2	0.3	0.8	...	0
...	...	...	...	...
Gene M	0.35	0.67		2.1

Procesamiento de datos: Imputación

- Se han elaborado modelos que toman en cuentan los ceros en cuentas (***zero-inflated models***)
  - El nivel de *zero-inflated* depende del ensayo o el protocolo
  - Un buen modelo puede ser **ensayo dependiente**, lo que funcionó para un proyecto no necesariamente funcionen en otros.
- Los modelos de imputación han demostrado **generar falsos positivos** y **disminuir la reproducibilidad** de marcadores célula específicos
- SAVER es un modelo basado en bayesiana diseñados para datos con UMIs y MAGIC construye una gráfica basada en afinidad de markov.

Procesamiento de datos: Selección de atributos

La meta es identificar genes con información biológica útil y quitar genes con ruido además de reducir el tamaño del *dataset*.

Por qué?

Por que queremos ver diferencias entre genes, células, si hay tipos celulares interesantes, etc.

Procesamiento de datos: Selección de atributos

La meta es identificar genes con información biológica útil y quitar genes con ruido además de reducir el tamaño del *dataset*.

- Hay atributos (features) que darán información bilógica (+)
- Otros atributos serán solo ruido (-)

Modificado de Peter Hickey

Procesamiento de datos: Selección de atributos

La meta es identificar genes con información biológica útil y quitar genes con ruido además de reducir el tamaño del *dataset*.

- Un enfoque sencillo es seleccionar los genes más variables acorde a su expresión.
  - Las transformaciones logarítmicas no logran una estabilización de la variancia
  - Uso de métrica ***deviance***, la cual cuantifica qué tan bien un gen dado se ajusta a un modelo nulo de expresión constante en todas las células. Se basa en UMIs.
- Podemos usar lo que sabemos de nuestro sistema y enfocarnos en genes relevantes.
- También existen métodos basados en *spike-ins* y en *dropouts*

Procesamiento de datos: Reducción de dimensiones

Objetivo: identificar grupos de células parecidas.

Problema: Nuestros datos tienen demasiados atributos como para hacerlo de forma eficiente (muchas células, muchos genes, más nuestro diseño experimental).

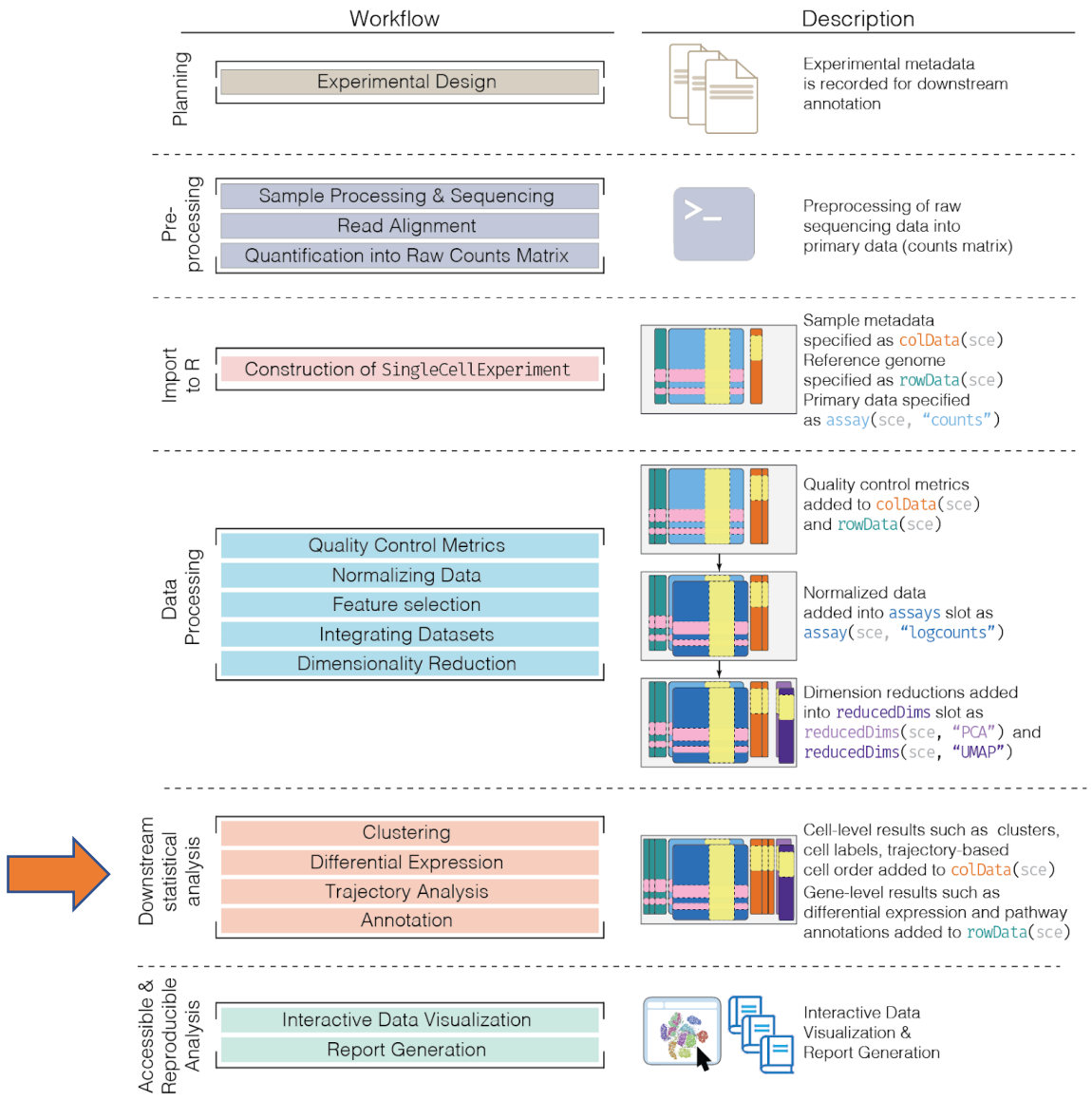
Procesamiento de datos: Reducción de dimensiones

- Genes distintos pueden estar correlacionados si el mismo proceso biológico los afecta.
- La meta es crear representaciones de los datos en dimensiones pequeñas pero que preservan una estructura significativa.
- Análisis de Componentes Principales (PCA)
- Para propósitos de visualización también existen:
  - T-distributed stochastic neighbor embedding(t-SNE)
  - Uniform Manifold Approximation and Projection (UMAP)

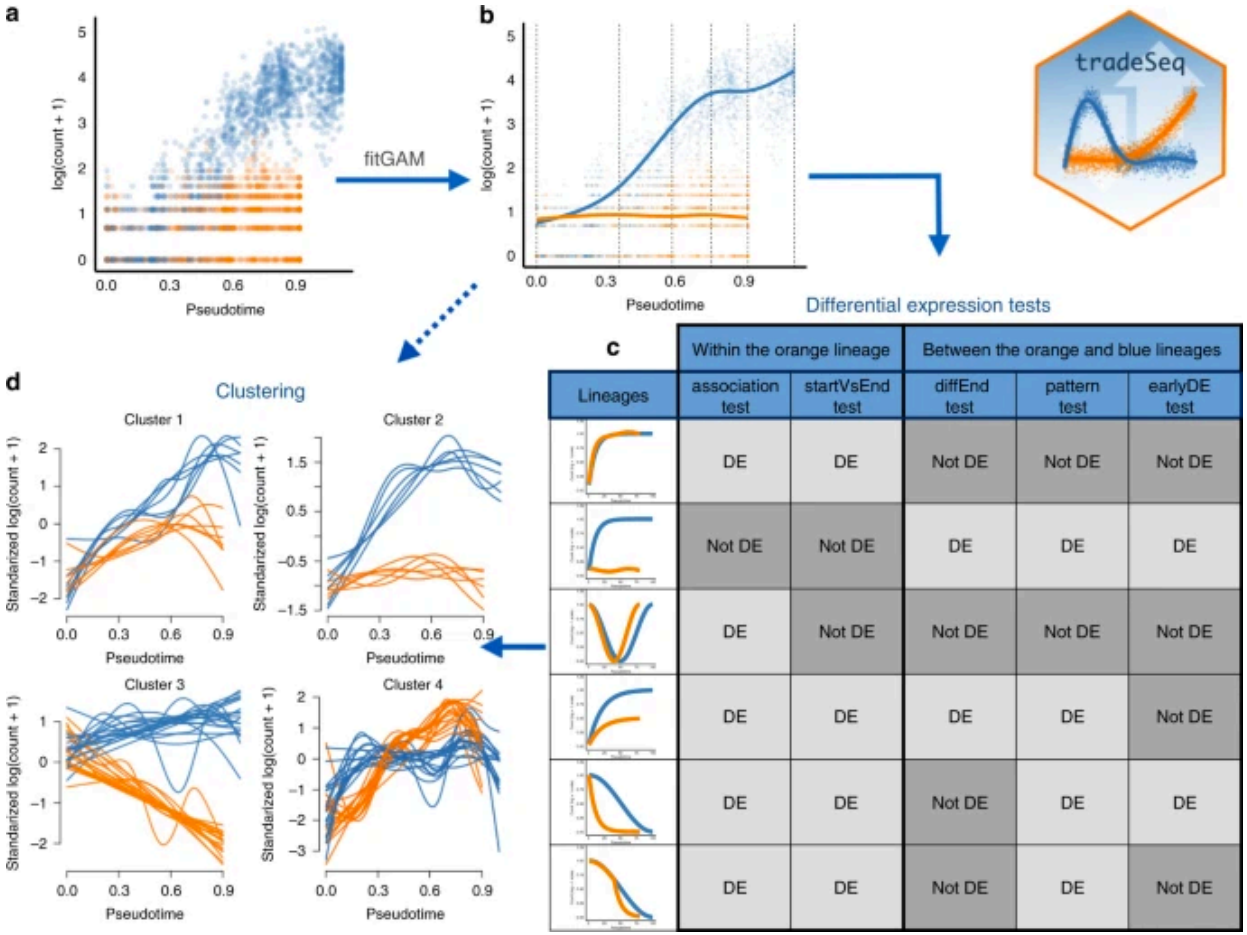
Procesamiento de datos: Integración de conjuntos de datos

En proyectos grandes de scRNA-seq es necesario generar los datos a través de múltiples *batches.*

- **La presencia de batches se puede detectar en un PCA** de valores de expresión logarítmicos
- Se han desarrollado métodos ***bespoke*** para la corrección de *batches* en datos de scRNA-seq que no requieren información a priori sobre la composición de la población
- Otros métodos que tratan de corregir los efectos de múltiples *batches* son **MNN** (*mutual nearest neighbor*) y **kBET** (*k-nearest neighbor batch effect test*)

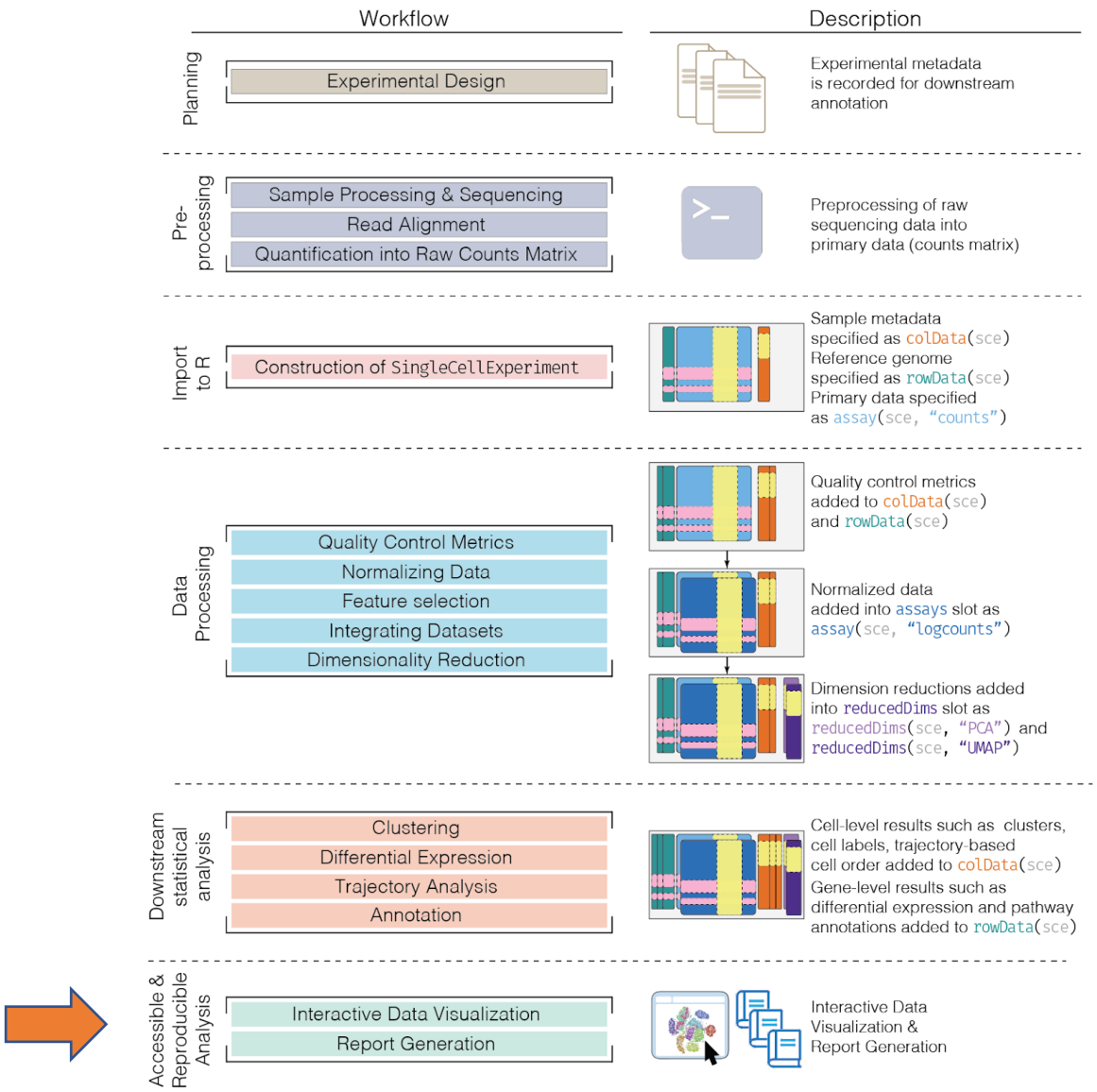

Workflow
Description
Planning
Experimental Design
Experimental metadata is recorded for downstream annotation
Pre-processing
Sample Processing & Sequencing
Read Alignment
Quantification into Raw Counts Matrix
Preprocessing of raw sequencing data into primary data (counts matrix)
Import to R
Construction 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

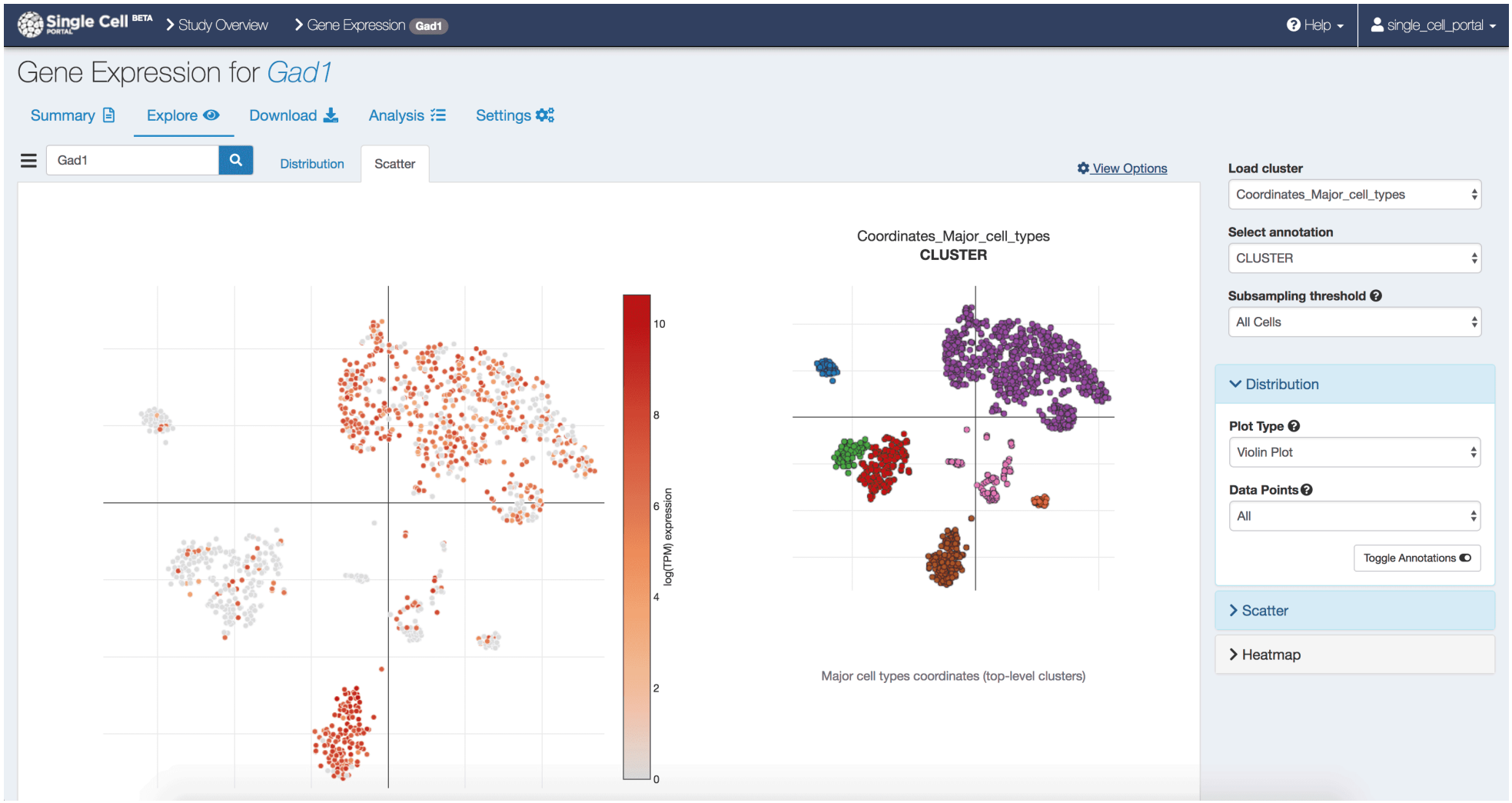
Análisis de trayectorias



Lineages	Within the orange lineage		Between the orange and blue lineages		
	association test	startVsEnd test	diffEnd test	pattern test	earlyDE test
	DE	DE	Not DE	Not DE	Not DE
	Not DE	Not DE	DE	DE	DE
	DE	Not DE	Not DE	Not DE	Not DE
	DE	DE	DE	DE	Not DE
	DE	DE	Not DE	DE	DE
	DE	DE	Not DE	DE	Not DE

https://www.nature.com/articles/s41467-020-14766-3


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



https://singlecell.broadinstitute.org/single_cell