

Universidad Abierta y a Distancia de México

DIVISIÓN DE CIENCIAS EXACTAS, INGENIERÍA Y TECNOLOGÍA

Licenciatura en Matemáticas

Normalización e integración de datos Single Cell RNA Sequencing en tejido intestinal

OMAR VARGAS HERNÁNDEZ
AL11503286

Asesora interna: DRA. MARÍA DEL ALBA PACHECO BLAS

Asesor externo: DR. ISRAEL AGUILAR ORDÓÑEZ

Querétaro, Qro., a 21 de noviembre de 2024

Universidad Abierta y a Distancia de México

Rectora

LILIAN KRAVZOV APPEL

Coordinación Académica y de Investigación

EDGAR ALCÁNTAR CORCHADO

Dirección de Ciencias Exactas, Ingeniería y Tecnología

DOLORES ALEJANDRA VASQUEZ CARBAJAL

Responsable del programa educativo de Matemáticas

CARLOS ALBERTO SERRATO HERNÁNDEZ

Docente en línea del proyecto terminal

DRA. MARÍA DEL ALBA PACHECO BLAS

Docentes en línea evaluadores del proyecto terminal

RAFAEL PACHECO ESPINOZA

ANAYANZI DELIA MARTÍNEZ HERNÁNDEZ

MARIA ELENA PACHECO CÓRDOVA

ALEJANDRA ZAMORA MEDINA

MANUEL LÓPEZ MATEOS

Institución receptora


Instituto Nacional de Medicina Genómica

Asesor externo

DR. ISRAEL AGUILAR ORDÓÑEZ

Índice general

Dedicatoria	vi
Agradecimientos	vii
Términos y bibliotecas usadas	x
Bioinformáticos	x
Matemáticos y Estadísticos	xi
Bibliotecas de R	xii
Resumen	xiv
1 Introducción	1
1.1. Condiciones del proyecto	4
1.2. Planteamiento del problema de investigación . . .	5
1.3. Pregunta de Investigación	8
1.4. Hipótesis	8
1.5. Objetivos	8
1.5.1. General	8
1.5.2. Específicos	8

2 Marco investigativo	9
3 Marco teórico	13
3.1. Parte 1: Biología molecular	13
3.1.1. Single Cell RNA Sequencing (scRNA-seq) .	13
3.1.2. Fuentes de variación en los datos de scRNA-Seq	15
3.2. Parte 2: Procesamiento de datos de scRNA-seq . .	16
3.2.1. El flujo de trabajo típico para el análisis de datos de secuenciación de scRNA-seq	16
3.2.2. Normalización de datos de scRNA-seq . . .	17
3.2.3. Integración de datos de scRNA-seq	19
3.2.4. Métricas de integración de datos de scRNA-seq	19
4 Metodología	25
4.1. Estrategia	25
4.2. Métodos	26
4.2.1. Datos de scRNA-seq	26
4.2.2. Normalización con la biblioteca de Seurat: LogNormalize	26
4.2.3. Integración con el algoritmo Harmony . . .	27
4.2.4. Criterios de selección de las métricas de evaluación de la integración	28
4.2.5. Métricas de evaluación de la integración . .	29
4.3. Repositorio GitHub 	30
4.4. Código L ^A T _E X	31

5 Resultados y análisis	32
5.1. Escenario 1, paciente 17	34
5.2. Escenario 1, paciente 19	37
5.3. Escenario 2, ambos pacientes	39
6 Conclusiones	45
6.1. Limitaciones	47
6.2. Perspectivas	48
6.3. Aportación	49
Referencias	51
Índice de figuras	53
Índice de tablas	54

Dedicatoria

Para mi pequeña y linda hija Fadia,
que desde que nació ha sido la luz de mi vida.

Agradecimientos

A Dios,
por el Universo, la Tierra, la Vida y la Existencia.

A Jesucristo,
por ser el camino a Dios y el perdón de los pecados.

A mi madre Fadia,
por darme la vida, sacrificarse por mí, apoyarme en todo momento, desde pequeño hasta adulto, sus consejos, corajes y lágrimas, por todo lo que sé de los valores morales, el bien y el mal, por acercarme a Dios y a mi hija, por mostrarme lo que vale la pena y lo que no, por apoyar este proyecto y mi formación profesional y siempre insistir en terminar, y claro, por ser mi primer amor.

A mi padre Marcos Gerardo,
por estar ahí, cerca de mí, apoyándome, dándome consejos y regaños, impulsándome y siempre estar a lado de mi madre,

amándonos a ella y a mí.

A mi pequeña y hermosa hija Fadia,
mi gran amor, que me ha inspirado desde que nació con su pequeña sonrisa y su caluroso abrazo, el brillo de sus ojos es una fuente inagotable de inspiración y poder, la responsabilidad de verla crecer y demostrarle que, aunque las cosas tardan en pasar, no debemos rendirnos, me impulsó siempre para terminar estos estudios, además de la gran alegría continua que tengo por su encantadora presencia.

A mi amigo y compadre Ricardo Guerra,
por el tiempo de amistad, discusiones, diseños experimentales, por las pláticas de política e innumerables temas, por su ayuda con el manejo de paquetes informáticos, y sus invaluable consejos profesionales para reducir mi curva de aprendizaje en esta nueva área del conocimiento.

A mi asesora interna María del Alba Pacheco,
por su tiempo, dedicación y consejo acerca de todas las partes de este proyecto, orientándome en todo momento y con toda paciencia para cumplir con el perfil de egreso.

A mi asesor externo, tutor y responsable del proyecto Israel Aguilar,

por confiar en mí, desde el principio, darme una oportunidad de trabajar en su equipo, participé en sus seminarios y cursos varios, aprendí mucho gracias a él, me capacitó con sus cursos de bioinformática, diseñó estrategias y retroalimentó en cada situación teórica y práctica a la cuál nos enfrentamos.

A los Docentes en línea evaluadores del proyecto terminal, Rafael Pacheco, Anayanzi Delia Martínez, Maria Elena Pacheco, Alejandra Zamora, Manuel López,
por la evaluación de la defensa oral del proyecto.

A la responsable del Programa de Participación Estudiantil del Instituto Nacional de Medicina Genómica, Alejandra Elizabeth Rangel,
por su ayuda invaluable en el registro, inscripción, alta y baja en el Instituto, así como por la gestión de las cartas de aceptación y terminación.

Al Instituto Nacional de Medicina Genómica (INMEGEN),
por la autorización de la realización de este proyecto.

A la Universidad Abierta y a Distancia de México (UnADM),
por ser el alma máter para la realización de la Licenciatura en Matemáticas.

Términos y bibliotecas usadas

Bioinformáticos

- **Expresión genética:** Conjunto de valores de expresión para todos los genes en una célula. Este perfil es único y permite identificar tipos celulares y estados biológicos.
- **scRNA-seq (single cell RNA sequencing):** Técnica de biología molecular que permite analizar la expresión génica a nivel de célula individual.
- **Matriz de conteos:** Representación matricial de datos de scRNA-seq, donde las filas son genes, las columnas son células, y los valores indican la expresión génica.
- **Lote (batch):** Conjunto de datos generado bajo condiciones experimentales similares, susceptible a variaciones técnicas.
- **Profundidad de secuenciación** Cantidad de veces que una secuencia de RNA es leída durante el proceso de secuenciación, determinando la precisión y sensibilidad en la detección de transcritos.

- **Normalización:** Proceso para ajustar los datos eliminando variaciones técnicas, como la profundidad de secuenciación.
- **Integración de Datos:** Combinación de matrices de conteo de diferentes lotes para permitir comparaciones válidas.

Matemáticos y Estadísticos

- **Coeficiente de variación:** Porcentaje que mide la variación relativa entre datos antes y después de la integración.
- **Clúster celular:** Conjunto de células con perfiles de expresión génica similares, agrupadas mediante métodos computacionales para identificar subpoblaciones o tipos celulares en análisis de scRNA-seq.
- **UMAP (Uniform Manifold Approximation and Projection):** Método de reducción de dimensionalidad que preserva relaciones locales entre datos en espacios de alta dimensión, utilizado para visualizar estructuras como clústeres celulares en scRNA-seq.
- **Dimensionalidad reducida:** Transformación de datos a un espacio con menos dimensiones (ejemplo: UMAP) para facilitar la visualización y el análisis.

Bibliotecas de R

Información acerca del entorno de programación, sistema operativo y bibliotecas cargadas para realizar el trabajo. Presento a continuación la salida del comando `sessionInfo()` en R:

```
R version 4.4.1 (2024-06-14)
Platform: x86_64-pc-linux-gnu
Running under: Ubuntu 24.04.1 LTS

Matrix products: default
BLAS:   /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.12.0
LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.12.0

locale:
 [1] LC_CTYPE=es_ES.UTF-8      LC_NUMERIC=C               LC_TIME=es_MX.UTF
    -8
 [4] LC_COLLATE=es_ES.UTF-8    LC_MONETARY=es_MX.UTF-8    LC_MESSAGES=es_ES
    .UTF-8
 [7] LC_PAPER=es_MX.UTF-8      LC_NAME=C                  LC_ADDRESS=C
[10] LC_TELEPHONE=C            LC_MEASUREMENT=es_MX.UTF-8 LC_IDENTIFICATION
    =C

time zone: America/Mexico_City
tzcode source: system (glibc)

attached base packages:
[1] stats      graphics  grDevices  utils      datasets  methods    base

other attached packages:
[1] ggplot2_3.5.1              openxlsx_4.2.7.1           scIntegrationMetrics_
    1.1
 [4] harmony_1.2.1              Rcpp_1.0.13-1              Seurat_5.1.0
 [7] SeuratObject_5.0.2         sp_2.1-4                   pacman_0.5.1
[10] rmarkdown_2.27

loaded via a namespace (and not attached):
 [1] RColorBrewer_1.1-3         rstudioapi_0.16.0          jsonlite_1.8.8
    magrittr_2.0.3
 [5] spatstat.utils_3.1-0       farver_2.1.2               vctr_0.6.5
    ROCR_1.0-11
 [9] spatstat.explore_3.3-3     htmltools_0.5.8.1          sctransform_0.4.1
    parallelly_1.38.0
[13] KernSmooth_2.23-24         htmlwidgets_1.6.4          ica_1.0-3
    plyr_1.8.9
[17] plotly_4.10.4              zoo_1.8-12                 igraph_2.1.1
    mime_0.12
```

[21] lifecycle_1.0.4 R6_2.5.1	pkgconfig_2.0.3	Matrix_1.7-1
[25] fastmap_1.2.0 shiny_1.8.1.1	fitdistrplus_1.2-1	future_1.34.0
[29] digest_0.6.36 tensor_1.5	colorspace_2.1-1	patchwork_1.3.0
[33] RSpectra_0.16-2 labeling_0.4.3	irlba_2.3.5.1	vegan_2.6-8
[37] progressr_0.15.0 httr_1.4.7	fansi_1.0.6	spatstat.sparse_3.1-0
[41] polyclip_1.10-7 compiler_4.4.1	abind_1.4-5	mgcv_1.9-1
[45] withr_3.0.2 MASS_7.3-61	fastDummies_1.7.4	R.utils_2.12.3
[49] permute_0.9-7 zip_2.3.1	tools_4.4.1	lmtest_0.9-40
[53] httpuv_1.6.15 R.oo_1.27.0	future.apply_1.11.3	goftest_1.2-3
[57] glue_1.7.0 grid_4.4.1	nlme_3.1-165	promises_1.3.0
[61] Rtsne_0.17 generics_0.1.3	cluster_2.1.6	reshape2_1.4.4
[65] gtable_0.3.5 tidyr_1.3.1	spatstat.data_3.1-2	R.methodsS3_1.8.2
[69] data.table_1.15.4 RcppAnnoy_0.0.22	utf8_1.2.4	spatstat.geom_3.3-3
[73] ggrepel_0.9.6 stringr_1.5.1	RANN_2.6.2	pillar_1.9.0
[77] spam_2.11-0 splines_4.4.1	RcppHNSW_0.6.0	later_1.3.2
[81] dplyr_1.1.4 deldir_2.0-4	lattice_0.22-5	survival_3.7-0
[85] tidyselect_1.2.1 knitr_1.47	miniUI_0.1.1.1	pbapply_1.7-2
[89] gridExtra_2.3 xfun_0.45	scattermore_1.2	RhpcBLASctl_0.23-42
[93] matrixStats_1.3.0 evaluate_0.24.0	stringi_1.8.4	lazyeval_0.2.2
[97] codetools_0.2-20 uwot_0.2.2	tibble_3.2.1	cli_3.6.3
[101] xtable_1.8-4 globals_0.16.3	reticulate_1.39.0	munSELL_0.5.1
[105] spatstat.random_3.3-2 parallel_4.4.1	png_0.1-8	spatstat.univar_3.0-1
[109] dotCall64_1.2 scales_1.3.0	listenv_0.9.1	viridisLite_0.4.2
[113] ggrridges_0.5.6 rlang_1.1.4	leiden_0.4.3.1	purrr_1.0.2
[117] cowplot_1.1.3		

Resumen

Este proyecto se enfoca en un área crítica de la Bioinformática: la integración de datos de Single Cell RNA Sequencing (scRNA-seq). La integración de datos es el proceso por el cual se filtran y limpian los datos de diferentes personas disminuyendo la variabilidad técnica entre estos para hacerlos comparables. Se evaluaron 8 métricas de la calidad de la integración de los datos, consistentes con las técnicas de análisis en nuestro laboratorio. Se generaron dos escenarios hipotéticos para evaluar que las métricas cumplieran con los criterios necesarios para su uso e interpretación. El primer escenario es construir 2 subconjuntos que provenían de un solo paciente y el segundo escenario con 2 conjuntos provenientes de 2 pacientes distintos. Para el escenario 1, se hicieron pruebas con datos del paciente 17 y del paciente 19, el coeficiente de variación calculado entre los datos no integrados y los datos integrados, fue menor al 8 % para todas las métricas. Para el escenario 2, solo 5 métricas mostraban coeficientes de variación mayores al 8 %; que era lo buscado, porque significa que la métrica podía discriminar entre escenarios. Al final, calculamos el coeficiente de variación entre los datos integrados de un mismo paciente y los datos integrados de pacientes distintos, escogimos los 3 valores más

altos (los que mejor discriminan escenarios) y por conveniencia interpretativa, seleccionamos una métrica que además estuviera normalizada (valores entre 0 y 1), de lo cual concluimos que la mejor métrica era norm iLSI. Con lo cual se garantiza que los resultados son comparables y válidos para futuros análisis e investigaciones en el área de biología y medicina.

Palabras clave: Perfil de transcripción, scRNA-seq, Matriz de conteos, Integración de datos, Métricas.

Capítulo 1

Introducción

En el campo de la bioinformática, la normalización e integración de datos de Single Cell RNA Sequencing (scRNA-seq) ha emergido como una herramienta crucial para comprender la complejidad y diversidad de los procesos biológicos a nivel celular. Este proyecto se centra en la aplicación de estas técnicas en el análisis de tejido intestinal, un área de estudio que ofrece importantes perspectivas tanto para la biología fundamental como para la investigación médica.

La integración de datos scRNA-seq es una meta valiosa para realizar metaanálisis, ya que permite combinar información de múltiples experimentos y obtener una visión más amplia de los procesos biológicos. La capacidad de analizar datos a nivel de célula individual proporciona una resolución sin precedentes de la expresión génica, lo que es esencial para desentrañar los mecanismos subyacentes en la fisiología y patología del tejido intestinal. Esto puede tener implicaciones significativas en la comprensión de enfermedades intestinales, como el cáncer colorrectal y las enfermedades inflamatorias intestinales, así como en el desarrollo de tratamientos personalizados.

Sin embargo, este proceso enfrenta grandes desafíos debido a las diversas fuentes de variación presentes en los datos. Estas variaciones pueden originarse de diferencias técnicas entre experimentos, como el uso de diferentes tecnologías de secuenciación, lotes de reactivos o condiciones de procesamiento de muestras, así como de variaciones biológicas intrínsecas. Para mitigar estas variaciones, se llevan a cabo varios pasos cruciales: la normalización de los datos, la corrección de efectos batch y, finalmente, la integración de las matrices de conteo.

La normalización de datos es el primer paso fundamental en el análisis de scRNA-seq. Este proceso ajusta las diferencias en la profundidad de secuenciación y otras fuentes técnicas de variación para permitir comparaciones válidas entre células. Sin una normalización adecuada, los análisis subsiguientes pueden ser sesgados y no reflejar la verdadera biología subyacente. A continuación, la corrección de efectos batch se enfoca en eliminar las variaciones que no están relacionadas con las diferencias biológicas reales pero que son introducidas por el procesamiento de muestras en diferentes momentos o condiciones. La integración de matrices de conteo, el paso final, combina los datos de múltiples experimentos en un único conjunto de datos cohesivo que puede ser analizado en su conjunto (KOTLOV et al. [2024](#)).

A pesar de la importancia de estos pasos, en nuestro laboratorio no conocemos una forma de saber si la normalización fue realizada correctamente. Además, tampoco tenemos claridad sobre cómo determinar si la integración está bien hecha,

lo cual puede estar relacionado con una normalización inadecuada. Este problema se agrava por la falta de métricas claras y objetivas que evalúen cada uno de estos procedimientos, dejando una brecha significativa en la validación de la calidad del análisis de datos de scRNA-seq (FORCATO, ROMANO y BICCIATO 2021).

El objetivo principal de este proyecto es proponer una métrica para evaluar la calidad del proceso de integración de matrices de conteo en el análisis de datos de scRNA-seq, usando como datos de entrada los de tejido intestinal de pacientes. Para lograr esto, se realizará una comparación de cuatro algoritmos de normalización y cuatro algoritmos de integración de datos, así como de tres métricas de evaluación de calidad preexistentes. Este enfoque permitirá identificar las estrategias más efectivas para asegurar que los análisis sean consistentes y reproducibles (ANDREATTA et al. 2024; MA et al. 2020).

La importancia de este trabajo radica no solo en su contribución al campo de la bioinformática, sino también en su potencial para mejorar la confiabilidad de los estudios de scRNA-seq. Esto es particularmente relevante para la investigación en biología y medicina, donde los resultados precisos y comparables son fundamentales para el avance del conocimiento y la aplicación clínica. El desarrollo de una métrica de evaluación de calidad específica no solo beneficiará a los investigadores en la validación de sus datos, sino que también establecerá un estándar que puede ser adoptado por la comunidad científica para garantizar la integridad de los análisis.

La estructura de este trabajo está organizada de la siguiente manera: En primer lugar, se contextualiza y plantea el problema, explicando la relevancia de la integración de datos scRNA-seq y los desafíos asociados. A continuación, se detallan los objetivos generales y específicos del proyecto. En la metodología, se describen los algoritmos y métricas evaluadas. Los resultados obtenidos se analizan en términos de su eficacia y robustez, seguidos de una discusión sobre las implicaciones y aplicaciones de estos hallazgos. Finalmente, se concluye con una reflexión sobre las direcciones futuras de investigación en este campo.

Bajo la supervisión del Dr. Israel Aguilar Ordóñez, este proyecto se está llevado a cabo en el Instituto Nacional de Medicina Genómica (INMEGEN). Confiamos en que los hallazgos de este estudio contribuirán significativamente a la mejora de las técnicas de análisis de datos scRNA-seq y, en última instancia, a un mejor entendimiento de los procesos biológicos en el tejido intestinal. Este avance no solo potenciará el conocimiento básico, sino que también abrirá nuevas vías para la investigación translacional y el desarrollo de terapias personalizadas.

1.1 Condiciones del proyecto

Este trabajo se realizó en el Instituto Nacional de Medicina Genómica (INMEGEN) y estuvo a cargo del Dr. Israel Aguilar Ordóñez, jefe del Departamento de Supercómputo, especialista en Bioinformática y que fungió como asesor externo del proyecto ante la UnADM.

Estuve dado de alta en el INMEGEN en el Programa de Participación Estudiantil bajo la modalidad de estancia voluntaria, con una duración de 7 meses, desde febrero del 2024 y hasta agosto del 2024, en calidad de estudiante de licenciatura y sin relación laboral alguna.

1.2 Planteamiento del problema de investigación

El cuerpo humano está conformado por células (aproximadamente 200 tipos distintos) que se organizan de maneras complejas para formar, secuencialmente, tejidos, órganos, sistemas y el organismo completo.

Aunque las células de un mismo cuerpo tienen la misma información genética (porque son todas clones de la primera célula que inició la formación de ese cuerpo), no expresan todos esos genes al mismo tiempo (en total hay aproximadamente 20,000 genes humanos), sino que, dependiendo del tipo celular, condiciones locales y su estado de salud (normal o enfermo), expresan diferentes genes.

Al conjunto de genes que las células expresan le llamamos perfil de expresión genética, o perfil transcripcional. Este perfil de expresión genética constituye en la realidad una huella molecular de la identidad de esa célula.

La técnica de biología molecular llamada Single Cell RNA Sequencing (scRNA-seq) determina los perfiles de expresión de cada una de las células provenientes de una muestra de un paciente.

Analizar esos datos permite determinar las identidades de esas células, que podrían ser células enfermas, como cancerosas o con otras patologías, que podrían reconocerse relativamente rápido usando la técnica de scRNA-seq.

Finalmente, lo que se obtiene al realizar la técnica de scRNA-seq es una matriz de conteos $M_{i \times j}$, cuyas filas son genes y las columnas son células, los valores constituyen la cantidad de expresión del gen i en la célula j .

MATRIZ DE CONTEOS			
	Célula 1	Célula 2	...
Gen1	18	0	
Gen2	1010	506	
Gen3	0	49	
Gen4	22	0	
...			

Figura 1 Matriz de conteos. Imagen modificada de: RNA-Seq. (2024). Transcriptome Sequencing Research & Industry News. <https://www.rna-seqblog.com/top-benefits-of-using-the-technique-of-single-cell-rna-seq/>.

Por cada paciente analizado se obtiene una matriz de conteos independiente, y, para poder hacer los procedimientos estadísticos, las matrices de conteo tienen que estar normalizadas o escaladas para finalmente ser integradas (unión de varias matrices de conteo en una matriz global de conteo).

La integración de datos scRNA-seq es una meta valiosa para realizar meta-análisis, ya que permite combinar información de múltiples experimentos y obtener una visión más amplia de los procesos biológicos.

Sin embargo, este proceso se enfrenta a grandes desafíos debido a las diversas fuentes de variación presentes en los datos. Para mitigar estas variaciones, se llevan a cabo varios pasos cruciales: la normalización de los datos, la corrección de efectos batch y, finalmente, la integración de las matrices de conteo.

A pesar de la importancia de estos pasos, en nuestro laboratorio no conocemos una forma de saber si la integración está bien hecha, lo cual puede estar relacionado con una normalización inadecuada.

Este problema se agrava por la falta de métricas claras y objetivas que evalúen cada uno de estos procedimientos, dejando una brecha significativa en la validación de la calidad del análisis de datos de scRNA-seq. Y este problema detona la realización de este proyecto.

1.3 Pregunta de Investigación

¿Cuál es la mejor métrica para evaluar la integración de datos scRNA-seq?

1.4 Hipótesis

Existe una manera eficaz de evaluar (mediante una escala numérica) la calidad de la integración de datos scRNA-seq que tenemos disponibles en el laboratorio.

1.5 Objetivos

1.5.1 General

Evaluar métricas que determinan la calidad de la integración de matrices de conteo en el análisis de datos de scRNA-seq.

1.5.2 Específicos

1. Ejecutar las bibliotecas de R necesarias para trabajar con Seurat, que es la paquetería de análisis de datos de scRNA-seq que se usa en el laboratorio.
2. Generar código en R para integrar datos scRNA-seq con el algoritmo Harmony sobre objetos Seurat.
3. Escoger una métrica y fundamentar biológica y matemáticamente su uso.

Capítulo 2

Marco investigativo

El análisis de la expresión genética de células únicas mediante la técnica de scRNA-seq ha emergido como una herramienta crucial en la comprensión de la heterogeneidad tumoral y la dinámica del microambiente inmune en sarcomas, como lo revela el estudio **“Single cell analysis reveals distinct immune landscapes in transplant and primary sarcomas that determine response or resistance to immunotherapy”** (WISDOM et al. [2020](#)). En esta investigación, se exploró el terreno complejo de los sarcomas, centrándose en la normalización de datos e integración para obtener una visión completa del escenario inmune que rodea a estos tumores.

La pregunta clave que guió esta investigación fue la siguiente: ¿Cómo difieren los escenarios inmunes de sarcomas trasplantados y primarios, y cómo estas diferencias inciden en la respuesta a la inmunoterapia? La hipótesis planteada sugirió que los distintos microambientes inmunes presentes en los sarcomas trasplantados y primarios pueden ser determinantes en su sensibilidad a la inmunoterapia. Además, se especuló que los pacientes cuyos sarcomas exhiben un fenotipo inmune si-

milar al de los tumores trasplantados podrían obtener mayores beneficios del bloqueo de PD-1 y la radioterapia.

Para abordar estas cuestiones, el estudio utilizó técnicas avanzadas de análisis unicelular, como scRNA-seq, para perfilar las células inmunes y tumorales presentes en sarcomas trasplantados y primarios. Sin embargo, el mero análisis de datos no fue suficiente; la normalización de datos e integración de información proveniente de diferentes conjuntos de datos resultaron ser aspectos cruciales para alcanzar una comprensión holística del complejo escenario inmune asociado a los sarcomas.

Entre los hallazgos más destacados se encontraron diferencias significativas en la infiltración de células inmunes entre sarcomas trasplantados y primarios, así como la identificación de marcadores inmunitarios asociados con la respuesta a la inmunoterapia en estos tumores. Estos resultados subrayan el potencial del análisis unicelular para prever la respuesta a la inmunoterapia en pacientes con sarcoma, lo que podría abrir nuevas puertas hacia enfoques terapéuticos más precisos y efectivos.

Desde una perspectiva clínica, los hallazgos de este estudio tienen implicaciones profundas para el tratamiento del sarcoma. No solo ayudan a comprender los mecanismos subyacentes que determinan la respuesta a la inmunoterapia en estos tumores, sino que también ofrecen una base sólida para el desarrollo de biomarcadores que puedan predecir esta respuesta. En última instancia, estos conocimientos permiten la formulación de estrategias de tratamiento más personalizadas y

efectivas para los pacientes con sarcoma, lo que representa un avance significativo en la lucha contra esta enfermedad.

Otro ejemplo lo tenemos en el estudio **“Single-cell atlases: shared and tissue-specific cell types across human organs”**, publicado en la revista *Nature Reviews Genetics* (ELMENTAITE et al. [2022](#)), se basa en el análisis de más de 2 millones de células individuales de 93 órganos humanos utilizando técnicas de secuenciación scRNA-seq. Este enfoque permitió a los investigadores obtener una comprensión detallada de la composición celular en una amplia variedad de tejidos.

El estudio se centró en la identificación de tipos celulares compartidos y específicos de tejidos en los órganos humanos, con el objetivo de construir atlas unicelulares completos. Para lograr esto, fue crucial aplicar técnicas de normalización y integración de datos. La normalización de datos permitió corregir las diferencias técnicas y biológicas entre las células, mientras que la integración de datos combinó información de múltiples conjuntos de datos para obtener una visión global del paisaje celular en los órganos humanos.

A través de este enfoque integrado, el estudio identificó más de 200 tipos celulares distintos en los órganos analizados, revelando la complejidad y diversidad de las células que componen el cuerpo humano. Además, se descubrieron nuevos tipos celulares especializados en cada órgano, así como tipos celulares compartidos que desempeñan funciones fundamentales en la circulación sanguínea y la respuesta inmune.

Estos hallazgos tienen importantes implicaciones para la investigación biomédica y la medicina clínica, ya que proporcionan una base sólida para comprender mejor las enfermedades y desarrollar terapias más precisas y efectivas. Los atlas unicelulares creados en este estudio no solo sirven como una referencia invaluable para futuras investigaciones sobre la biología humana y la salud, sino que también destacan la importancia de la integración de datos y la normalización en el análisis de datos de secuenciación de RNA unicelular.

Capítulo 3

Marco teórico

3.1 Parte 1: Biología molecular

3.1.1 Single Cell RNA Sequencing (scRNA-seq)

La técnica scRNA-seq (secuenciación de RNA en células individuales) es un método de biología molecular que permite analizar la expresión génica a nivel de una sola célula. Los pasos principales para realizarla son (HAQUE et al. [2017](#)):

1. **Aislamiento de células individuales:** Se emplean dispositivos microfluídicos para capturar células individualmente en gotas o pozos.
2. **Captura y lisis celular:** Cada célula individual se encapsula o se coloca en un pozo separado, donde se lisa (rompe) para liberar su contenido de RNA.
3. **Transcripción inversa y amplificación:** El RNA mensajero (mRNA) liberado se convierte en ADN complementario (cDNA) mediante la transcripción inversa. El cDNA se

amplifica para generar suficientes copias para la secuenciación.

4. **Etiquetado y secuenciación:** Se agregan códigos de barras a cada molécula de cDNA para rastrear el origen celular durante el análisis de secuenciación. La biblioteca de cDNA etiquetada se somete a secuenciación masiva en paralelo utilizando secuenciadores de alto rendimiento.
5. **Análisis Bioinformático:** Los datos de secuenciación se procesan para mapear las lecturas a un genoma de referencia y cuantificar la expresión génica. Posteriormente, Se utilizan algoritmos y herramientas bioinformáticas para agrupar células con perfiles de expresión génica similares, identificar subpoblaciones celulares, y estudiar rutas y redes de señalización celulares.

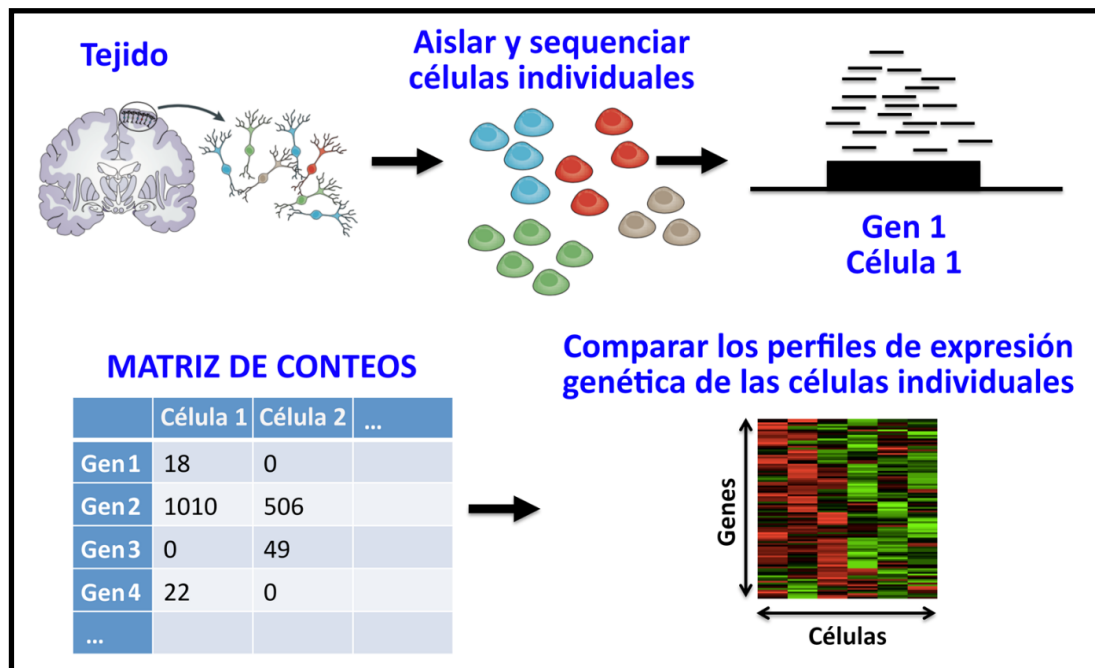


Figura 2 Pasos principales en la técnica de scRNA-seq. Imagen modificada de: RNA-Seq. (2024). Transcriptome Sequencing Research & Industry News. <https://www.rna-seqblog.com/top-benefits-of-using-the-technique-of-single-cell-rna-seq/>.

3.1.2 Fuentes de variación en los datos de scRNA-Seq

Existen varias fuentes de variación en los datos de scRNA-seq, lo cual es importante tener en cuenta durante el análisis para obtener resultados precisos. En la tabla siguiente se separan las fuentes de variación en dos categorías principales: origen biológico y origen no biológico, y dentro de cada categoría se distinguen varios ejemplos de variación (ANDREWS et al. [2020](#)).

Origen biológico	Origen no biológico
Variación biológica	Variación técnica
Contaminación celular	Efectos batch
Variación ambiental	Ruido de medición

Tabla 1 Fuentes de variación en los datos de scRNA-seq. Ver el texto.

Las principales fuentes de variación de origen biológico son:

- **Variación biológica:** Proviene de la propia biología de las células y de su entorno.
- **Contaminación celular:** Se refiere a la presencia de material genético de otras células en una muestra individual. Esto puede ocurrir durante la manipulación y el procesamiento de las muestras, lo que lleva a una mezcla de señales genéticas de múltiples células.
- **Variación ambiental:** Incluye las diferencias en la expresión génica causadas por variaciones en el microambiente celular. Factores como la disponibilidad de nutrientes, los niveles de oxígeno, y las señales químicas externas pueden influir en la actividad génica de manera significativa.

Las principales fuentes de variación de origen no biológico son:

- **Variación técnica:** Esta variación es introducida por los procesos técnicos y metodológicos utilizados durante la preparación y análisis de las muestras.
- **Efectos batch:** Se refiere a las variaciones sistemáticas introducidas durante el procesamiento de diferentes lotes de muestras. Estas variaciones pueden ser causadas por diferencias en los reactivos, las condiciones experimentales, el equipo utilizado, o los técnicos que manejan las muestras.
- **Ruido de medición:** Es la variabilidad introducida por ineficiencias y errores en las técnicas de medición, como la transcripción inversa, la amplificación por PCR, y la secuenciación. Factores técnicos como la eficiencia variable de las enzimas y la captura incompleta de mRNA contribuyen a este ruido.

3.2 Parte 2: Procesamiento de datos de scRNA-seq

3.2.1 El flujo de trabajo típico para el análisis de datos de secuenciación de scRNA-seq

El procesamiento de datos de secuenciación de scRNA-seq es un proceso complejo que incluye varias etapas críticas para asegurar la calidad y la comparabilidad de los datos, la manera típica de hacerlo sería:

1. **Preprocesamiento:** Filtrado de células y genes de baja calidad, eliminación de dobles (doublets) y evaluación de calidad.
2. **Normalización:** Ajuste de la matriz de conteos para comparaciones entre células.
3. **Identificación y corrección de efectos batch:** Uso de métodos de bibliotecas como ComBat, Harmony o Seurat para eliminar variaciones técnicas no deseadas.
4. **Reducción de dimensionalidad:** Técnicas como PCA, t-SNE o UMAP para visualizar y analizar la variabilidad en los datos.
5. **Clustering e identificación de tipos celulares:** Agrupamiento de células similares y anotación de clústeres para identificar subpoblaciones celulares.
6. **Integración de datos:** Combinar datos de múltiples experimentos utilizando los métodos incluidos en bibliotecas como Seurat, Scanorama o Harmony (Lu et al. [2023](#); Malte D LUECKEN y THEIS [2019](#)).

3.2.2 Normalización de datos de scRNA-seq

La normalización corrige variaciones técnicas que pueden surgir debido a la eficiencia de captura de RNA, profundidad de secuenciación, y otros artefactos experimentales (LYTAL, RAN y AN [2020](#)). Los objetivos principales de la normalización son entonces calcular:

- **Eficiencia de captura de RNA:** Se refiere a la proporción de RNA mensajero que es capturada y convertida en cDNA durante el proceso de secuenciación.
- **Profundidad de secuenciación:** Indica la cantidad de veces que una secuencia de RNA es leída durante el proceso de secuenciación, influyendo en la precisión y sensibilidad de la detección de transcritos.
- **Biblioteca de secuenciación:** Conjunto de fragmentos de cDNA preparados para ser secuenciados, representando las moléculas de RNA originales de la muestra.

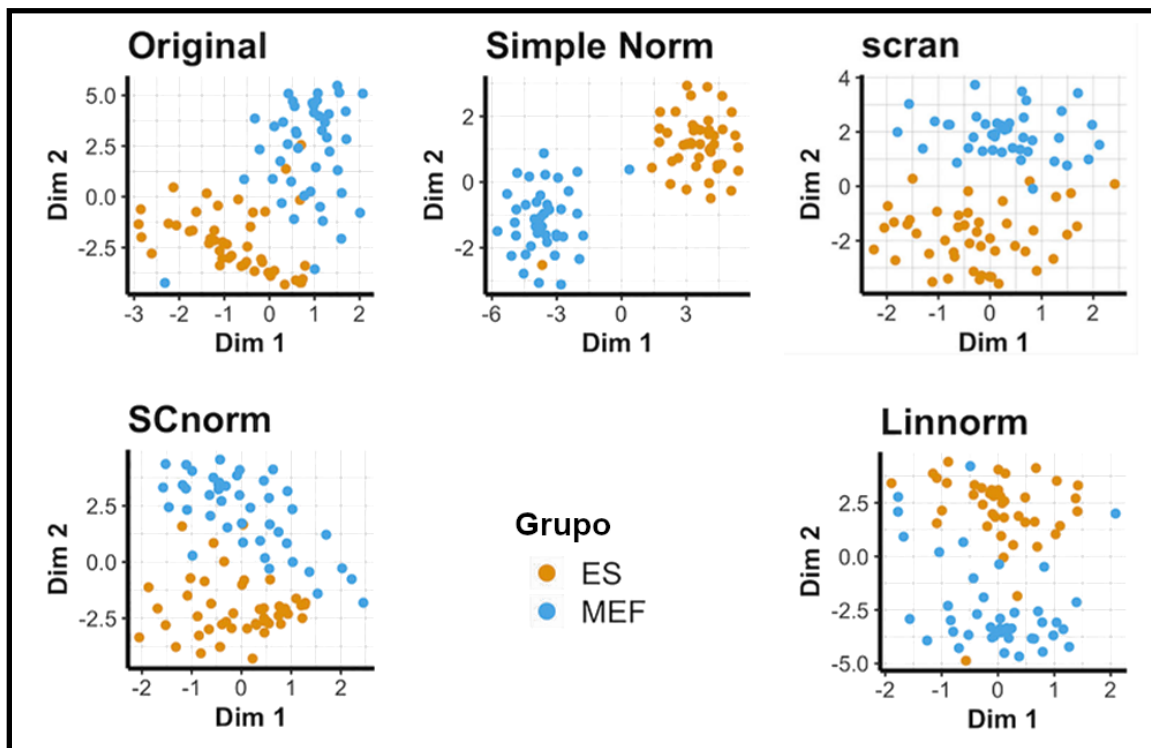


Figura 3 Diferentes algoritmos de normalización de datos de scRNA-seq. Se comparan diferentes algoritmos de normalización, los datos originales están en el primer panel. Los gráficos de dispersión se realizan después de la reducción de dimensionalidad (Dim1 y Dim2). ES y MEF son tipos celulares de embrión de ratón. Modificado de (LYTAL, RAN y AN [2020](#)).

3.2.3 Integración de datos de scRNA-seq

La integración de datos de secuenciación de scRNA-seq implica combinar datos provenientes de diferentes experimentos, lotes o tecnologías para analizarlos conjuntamente (FORCATO, ROMANO y BICCIATO 2021). Los objetivos principales de la integración son:

- **Conservar variabilidad biológica:** Mantener las diferencias biológicas genuinas mientras se eliminan las variaciones técnicas.
- **Facilitar comparaciones:** Permitir la comparación directa entre células de diferentes experimentos o condiciones.
- **Mejorar la robustez del análisis:** Aumentar la robustez y la reproducibilidad del análisis de datos integrando múltiples fuentes de datos.

Se puede ver un resumen en la Figura 4.

3.2.4 Métricas de integración de datos de scRNA-seq

Se evaluaron 8 métricas para determinar la calidad de la integración de los datos de scRNA-seq (ANDREATTA et al. 2024; KORSUNSKY et al. 2019), la biblioteca en lenguaje R se puede encontrar en: <https://github.com/carmonalab/scIntegrationMetrics>.

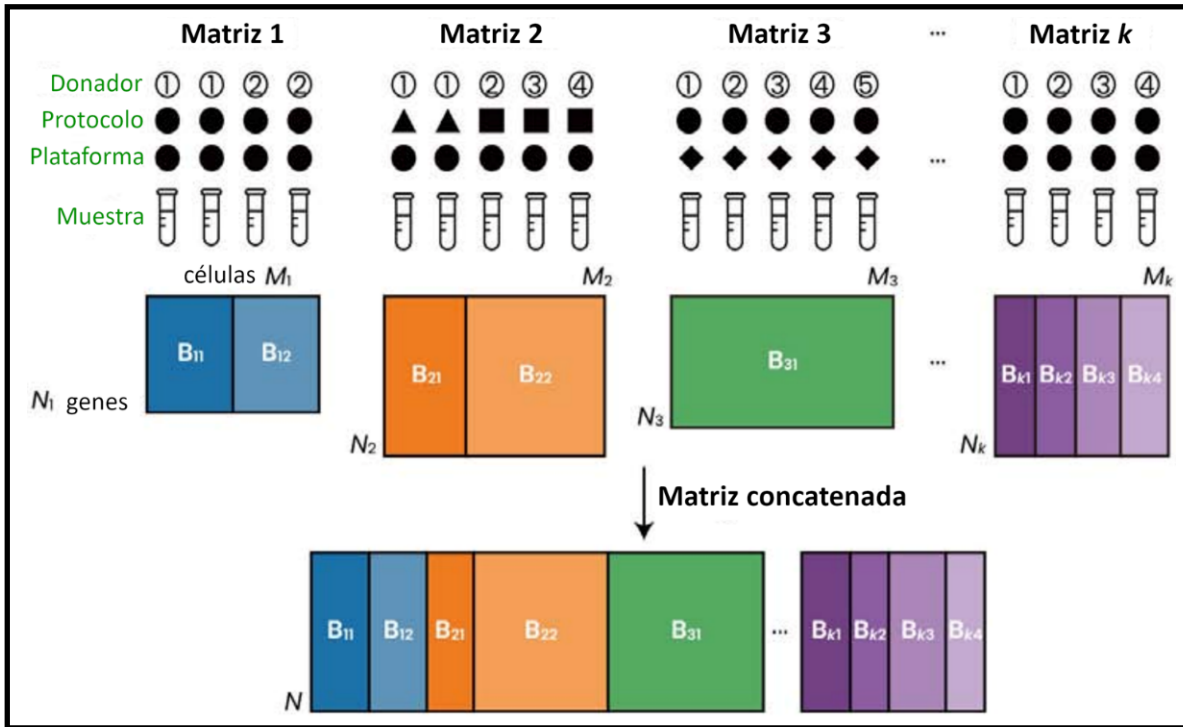


Figura 4 Integración de datos de scRNA-seq. Esquema de las fuentes de variación: por donantes (Matriz 1), protocolos de preparación de muestras (Matriz 2), plataformas de secuenciación (Matriz 3) y muestras individuales (Matriz k). Al último se muestra la matriz concatenada (integrada). Las filas son genes y las columnas células. Modificado de (Ryu et al. 2023).

Primero se describen las variables que se ocuparán y posteriormente una explicación y fórmula para calcular cada métrica.

i : Índice para una célula.

j : Índice para otra célula.

k : Índice para un tipo celular.

k' : Índice para un tipo celular diferente a k .

b : Índice para un lote.

N : Número de vecinos más cercanos considerados para el cálculo de LISI.

K : Número total de tipos celulares.

B : Número total de lotes.

N_k : Número de células del tipo celular k .

B_k : Número de lotes presentes en las células del tipo celular k .

$n_{b(i)}$: Número de vecinos de la célula i que pertenecen al lote b .

$n_{k(i)}$: Número de vecinos de la célula i que pertenecen al tipo celular k .

$d(i, j)$: Distancia euclidiana entre la célula i y la célula j en el espacio de expresión génica reducido.

a_i : Promedio de las distancias entre la célula i y todas las demás células del mismo tipo celular.

b_i : Mínima distancia promedio entre la célula i y todas las células de cualquier otro tipo celular.

1. **iLISI (Integration Local Inverse Simpson's Index)**: Evalúa la mezcla de lotes a nivel local. Para cada célula, examina su vecindario más cercano y calcula la diversidad de lotes presentes. Un valor alto de iLISI indica que las células de diferentes lotes se mezclan bien, lo cual es deseable después de la integración. Un valor bajo sugiere que las células de un mismo lote tienden a agruparse, indicando una integración deficiente.

$$iLISI_i = \frac{1}{\sum_b \left(\frac{n_{b(i)}}{N} \right)^2}$$

2. **norm_iLISI (Normalized iLISI):** Es simplemente el iLISI normalizado entre 0 y 1 para facilitar la comparación entre experimentos con diferente número de lotes. La normalización permite interpretar la métrica independientemente del número de lotes presentes.

$$\text{norm_iLISI}_i = \frac{\text{iLISI}_i - 1}{B - 1}$$

3. **CiLISI (Cell-type specific iLISI):** Similar al iLISI, pero calcula la mezcla de lotes dentro de cada tipo celular. Esto proporciona una visión más granular de la integración, mostrando si la mezcla de lotes es uniforme en todos los tipos celulares o si hay subpoblaciones específicas con problemas de integración.

Para una célula i del tipo celular k :

$$\text{CiLISI}_i = \frac{\text{iLISI}_i - 1}{B_k - 1}$$

Donde iLISI_i se calcula considerando solo las células del tipo celular k .

4. **CiLISI_means:** Calcula el promedio de los valores de CiLISI para todos los tipos celulares, proporcionando una métrica global de la mezcla de lotes considerando las diferencias entre tipos celulares.

$$\text{CiLISI_means} = \frac{1}{K} \sum_k \left(\frac{1}{N_k} \sum_{i \in k} \text{CiLISI}_i \right)$$

5. **norm_cLISI (Normalized cell-type LISI):** Esta métrica evalúa la separación entre los tipos celulares. A diferencia de iLISI, un valor alto de norm_cLISI es deseable, ya que indica que las células del mismo tipo celular se agrupan juntas y están bien separadas de otros tipos celulares. Un valor bajo sugiere que las células de diferentes tipos celulares se mezclan, lo que podría indicar una sobrecorrección durante la integración o una definición deficiente de los tipos celulares.

$$\text{norm_cLISI}_i = 1 - \frac{\text{cLISI}_i - 1}{K - 1}$$

Donde:

$$\text{cLISI}_i = \frac{1}{\sum_k \left(\frac{n_k(i)}{N} \right)^2}$$

6. **norm_cLISI_means:** Calcula el promedio de los valores de norm_cLISI para todos los tipos celulares, proporcionando una métrica global de la separación entre tipos celulares.

$$\text{norm_cLISI_means} = \frac{1}{K} \sum_k \left(\frac{1}{N_k} \sum_{i \in k} \text{norm_cLISI}_i \right)$$

7. **celltype_ASW (Cell-type Average Silhouette Width):** También mide la separación entre tipos celulares, pero utilizando el coeficiente de Silhouette. Para cada célula, el coeficiente de Silhouette compara la distancia media a las células de su propio tipo celular con la distancia media al

tipo celular más cercano. Un valor alto indica que la célula está bien ubicada dentro de su cluster (tipo celular), mientras que un valor bajo o negativo sugiere que la célula podría estar mal clasificada. `celltype_ASW` calcula el promedio del coeficiente de Silhouette para todas las células dentro de cada tipo celular.

Para una célula i del tipo celular k :

$$\text{celltype_ASW}_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}$$

Donde:

$$a_i = \frac{1}{N_k - 1} \sum_{j \in k, j \neq i} d(i, j)$$

$$b_i = \min_{k' \neq k} \left\{ \frac{1}{N_{k'}} \sum_{j \in k'} d(i, j) \right\}$$

8. **celltype_ASW_means:** Calcula el promedio de los valores de `celltype_ASW` para todos los tipos celulares, proporcionando una métrica global de la separación entre tipos celulares basada en la silueta de las células.

$$\text{celltype_ASW_means} = \frac{1}{K} \sum_k \left(\frac{1}{N_k} \sum_{i \in k} \text{celltype_ASW}_i \right)$$

Capítulo 4

Metodología

4.1 Estrategia

La estrategia para evaluar las métricas que miden la calidad de la integración de datos scRNA-seq fue de la siguiente forma:

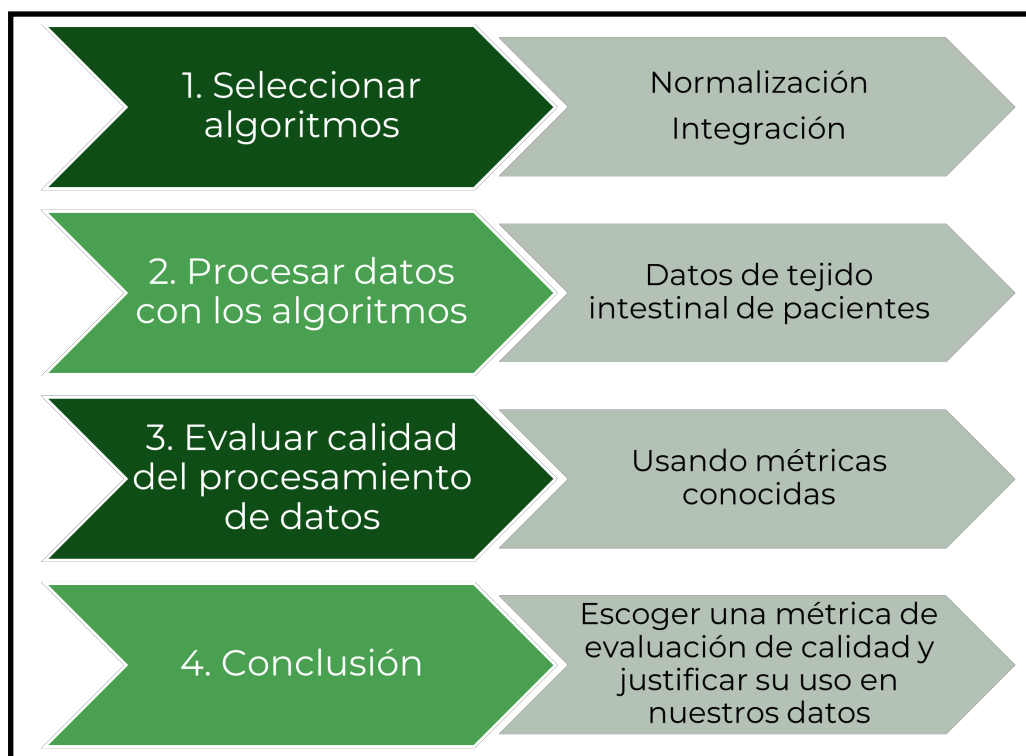


Figura 5 Metodología. Esquema representativo de las 4 fases metodológicas, explicación en el apartado de Marco teórico, sección 3.2.1

4.2 Métodos

4.2.1 Datos de scRNA-seq

Los datos de scRNA-seq se obtuvieron de muestras intestinales de dos pacientes (nombrados en este trabajo como paciente 17 y paciente 19, por protección de datos personales) del Instituto Nacional de Medicina Genómica a los que se les había realizado la técnica previamente.

4.2.2 Normalización con la biblioteca de Seurat: LogNormalize

La biblioteca Seurat es una herramienta ampliamente utilizada para el análisis de datos scRNA-seq que ofrece varios métodos de normalización, pero la que usaremos es:

LogNormalize: Ajusta los conteos de cada célula por el total de moléculas detectadas, multiplica por un factor de escala (normalmente 10,000), y luego aplica una transformación logarítmica.

$$Nx_{(i,j)} = \ln \left[\frac{(1 + x_{(i,j)}) * scale.factor}{\sum_i (1 + x_{(i,j)})} \right]$$

Donde:

i = Índice para un gen.

j = Índice para una célula.

$Nx_{(i,j)}$ = valor normalizado para el gen i de la célula j en la matriz de conteos.

\ln = Logaritmo natural.

$x_{(i,j)}$ = conteo original del gen i de la célula j en la matriz de conteos, se le suma 1 para evitar valores de 0.

$scale.factor$ = factor de escala (valor predeterminado: 10000).

$\sum_i(1 + x_{(i,j)})$: suma de los conteos de todos los genes i de la célula j .

4.2.3 Integración con el algoritmo Harmony

El algoritmo Harmony funciona proyectando los datos a un espacio de baja dimensionalidad donde las células de diferentes lotes se mezclan, al tiempo que preserva la variabilidad biológica (KORSUNSKY et al. [2019](#)). A continuación se presenta una explicación general del funcionamiento del algoritmo Harmony:

1. **Reducción de dimensionalidad:** Harmony comienza con una matriz de expresión génica (matriz de conteos) y aplica un método de reducción de dimensionalidad, como PCA o UMAP, para representar las células en un espacio de menor dimensión. Esto facilita el cálculo y la visualización.
2. **Clustering inicial:** Se realiza una agrupación inicial de las células en clústeres, generalmente utilizando el algoritmo de Louvain o Leiden. Este paso proporciona una primera aproximación de las poblaciones celulares presentes.

3. **Iteración y corrección de efectos de lote:** El algoritmo itera a través de los siguientes pasos:

Cálculo de centroides: Para cada clúster, se calcula el centroide en el espacio de baja dimensionalidad, tanto para el conjunto de datos completo como para cada lote individual.

Estimación de la corrección del lote: Harmony estima la corrección necesaria para alinear los centroides de cada lote con el centroide global del clúster. Esta corrección se basa en una transformación lineal (rotación y traslación).

Corrección de las coordenadas celulares: Las coordenadas de cada célula en el espacio de baja dimensionalidad se ajustan según la corrección del lote calculada.

Nueva agrupación en clústeres: Se realiza una nueva agrupación en clústeres con las coordenadas corregidas.

4. **Convergencia:** El proceso iterativo continúa hasta que la asignación de clústeres se estabiliza, indicando que se ha alcanzado la convergencia.

4.2.4 Criterios de selección de las métricas de evaluación de la integración

Existen muchos algoritmos para integrar los datos de scRNA-seq y evaluar esta integración, como se puede observar en el artículo titulado "**Benchmarking atlas-level data integration in single-cell genomics**" (Malte D. LUECKEN et al. [2021](#)), en el cual se hace un estudio de la mayoría de los algoritmos para

integración de datos scRNA-seq hasta el año 2021, y se hace una lista de los mejores algoritmos de integración que se seleccionaron a través de múltiples pruebas en diferentes modelos.

En este exhaustivo trabajo se determinó que el algoritmo Harmony está dentro de los 5 mejores, además era compatible con la biblioteca Seurat en R, que es la que se utiliza en el laboratorio de manera habitual, por ese motivo se seleccionó.

Con respecto a las métricas que evalúan la calidad de la integración de datos scRNA-seq, se acotó a aquellas métricas que cumplieran con lo siguiente:

- Que estuvieran ya implementadas en lenguaje R.
- Que fueran compatibles con la biblioteca Seurat.
- Que fueran compatibles con el algoritmo Harmony (algoritmo de integración de datos scRNA-seq).
- Que pudieran discriminar entre la integración de 2 subconjuntos de una matriz de conteos provenientes del mismo paciente y la integración de dos matrices de conteo provenientes de 2 pacientes distintos.
- Que los valores que calculara la métrica estuvieran acotados entre 0 y 1, a esto se le llama métrica normalizada.

4.2.5 Métricas de evaluación de la integración

Las 8 métricas seleccionadas con los criterios desarrollados en la sección anterior [4.2.4](#) fueron las siguientes:

1. **iLISI** (Integration Local Inverse Simpson's Index).
2. **norm_iLISI** (Normalized iLISI).
3. **CiLISI** (Cell-type specific iLISI).
4. **CiLISI_means**.
5. **norm_cLISI** (Normalized cell-type LISI).
6. **norm_cLISI_means**.
7. **celltype_ASW** (Cell-type Average Silhouette Width).
8. **celltype_ASW_means**.

La descripción detallada de cada métrica se encuentra en la sección [3.2.4 Métricas para evaluar la calidad de la integración de los datos de scRNA-seq](#).

4.3 Repositorio GitHub

El código en lenguaje R que se utilizó a lo largo de este proyecto está disponible en el siguiente repositorio de GitHub:

<https://github.com/osaadn/metricas>

También incluye las imágenes y los datos originales de los pacientes.

4.4 Código L^AT_EX

Disponible en:

<https://www.overleaf.com/read/krgtsqqpfsfz#e60394>

Capítulo 5

Resultados y análisis

En el semestre 2024-1 se realizaron todos los cursos necesarios para poder instalar y programar en el lenguaje R, junto con otros cursos para el entendimiento y análisis de los datos de la técnica de scRNA-seq, los cuáles fueron los siguientes:

- Curso para creación de máquinas virtuales.
- Curso de Linux Básico.
- Curso básico de programación en R.
- Curso intermedio de programación en R.
- Curso básico de análisis de datos de scRNA-seq.
- Curso intermedio de análisis de datos de scRNA-seq.
- Curso de normalización e integración de datos de scRNA-seq con la biblioteca Seurat.

Una vez terminada la fase de cursos y preparativos, en el semestre 2024-2 se realizó el código de programación en R. Se

inició generando 2 escenarios para la evaluación de las métricas de calidad de la integración.

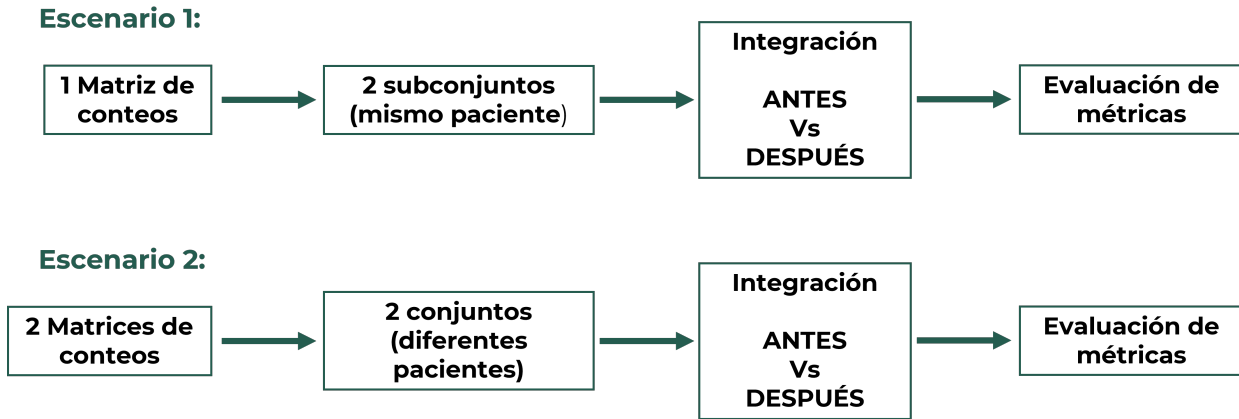


Figura 6 Escenarios para la evaluación de las métricas de integración. En el escenario 1 (arriba) se construyeron subconjuntos del mismo paciente; en el escenario 2 (abajo) se usaron matrices de conteos de dos pacientes distintos.

- **Escenario 1:** Tomar la matriz de conteos de 1 solo paciente, construir 2 subconjuntos con ella, calcular las métricas antes y después de la integración. Se espera que la integración sea muy buena porque los subconjuntos en realidad solo son una división artificial de la misma muestra que proviene del mismo individuo. Se usará como una cota superior con la que compararemos el escenario 2.
- **Escenario 2:** Tomar 2 matrices de conteo de 2 pacientes diferentes, integrar los datos y evaluar métricas. Se espera que la integración ahora no sea tan buena porque las matrices de conteo provienen de distintos pacientes y tendrán más variabilidad biológica asociada.

Se utilizaron datos de scRNA-seq de muestras de tejido intestinal de dos pacientes del Instituto Nacional de Medicina

Genómica. Estos pacientes fueron nombrados como paciente 17 (pt17) y paciente 19 (pt19) por la protección de datos personales; los resultados se muestran a continuación.

5.1 Escenario 1, paciente 17

Lo que se observa en la figura 7 es la distribución de las células obtenidas de dos subconjuntos del mismo paciente (17), sin integrar los datos, en un espacio reducido de dos dimensiones hecho mediante la técnica de reducción de dimensionalidad UMAP.

Nótese que en cada nube de puntos se pueden observar tanto puntos del subconjunto 1 pt17s1 (rojos) como del subconjunto 2 pt17s2 (verdes).

Y lo que se puede ver en la figura 8 es la distribución de las células (datos) obtenidas de dos subconjuntos del mismo paciente (las mismas de la figura 7), integrando ahora los datos con el algoritmo Harmony, en un espacio reducido de dos dimensiones hecho mediante la reducción de dimensionalidad UMAP.

A lo que hay que prestar atención es a la distribución de las células (puntos) de ambos subconjuntos pt17s1 (rojos) y pt17s2 (verdes), y podemos observar que estos se distribuyen de manera homogénea en cualquier región del plano.

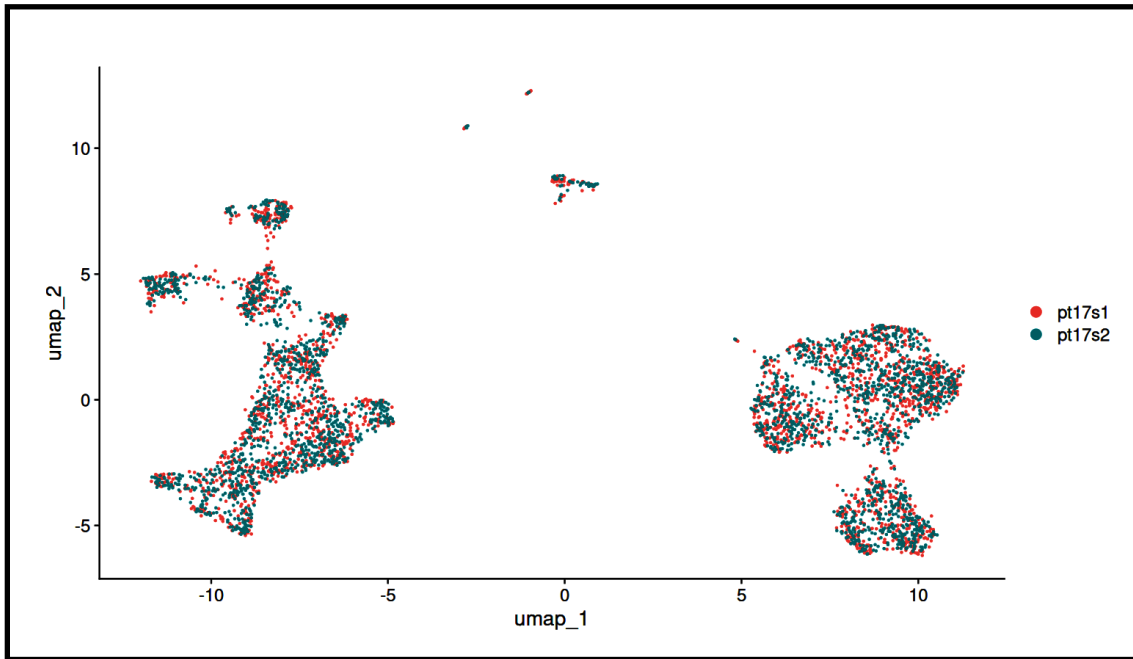


Figura 7 Escenario 1, paciente 17. Distribución de los datos ANTES de la integración. Se realizó la reducción de dimensionalidad (espacio reducido) con el método UMAP y después se graficaron las dos dimensiones principales, cada punto representa a una célula, y los distintos colores representan el origen de esos datos, no se hizo la integración de los datos. pt17s1=paciente 17 subconjunto 1, pt17s2=paciente 17 subconjunto 2.

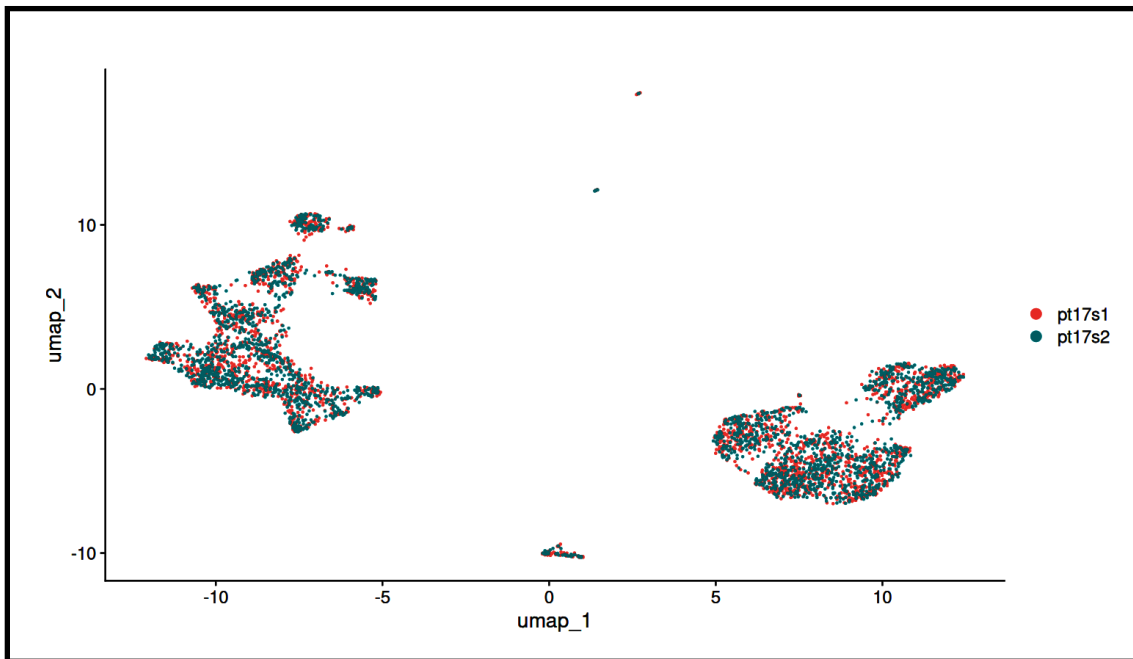


Figura 8 Escenario 1, paciente 17. Distribución de los datos DESPUÉS de la integración. Se realizó la reducción de dimensionalidad (espacio reducido) con UMAP, posteriormente se integraron los datos con el algoritmo Harmony y finalmente se graficaron las dos dimensiones principales, cada punto representa a una célula, y los distintos colores representan el origen de esos datos. pt17s1=paciente 17 subconjunto 1, pt17s2=paciente 17 subconjunto 2.

La pregunta central de este trabajo es ¿qué tan bien integrados están los datos? que es equivalente a preguntarnos si ¿la distribución de los puntos (de diferentes muestras) en el espacio reducido es homogénea?.

Estas cuestiones las podemos evaluar con las "métricas de evaluación de la integración" definidas en el apartado 4.2.5 del capítulo de Metodología. Calculando para nuestro escenario 1 con el paciente 17, tenemos lo siguiente:

Paciente 17 (pt17)	iLISI	norm iLISI	CiLISI	CiLISI means	norm cLISI	norm cLISI means	celltype ASW	celltype ASW means
ANTES	1.9297	0.9297	0.9281	0.8687	0.9479	0.8888	0.4679	0.5630
DESPUÉS	1.9383	0.9383	0.9394	0.8850	0.9540	0.9249	0.4553	0.5197
Coeficiente de variación	0.45%	0.93%	1.22%	1.88%	0.64%	4.05%	2.70%	7.69%

Tabla 2 Métricas de evaluación de la integración, subconjuntos paciente 17. Se calcularon 8 métricas (columnas) para los datos no integrados (ANTES) y para los datos integrados (DESPUÉS) con el algoritmo Harmony, y se calculó el coeficiente de variación porcentual, ver texto.

Observamos en la última fila de la tabla 2 los coeficientes de variación porcentual:

$$\left| \frac{\text{DESPUES} - \text{ANTES}}{\text{ANTES}} \right| * 100 \%$$

En nuestro caso este coeficiente se refiere a la diferencia porcentual entre no hacer la integración de los datos (ANTES) y sí hacer la integración (DESPUÉS) de los datos con el algoritmo Harmony.

Estos coeficientes tienen un valor pequeño, que en ninguna métrica evaluada rebasa el 8 % de variación, este resultado significa que realizar la integración de los datos no tiene un efecto importante en la distribución de los mismos, lo cual visualmente ya lo habíamos detectado en las figuras 7 y 8.

Aunque este resultado ya se esperaba de antemano porque los dos conjuntos de datos, el pt17s1 y el pt17s2, en realidad provienen de un solo paciente que tiene originalmente una única matriz de conteos.

5.2 Escenario 1, paciente 19

Ahora haremos el mismo análisis para el paciente 19; lo que se observa en la figura 9 es la distribución de las células (datos) obtenidas de dos subconjuntos del mismo paciente (19), sin integrar los datos, en un espacio reducido de dos dimensiones hecho mediante la técnica de reducción de dimensionalidad UMAP.

Y posteriormente, un análisis similar, pero después de la integración de los datos con el algoritmo Harmony que se puede ver en la figura 10.

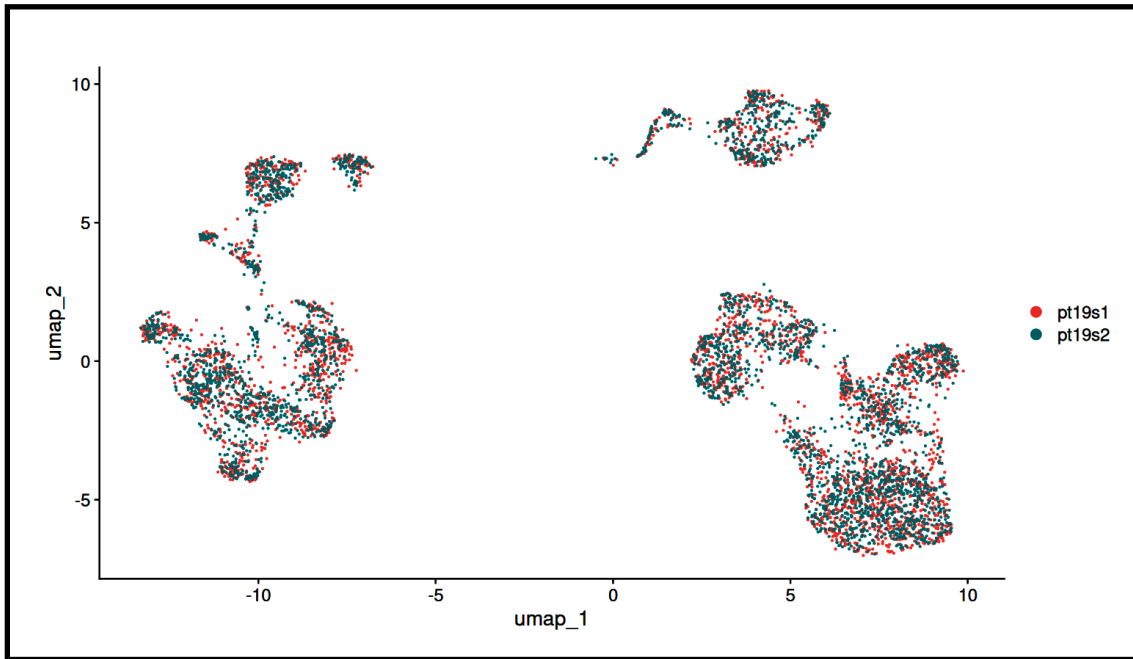


Figura 9 Escenario 1, paciente 19. Distribución de los datos ANTES de la integración. Se realizó la reducción de dimensionalidad (espacio reducido) con el método UMAP y después se graficaron las dos dimensiones principales, cada punto representa a una célula, y los distintos colores representan el origen de esos datos, no se hizo la integración de los datos. pt19s1=paciente 19 subconjunto 1, pt19s2=paciente 19 subconjunto 2.

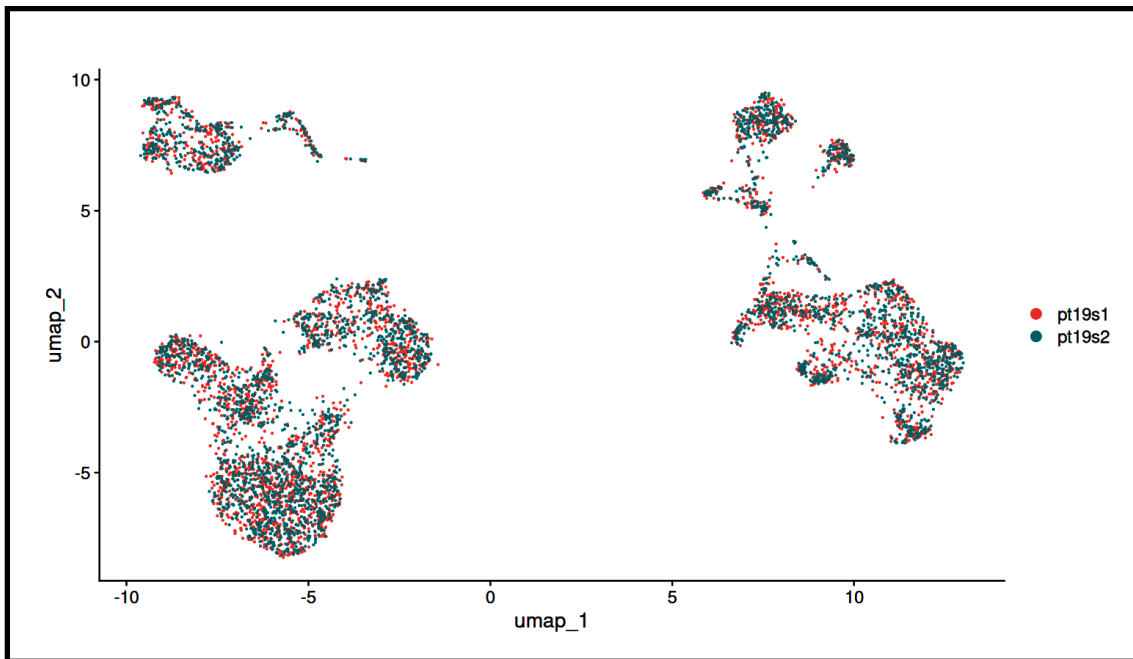


Figura 10 Escenario 1, paciente 19. Distribución de los datos DESPUÉS de la integración. Se realizó la reducción de dimensionalidad (espacio reducido) con UMAP, posteriormente se integraron los datos con el algoritmo Harmony y finalmente se graficaron las dos dimensiones principales, cada punto representa a una célula, y los distintos colores representan el origen de esos datos. pt19s1=paciente 19 subconjunto 1, pt19s2=paciente 19 subconjunto 2.

Notamos que en ambos casos, ANTES y DESPUÉS de integrar los datos con el algoritmo Harmony (figuras 9 y 10 respectivamente), los datos se distribuyen homogéneamente, lo cual también es visible en la tabla 3, donde calculamos las 8 métricas para evaluar la integración.

Paciente 19 (pt19)	iLISI	norm iLISI	CiLISI	CiLISI means	norm cLISI	norm cLISI means	celltype ASW	celltype ASW means
ANTES	1.9379	0.9379	0.9386	0.9387	0.9523	0.9442	0.3781	0.4405
DESPUÉS	1.9481	0.9481	0.9502	0.9516	0.9498	0.9380	0.3669	0.4385
Coefficiente de variación	0.53%	1.09%	1.24%	1.38%	0.27%	0.66%	2.98%	0.45%

Tabla 3 Métricas de evaluación de la integración, subconjuntos paciente 19. Se calcularon 8 métricas (columnas) para los datos no integrados (ANTES) y para los datos integrados (DESPUÉS) con el algoritmo Harmony, y se calculó el coeficiente de variación porcentual.

A diferencia de lo observado anteriormente con el paciente 17 (tabla 2), los coeficientes de variación esta vez no rebasan el 3 %, eso nos ayudó a definir un valor máximo de variación porcentual que estaríamos dispuestos a tolerar para considerar que las muestras provenían del mismo paciente, **lo establecimos en 10 %**.

5.3 Escenario 2, ambos pacientes

En este momento nos dispusimos a realizar una integración de datos conforme a lo que se hace en realidad cuando se trabaja experimentalmente con los pacientes; es decir, que no trabajamos más con subconjuntos de una única matriz de conteos

de un solo paciente, sino que lo hicimos con pacientes distintos.

Los resultados del escenario 1 (paciente 17 o 19) solo eran necesarios para determinar un coeficiente de variación y un valor máximo para cada una de las métricas de evaluación de la integración que fueron definidas en la sección 4.2.5 del capítulo de Metodología.

En la figura 11 vemos la distribución de los datos antes y después de la integración con el algoritmo Harmony.

También se puede ver en la figura 11 que los datos de cada paciente no están distribuidos homogéneamente, hay regiones donde se observa una mayoría de puntos rojos (células del paciente 17) y otras donde hay mayoría de puntos verdes (células del paciente 19) lo cual es esperado porque son dos pacientes distintos y los datos aún no se han integrado.

Cuando en la figura 12 integramos los datos, ahora podemos ver algunas regiones donde ambos conjuntos de puntos se distribuyen homogéneamente y otras regiones donde eso no sucede; esto, desde el punto de vista biológico, se puede deber a que las muestras de los pacientes no poseen los mismos tipos celulares (regiones de puntos o clústers en el gráfico).

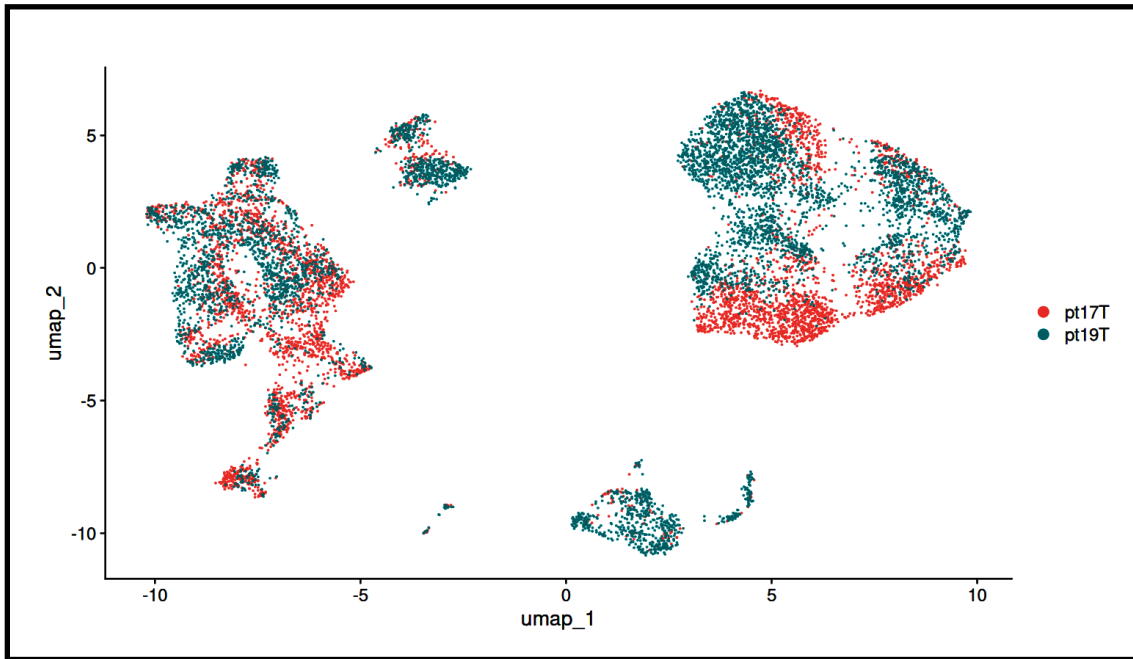


Figura 11 Escenario 2. Distribución de los datos ANTES de la integración. Se realizó la reducción de dimensionalidad (espacio reducido) con el método UMAP y después se graficaron las dos dimensiones principales, cada punto representa a una célula, y los distintos colores representan el origen de esos datos, no se hizo la integración de los datos. pt17T=paciente 17, pt19T=paciente 19.

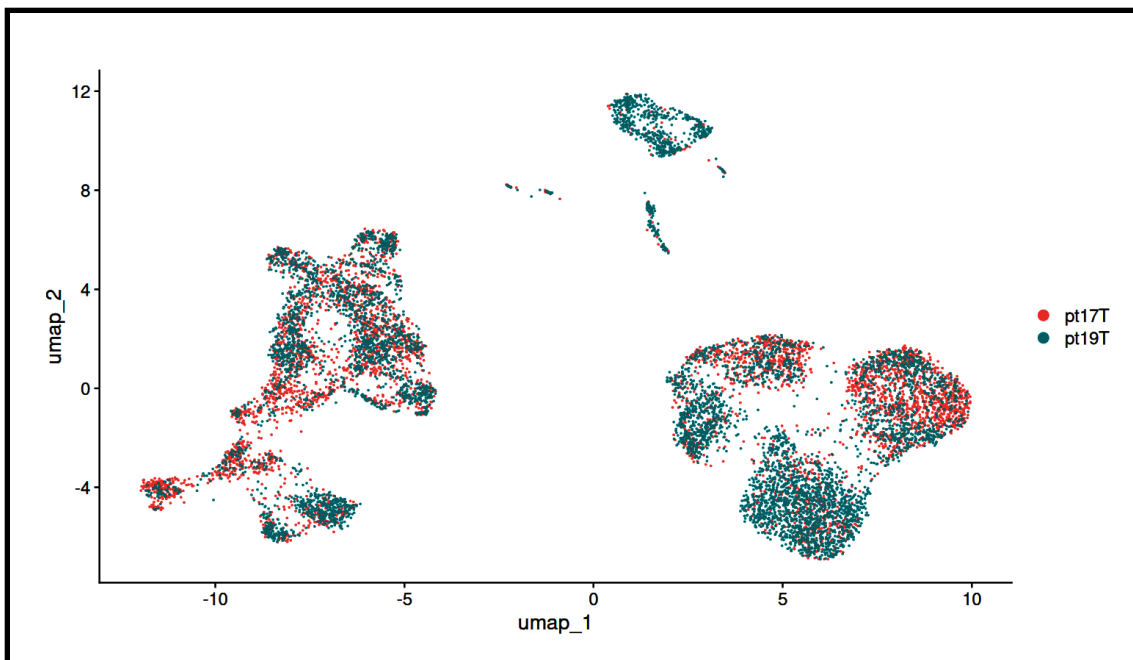


Figura 12 Escenario 2. Distribución de los datos DESPUÉS de la integración. Se realizó la reducción de dimensionalidad (espacio reducido) con UMAP, posteriormente se integraron los datos con el algoritmo Harmony y finalmente se graficaron las dos dimensiones principales, cada punto representa a una célula, y los distintos colores representan el origen de esos datos. pt17T=paciente 17, pt19T=paciente 19.

Calculamos las métricas que avalúan la integración y obtuvimos lo siguiente:

Pacientes 17 y 19	iLISI	norm iLISI	CiLISI	CiLISI means	norm cLISI	norm cLISI means	celltype ASW	celltype ASW means
ANTES	1.4488	0.4488	0.4443	0.4870	0.9503	0.9352	0.4019	0.4371
DESPUÉS	1.6888	0.6888	0.6874	0.6680	0.9512	0.9268	0.4617	0.3959
Coeficiente de variación	16.57%	53.48%	54.73%	37.16%	0.10%	0.89%	14.89%	9.42%

Tabla 4 Métricas de evaluación de la integración, ambos pacientes. Se calcularon 8 métricas (columnas) para los datos no integrados (ANTES) y para los datos integrados (DESPUÉS) con el algoritmo Harmony, y se calculó el coeficiente de variación porcentual.

Como asignamos un valor de 10 % de variación como máximo para considerar que 2 muestras provenían del mismo paciente, descartamos las métricas que tenían valores absolutos menores al 10 % a la hora que integramos los datos de ambos pacientes a la vez, las métricas ELIMINADAS fueron: norm CLISI, norm CLISI means y celltype ASW means.

Para las 5 métricas restantes calculamos los siguientes valores:

- **Paciente 17 subconjuntos:** valor de las métricas con los datos integrados de los subconjuntos del paciente 17, los mismos que en la tabla 2.
- **Paciente 19 subconjuntos:** valor de las métricas con los datos integrados de los subconjuntos del paciente 19, los mismos que en la tabla 3.

- **PROMEDIO subconjuntos:** Se calcula el promedio de "Paciente 17 subconjuntos" y "Paciente 19 subconjuntos".
- **Paciente 17 y Paciente 19:** valor de las métricas con los datos integrados de los pacientes 17 y 19, los mismos que en la tabla 4.
- **Coefficiente de variación:** variación porcentual de la DIFERENCIA y el valor de "Paciente 17 y Paciente 19".

Muestras	iLISI	norm iLISI	CiLISI	CiLISI means	celltype ASW
Paciente 17 subconjuntos	1.9383	0.9383	0.9394	0.8850	0.5197
Paciente 19 subconjuntos	1.9481	0.9481	0.9502	0.9516	0.4385
PROMEDIO subconjuntos	1.9432	0.9432	0.9448	0.9183	0.4791
Paciente 17 y Paciente 19	1.6888	0.6888	0.6874	0.6680	0.3959
Coefficiente de variación	15.07%	36.94%	37.44%	37.47%	21.01%

Tabla 5 Métricas de evaluación de la integración, selección de las mejores. Se calcularon las 5 métricas seleccionadas previamente para los datos integrados de diferentes muestras, se calculó el promedio de las métricas de las muestras de subconjuntos y el valor obtenido al integrar a ambos pacientes a la vez, para finalmente, calcular el coeficiente de variación porcentual entre ambas condiciones.

Con los datos que se muestran en la tabla 5, seleccionamos aquellas métricas con los coeficientes de variación más grandes, porque eso significa en la práctica que la métrica es capaz

de diferenciar efectivamente entre muestras tomadas del mismo paciente (subconjuntos) y muestras tomadas de dos pacientes distintos. Las 3 métricas con mayores valores fueron: norm iLISI, CiLISI y CiLISI means.

Finalmente, por conveniencia experimental, como en el futuro se compararán muchos pacientes distintos, es preferible tener una métrica normalizada, es decir que sus valores solo fluctúen entre 0 y 1. Por lo que escogimos como la mejor métrica a:

norm iLISI

Capítulo 6

Conclusiones

Se lograron ejecutar las bibliotecas (lenguaje de programación: R) necesarias para trabajar con la paquetería Seurat, que es la que se usa en el laboratorio para el análisis de datos de scRNA-seq, en un ambiente Linux.

Se seleccionó el algoritmo Harmony para la integración de datos scRNA-seq, debido a que diferentes autores mencionan que Harmony es uno de los más eficientes y rápidos que son compatibles con la paquetería Seurat.

Se encontraron 8 métricas diseñadas para evaluar la integración de datos scRNA-seq con el algoritmo Harmony. Las 8 métricas se encuentran explicadas y desarrolladas en la subsección [3.2.4](#) del capítulo de Marco teórico.

Se construyeron 2 escenarios hipotéticos para evaluar las 8 métricas:

- **Escenario 1:** Tomar la matriz de conteo de 1 solo paciente, construir 2 subconjuntos con ella, calcular las métricas antes y después de integrar los datos y evaluar.

- **Escenario 2:** Tomar 2 matrices de conteo de 2 pacientes diferentes, calcular las métricas antes y después de integrar los datos y evaluar.

Para el escenario 1, usando a cualquiera de los pacientes, se observó lo esperado, las métricas calculadas ANTES o DESPUÉS de la integración tenían valores muy similares, el coeficiente de variación porcentual de todas las métricas en este escenario siempre fue menor al 8 %, ver tablas 2 y 3 para los pacientes 17 y 19 respectivamente.

En el escenario 1, los datos con los que se calculan las métricas provienen del mismo paciente; es decir, los datos son subconjuntos de la misma matriz de conteos original, por lo cual hacer o no la integración de datos debería ser irrelevante, y los coeficientes de variación calculados pueden ser considerados como el **error del algoritmo Harmony** al integrar datos que por su naturaleza están bien integrados desde el principio.

Con respecto al escenario 2, al calcular las métricas ANTES y DESPUÉS de la integración, nos percatamos de que había 3 métricas cuyos coeficiente de variación porcentual eran menores que el **error del algoritmo Harmony** del 8 % (ver tabla 4), por lo que estas métricas fueron eliminadas del análisis, la interpretación es que son incapaces de discriminar entre el escenario 1 y el escenario 2.

Finalmente, calculamos un último coeficiente de variación, que-
ríamos comparar las métricas promedio obtenidas en el esce-

nario 1 y las métricas en el escenario 2, ambas después de la integración de datos, para escoger la métrica que tuviera el mayor coeficiente de variación entre estos dos escenarios, logrando nuestro objetivo general.

Por lo que pudimos escoger una métrica que cumplía los siguientes criterios:

- ✓ Valores altos cuando se mide la integración de datos que provienen de subconjuntos del mismo paciente.
- ✓ Capaz de diferenciar entre la integración de subconjuntos del mismo paciente y pacientes diferentes.
- ✓ Variaciones mayores al 8 % para la integración de datos de pacientes distintos.
- ✓ Métrica normalizada para tener valores entre 0 y 1.

Y concluimos, **cumpliendo así con nuestra pregunta de investigación**, que la mejor métrica para evaluar la calidad de la integración de datos scRNA-seq en nuestros datos es:

norm iLISI

6.1 Limitaciones

- Se necesitan hacer más pruebas con distintos pacientes para determinar el error del algoritmo Harmony.

- Al usar subconjuntos de una única matriz de conteos, se obtienen los valores máximos de las métricas; es decir, una cota superior, pero no tenemos una cota inferior, se podrían utilizar matrices de conteos artificiales o provenientes de muestras de diferentes animales, para determinar la peor integración posible.
- Definir una región de decisión para calificar si un conjunto de datos están bien integrado o no requiere más pruebas con diferentes escenarios previamente probados por otros investigadores y que ya se consideren en la teoría como bien integrados.

6.2 Perspectivas

- Evaluar las métricas usando escenarios realizados con más de 2 subconjuntos, para determinar ¿cuál es el tamaño mínimo de una matriz de conteos donde todavía se puede hacer una buena integración?
- Evaluar las métricas en regiones delimitadas del espacio reducido, bajo el entendido de que cada región en el espacio reducido se debe corresponder, biológicamente, con un tipo celular específico.
- Explorar las aplicaciones de los grupos algebraicos en matrices de conteo es una perspectiva prometedora. Las operaciones dentro de un grupo podrían modelar transformaciones entre estados celulares, como transiciones entre subpoblaciones. Esto podría facilitar la identificación de es-

estructuras invariantes en los datos, útiles para definir clústeres celulares de manera más precisa o para proponer nuevas métricas que aprovechen estas propiedades matemáticas.

- Analizar si las métricas utilizadas pueden agruparse en categorías basadas en las propiedades del espacio métrico que inducen sobre los datos. Por ejemplo, métricas como `norm_iLISI` que se centran en la mezcla de lotes podrían clasificarse como métricas de homogeneidad, mientras que otras como `celltype_ASW` podrían pertenecer a métricas de separación. Además, explorar si estos espacios métricos cumplen propiedades como completitud, convexidad o continuidad podría ofrecer nuevas herramientas para seleccionar métricas según los objetivos biológicos del análisis.

6.3 Aportación

Muchos de los investigadores que realizan análisis bioinformáticos con datos de scRNA-seq, siguen flujos de trabajo de manera automatizada y definida por terceras personas, sin reparar en la calidad de la información utilizada.

Sin importar lo sofisticados que puedan ser estos análisis de datos scRNA-seq, si los datos originalmente no están bien integrados, no resultarían comparables, y los resultados de estos análisis serían dudosos.

Este trabajo ayuda a cuantificar la integración de los datos de scRNA-seq, proporcionando un valor numérico que mide la calidad de la integración de los datos en una escala fácilmente comprensible entre 0 y 1, este valor se corresponde con asegurar que los pacientes elegidos pertenecen a la misma población, la que fue previamente definida por criterios de inclusión y exclusión de carácter médico.

Referencias

- ANDREATTA, Massimo, Léonard HÉRAULT, Paul GUEGUEN, David GFELLER, Ariel J. BERENSTEIN y Santiago J. CARMONA (ene. de 2024). «Semi-supervised integration of single-cell transcriptomics data». En: *Nature Communications* 2024 15:1 15 (1), págs. 1-13. ISSN: 2041-1723. DOI: [10.1038/s41467-024-45240-z](https://doi.org/10.1038/s41467-024-45240-z). URL: <https://www.nature.com/articles/s41467-024-45240-z>.
- ANDREWS, Tallulah S., Vladimir Yu KISELEV, Davis MCCARTHY y Martin HEMBERG (dic. de 2020). «Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data». En: *Nature Protocols* 2020 16:1 16 (1), págs. 1-9. ISSN: 1750-2799. DOI: [10.1038/s41596-020-00409-w](https://doi.org/10.1038/s41596-020-00409-w). URL: <https://www.nature.com/articles/s41596-020-00409-w>.
- ELMENTAITE, Rasa, Cecilia Domínguez CONDE, Lu YANG y Sarah A. TEICHMANN (jul. de 2022). «Single-cell atlases: shared and tissue-specific cell types across human organs». En: *Nature Reviews Genetics* 23 (7), págs. 395-410. ISSN: 1471-0056. DOI: [10.1038/s41576-022-00449-w](https://doi.org/10.1038/s41576-022-00449-w). URL: <https://www.nature.com/articles/s41576-022-00449-w>.
- FORCATO, Mattia, Oriana ROMANO y Silvio BICCIATO (ene. de 2021). «Computational methods for the integrative analysis of single-cell data». En: *Briefings in bioinformatics* 22 (1), págs. 20-29. ISSN: 1477-4054. DOI: [10.1093/BIB/BBAA042](https://doi.org/10.1093/BIB/BBAA042). URL: <https://pubmed.ncbi.nlm.nih.gov/32363378/>.
- HAQUE, Ashraf, Jessica ENGEL, Sarah A. TEICHMANN y Tapio LÖNNBERG (ago. de 2017). «A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications». En: *Genome Medicine* 9 (1), págs. 1-12. ISSN: 1756994X. DOI: [10.1186/S13073-017-0467-4](https://doi.org/10.1186/S13073-017-0467-4). URL: <https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-017-0467-4>.
- KORSUNSKY, Ilya, Nghia MILLARD, Jean FAN, Kamil SLOWIKOWSKI, Fan ZHANG, Kevin WEI, Yuriy BAGLAENKO, Michael BRENNER, Po ru LOH y Soumya RAYCHAUDHURI (nov. de 2019). «Fast, sensitive and accurate integration of single-cell data with Harmony». En: *Nature Methods* 2019 16:12 16 (12), págs. 1289-1296. ISSN: 1548-7105. DOI: [10.1038/s41592-019-0619-0](https://doi.org/10.1038/s41592-019-0619-0). URL: <https://www.nature.com/articles/s41592-019-0619-0>.
- KOTLOV, Nikita, Kirill SHAPOSHNIKOV, Cagdas TAZEARSAN, Madison CHASSE, Artur BAI-SANGUROV, Svetlana PODSVIROVA, Dawn FERNANDEZ, Mary ABDU, Leznath KANEUNYENYE, Kelley MORGAN, Ilya CHEREMUSHKIN, Pavel ZEMSKIY, Maxim CHELUSHKIN, Maria SOROKINA, Ekaterina BELOVA, Svetlana KHORKOVA, Yaroslav LOZINSKY, Katerina NUZH-DINA, Elena VASILEVA, Dmitry KRAVCHENKO, Kushal SURYAMOHAN, Krystle NOMIE, John CURRAN, Nathan FOWLER y Alexander BAGAEV (mar. de 2024). «Procrustes is

- a machine-learning approach that removes cross-platform batch effects from clinical RNA sequencing data». En: *Communications Biology* 2024 7:1 7 (1), págs. 1-14. ISSN: 2399-3642. DOI: [10.1038/s42003-024-06020-z](https://doi.org/10.1038/s42003-024-06020-z). URL: <https://www.nature.com/articles/s42003-024-06020-z>.
- LU, Junru, Yuqi SHENG, Weiheng QIAN, Min PAN, Xiangwei ZHAO y Qinyu GE (mayo de 2023). «scRNA-seq data analysis method to improve analysis performance». En: *IET Nanobiotechnology* 17 (3), págs. 246-256. ISSN: 1751-875X. DOI: [10.1049/NBT2.12115](https://doi.org/10.1049/NBT2.12115). URL: <https://onlinelibrary.wiley.com/doi/full/10.1049/nbt2.12115> %20https://onlinelibrary.wiley.com/doi/abs/10.1049/nbt2.12115%20https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/nbt2.12115.
- LUECKEN, Malte D y Fabian J THEIS (jun. de 2019). «Current best practices in single-cell RNA-seq analysis: a tutorial». En: *Molecular Systems Biology* 15 (6), pág. 8746. ISSN: 1744-4292. DOI: [10.15252/MSB.20188746](https://doi.org/10.15252/MSB.20188746) / SUPPL _ FILE / MSB188746 - SUP - 0003 - DATASETEV1.ZIP. URL: <https://www.embopress.org/doi/10.15252/msb.20188746>.
- LUECKEN, Malte D., M. BÜTTNER, K. CHAICHOOMPU, A. DANESE, M. INTERLANDI, M. F. MUELLER, D. C. STROBL, L. ZAPPIA, M. DUGAS, M. COLOMÉ-TATCHÉ y Fabian J. THEIS (dic. de 2021). «Benchmarking atlas-level data integration in single-cell genomics». En: *Nature Methods* 2021 19:1 19 (1), págs. 41-50. ISSN: 1548-7105. DOI: [10.1038/s41592-021-01336-8](https://doi.org/10.1038/s41592-021-01336-8). URL: <https://www.nature.com/articles/s41592-021-01336-8>.
- LYTAL, Nicholas, Di RAN y Lingling AN (feb. de 2020). «Normalization Methods on Single-Cell RNA-seq Data: An Empirical Survey». En: *Frontiers in Genetics* 11, pág. 501166. ISSN: 16648021. DOI: [10.3389/FGENE.2020.00041](https://doi.org/10.3389/FGENE.2020.00041) / BIBTEX. URL: <https://github.com/shka/R-SAMstrt/archive/0.99.0.tar.gz>.
- MA, Ying, Shiquan SUN, Xuequn SHANG, Evan T. KELLER, Mengjie CHEN y Xiang ZHOU (mar. de 2020). «Integrative differential expression and gene set enrichment analysis using summary statistics for scRNA-seq studies». En: *Nature Communications* 2020 11:1 11 (1), págs. 1-13. ISSN: 2041-1723. DOI: [10.1038/s41467-020-15298-6](https://doi.org/10.1038/s41467-020-15298-6). URL: <https://www.nature.com/articles/s41467-020-15298-6>.
- RYU, Yeonjae, Geun Hee HAN, Eunsoo JUNG y Daehee HWANG (feb. de 2023). *Integration of Single-Cell RNA-Seq Datasets: A Review of Computational Methods*. DOI: [10.14348/molcells.2023.0009](https://doi.org/10.14348/molcells.2023.0009).
- WISDOM, Amy J., Yvonne M. MOWERY, Cierra S. HONG, Jonathon E. HIMES, Barzin Y. NABET, Xiaodi QIN, Dadong ZHANG, Lan CHEN, Hélène FRADIN, Rutulkumar PATEL, Alex M. BASSIL, Eric S. MUISE, Daniel A. KING, Eric S. XU, David J. CARPENTER, Collin L. KENT, Kimberly S. SMYTHE, Nerissa T. WILLIAMS, Lixia LUO, Yan MA, Ash A. ALIZADEH, Kouros OWZAR, Maximilian DIEHN, Todd BRADLEY y David G. KIRSCH (dic. de 2020). «Single cell analysis reveals distinct immune landscapes in transplant and primary sarcomas that determine response or resistance to immunotherapy». En: *Nature Communications* 11 (1), pág. 6410. ISSN: 2041-1723. DOI: [10.1038/s41467-020-19917-0](https://doi.org/10.1038/s41467-020-19917-0). URL: <https://www.nature.com/articles/s41467-020-19917-0>.

Índice de figuras

1. Matriz de conteos.	6
2. Pasos principales en la técnica de scRNA-seq.	14
3. Diferentes algoritmos de normalización de datos de scRNA-seq.	18
4. Integración de datos de scRNA-seq.	20
5. Metodología.	25
6. Escenarios para la evaluación de las métricas.	33
7. Escenario 1, paciente 17. ANTES de la integración.	35
8. Escenario 1, paciente 17. DESPUÉS de la integración.	35
9. Escenario 1, paciente 19. ANTES de la integración.	38
10. Escenario 1, paciente 19. DESPUÉS de la integración.	38
11. Escenario 2, ambos pacientes. ANTES de la integración.	41
12. Escenario 2, ambos pacientes. DESPUÉS de la integración.	41

Índice de tablas

1. Fuentes de variación en los datos de scRNA-seq. . . .	15
2. Métricas, subconjuntos paciente 17.	36
3. Métricas, subconjuntos paciente 19.	39
4. Métricas, ambos pacientes.	42
5. Métricas, selección de las mejores.	43