Universitat Oberta de Catalunya

uoc.edu

# Proporciones celulares de las lesiones precursoras del cáncer gástrico definidas por deconvolución de transcriptómica en tejido y transcriptómica *single-cell*.

Sergio Lario García

Màster en Bioinformàtica i Bioestadística

Subárea 5 epigenética y cáncer

Consultora:

Izaskun Mallona González

Profesora responsable de la assignatura:

Laura Calvet Liñan

Fecha Entrega: 2 de junio de 2022.

FICHA DEL TRABAJO FINAL

| | |
|---|---|
| Título del trabajo: | Proporciones celulares de las lesiones precursoras del cáncer gástrico definidas por deconvolución de transcriptómica en tejido y transcriptómica single-cell. |
| Nombre del autor: | Sergio Lario García |
| Nombre del consultor/a: | Izaskun Mallona González |
| Nombre del PRA: | Laura Calvet Liñan |
| Fecha de entrega (mm/aaaa): | 06/2022 |
| Titulación: | Máster universitario en Bioinformática y bioestadística UOC-UB |
| Área del Trabajo Final: | ÁREA 3, Subárea 5 epigenética y cáncer |
| Idioma del trabajo: | Castellano |
| Número de créditos: | 15 |
| Palabras clave | Deconvolución, single cell-RNAseq, lesiones precursoras del cáncer gástrico-. |

**Resumen del Trabajo (máximo 250 palabras):** Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.

Las lesiones precursoras del cáncer gástrico (LPCG) preceden al desarrollo de los adenocarcinomas. En el estómago, las LPCG se caracterizan por la aparición de metaplasia intestinal, en la que hay una sustitución de tipos celulares. El estudio del transcriptoma en biopsias gástricas presenta el inconveniente de su interpretación. Un gen desregulado podría estarlo porque está desregulado en algún tipo celular concreto, o porque se ha alterado el número de células que lo expresan. Aquí, usando la deconvolución, determinamos las proporciones celulares de un estudio transcriptómico de pacientes con LPGC. Las proporciones celulares, además de ser un factor de confusión, podrían tener relevancia en cuanto a la evolución de las LPGC.

A partir de experimentos de single-cell RNAseq usamos Seurat para extraer la matriz de identidades celulares. Esta matriz se extrajo después de reanalizar los experimentos de *single-cell* y fue la entrada para algoritmo de deconvolución (CIBERSORTx).

Hemos comprobado que en las LPCG existe un aumento de células metaplásicas intestinales, concretamente enterocitos y células caliciformes. Además, hemos visto que los enterocitos se pueden asociar a cuatro *clusters* con marcadores diferentes. Por último, el aumento de células intestinales se acompaña de la desaparición de células D, secretoras de somatostatina.

Concluimos que es viable la determinación de proporciones celulares a partir de datos -ómicos.

Abstract (in English, 250 words or less):

Precursor lesions of gastric cancer (PLGC) precede the development of adenocarcinomas. In the stomach, PLGCs are characterized by the presence of intestinal metaplasia, in which one type of differentiated cell is substituted by another type. One disadvantage of transcriptomic analysis in gastric biopsies is its interpretation. An altered gene could be deregulated because it is actually deregulated in a particular cell type, or because the number of cells expressing it has been altered. Here, using deconvolution, we determined the cell proportions from a transcriptomic study of patients with LPGC. Cell proportions, in addition to being a confounding factor for transcriptomic studies, could have relevance in terms of the progression of LPGCs.

Using single-cell RNAseq experiments and Seurat we extracted the matrix of cell signatures. Importantly, the signature matrix was obtained after reanalyzing the single-cell experiments. The signature matrix was used as the input for the deconvolution algorithm CIBERSORTx.

We have found that in LPCGs there is an increase of intestinal metaplastic cells, namely enterocytes and goblet cells. In addition, we show that enterocytes can be associated with four independent clusters with different markers. Finally, the increase in intestinal cells is accompanied by the loss of somatostatin-secreting D cells. We conclude that it is feasible to determine cellular proportions from -omics data.

# Contenido

Índice de figuras y Tablas

# 1 Resumen

Las lesiones precursoras del cáncer gástrico (LPCG) preceden al desarrollo de los adenocarcinomas y su presencia se asocia a un mayor riesgo de padecer la enfermedad. En el estómago, las LPCG se caracterizan por la aparición de metaplasia intestinal. En la metaplasia intestinal hay una sustitución de tipos celulares (de secretores a absortivos). Los estudios transcriptómicos han identificado muchos de los genes diferencialmente expresados en estos procesos. En humanos, el estudio del transcriptoma en biopsias gástricas tiene la ventaja que se analizan miles de genes, pero presenta el inconveniente de su interpretación. Un gen desregulado podría estarlo porque está desregulado en algún tipo celular concreto, o porque se ha alterado el número de células que lo expresan. En el presente trabajo, usando la deconvolución, determinaremos las proporciones celulares en un dataset obtenido de un estudio transcriptómico (microarrays) de pacientes con LPGC. El análisis nos permitirá conocer cuáles son las proporciones de los tipos celulares presentes en estas lesiones. Tiene un interés múltiple, ya que las proporciones celulares son un factor de confusión en los estudios de expresión génica, pero también podrían tener relevancia en cuanto a la evolución de las LPGC.

A partir de experimentos de single-cell RNAseq de biopsias de pacientes con lesiones precursoras, usamos Seurat para extraer la matriz de identidades celulares. La matriz de identidades se extrajo después de reanalizar los experimentos de *single-cell*. Esta matriz de identidades fue la entrada para algoritmo de deconvolución (CIBERSORTx) para identificar los tipos celulares de otro experimento, con mayor número de pacientes con LPCG, realizado con microarrays.

Usando esta metodología, hemos comprobado que existe un aumento de células metaplásicas intestinales, concretamente enterocitos y células caliciformes. Además, hemos visto que los enterocitos se pueden asociar a cuatro *clusters* con marcadores diferentes, con significado incierto. Por último, el aumento de células intestinales se acompaña de la desaparición de células D, secretoras de somatostatina.

Concluimos que es viable la determinación de proporciones celulares a partir de datos -ómicos. Finalmente, se discuten los métodos disponibles, las ventajas y limitaciones de la metodología utilizada y se sugieren varios escenarios de investigación futura.

# 2 Introducción

## 2.1 Contexto y justificación del Trabajo

El presente TFM se enmarca en la Subárea 5 sobre epigenética y cáncer. Uno de los puntos de interés en esta subárea es el análisis de datos generados por tecnologías *single-cell*. Éste está siendo un ámbito que está creciendo muy rápidamente desde su primera publicación en 2009[1]. Esto es porque tiene el potencial ayudar a responder cuestiones clave de la biología celular de las que hasta el momento se tenían respuestas parciales. Estas cuestiones son: la clasificación celular, el linaje y diferenciación celular, las interacciones celulares en los tejidos, las variaciones fenotípicas de una estirpe celular o discernir qué relación tienen estos cambios fenotípicos con las enfermedades[2].

Todos estos puntos están asimismo relacionados con las lesiones precursoras del cáncer (LPC). Las LPC preceden al desarrollo de los adenocarcinomas y su presencia se asocia a un mayor riesgo de cáncer[3]. Afortunadamente no todas las LPC progresan, en la mayoría de casos permanecen estables o incluso regresan. En este escenario, las tecnologías *single-cell* pueden ayudar a: *i)* comprender estos procesos biológicos, *ii)* estratificar a los pacientes en función del riesgo de progresión, y *iii)* diseñar estrategias de prevención.

En el presente TFM aprovecharemos dos datasets[4,5] obtenidos por técnicas transcriptómicas en muestras de pacientes con lesiones precursoras de cáncer gástrico (LPCG). El análisis nos permitirá conocer cuáles son las proporciones de los tipos celulares presentes en estas LPCG. Tiene un interés múltiple, ya que las proporciones celulares son un factor de confusión en los estudios de expresión génica y podrían tener relevancia en cuanto a la evolución de las LPGC.

## 2.2 Objetivos del Trabajo

Objetivo general:

Determinar las proporciones celulares de un estudio transcriptómico en microarrays realizado en biopsias de pacientes en las fases iniciales de la cascada de Correa.

Los objetivos específicos son los siguientes:

- Obtener la matriz de firmas génicas a partir de un estudio de transcriptómica *single-cell* realizado en muestras de mucosa gástrica[5] .
- Por medio de una técnica de deconvolución, y usando la matriz de firmas génicas del objetivo 1, determinar las proporciones celulares de la mucosa gástrica de pacientes con y sin LPCG de un estudio transcriptómico realizado en microarrays[4].
- Con las proporciones celulares derivadas del objetivo 2, realizar un análisis descriptivo de las proporciones celulares durante la cascada de Correa (NAG, CAG, AT, IM).

## 2.3 Enfoque y método seguido

Existen más de 50 métodos de deconvolución. Entran en dos categorías[6]:

*Supervisados:* usan un conocimiento previo en forma de proporciones celulares o de matrices de firmas génicas. La deconvolución se basa comúnmente en *linear least squares* (LLS) o en *support vector regression* (SVR).

*No supervisados:* tratan el problema como una reduccionalidad de dimensiones por *principal component analysis* (PCA), o bien por *non-negative matrix factorization* (NNMF) o modelos Bayesianos.

Los métodos supervisados generalmente presentan mejores valores de rendimiento[6]. Además, aprovechamos que disponemos en la bibliografía y en las bases de datos públicas (GEO) de estudios transcriptómicos de célula única de las LPCG[5] y del adenocarcinoma gástrico[7] para obtener las matrices de firmas génicas.

Consideramos, por tanto, que un método supervisado que utilice la matriz de firmas obtenida del estudio *single-cell* es la metodología más apropiada para alcanzar el objetivo principal. Se propone el uso conjunto de Seurat para el análisis *single-cell* y CIBERSORTx para la deconvolución.

### 2.4 Planificación del Trabajo

Se recoge a continuación las tareas ideadas para completar los objetivos. Los hitos corresponden a resultados parciales o finales en forma de tablas, resúmenes o interpretación. El Anexo 1 muestra la información en el tiempo, según los plazos de las PEC, en formato de diagrama de Gantt. (El diagrama de Gantt también incluye la planificación de PEC-1, -4 y -5).

Para el objetivo 1, (*PEC2*):

1. Familiarizarse con el paquete Seurat por seguimiento de su tutorial[8].
2. Obtención programática de los datos *single-cell* de GSE134520 con el paquete GEOquery.

Usando *Seurat*:

3. Realizar un control de calidad de las células y filtrado de las células que no cumplan criterios.
4. Normalización de los datos.
5. *Feature selection*: identificación de los genes (*features*) con alta variabilidad.
6. Transformación por escalado (media= 0, varianza=1).
7. Reducción lineal de la dimensionalidad por PCA usando los genes de *Feature selection* y selección del número de componentes principales. **HITO**
8. Agrupamiento (*cluster*) de las células y determinación de la resolución.
9. Reducción no lineal de la dimensionalidad (UMAP/tSNE).
10. Identificación de los genes diferencialmente expresados.
11. Identificación de los *clusters* por según la información de los autores y bibliografía.
12. Exportación de la tabla con las identidades celulares y sus marcadores. **HITO**
13. Interpretación de los resultados obtenidos. **HITO**

Para el objetivo 2, (*PEC3*):

14. Familiarizarse con algún paquete de deconvolución (AntigeneS, EPIC, CIBERSORTx).
15. Obtención programática de los datos de microarray de E-MTAB-8889 con el paquete ArrayExpress.
16. Normalización y filtrado de los datos.
17. Rstudio: Generación de la matriz *single-cell* para corrección *batch* tipo S.
18. CIBERSORTx: análisis de proporciones celulares usando las matrices del paso anterior. **HITO**

Para el objetivo general, (*PEC3*):

19. Importación en Rstudio de las tablas generadas por CIBERTSORTx.
20. Estadística descriptiva de los tipos celulares en cada paso de la cascada de Correa. **HITO**

Desde su concepción inicial se ha modificado la planificación. Especialmente para el objetivo 2, la deconvolución. Además, y con tal de limitar la extensión de resultados, la estadística referente a los tipos celulares obtenidos en relación a la cascada carcinogénica se ha limitado al estudio de algunos de ellos.

Se adjunta el diagrama de Gantt como figura en el Anexo 1. Este diagrama se ha creado con versión gratuita de TeamGantt.

### 2.5 Breve sumario de contribuciones y productos obtenidos

El producto se traduce en la presente memoria y en una serie de scripts en R que están disponibles en los anexos y, bajo solicitud, en github (https://github.com/lariosergio/TFM_scGastric).

### 2.6 Breve descripción de los otros capítulos de la memoria

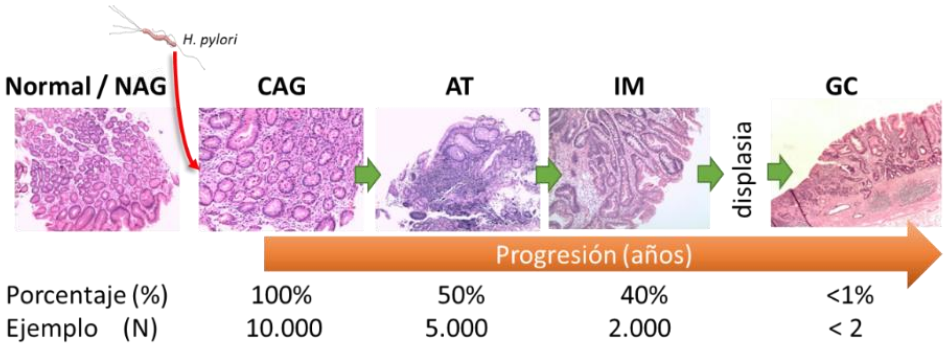
El capítulo “Material y métodos” se describen los datasets y las herramientas informáticas que se han utilizado. En “Resultados” se explican en detalle los resultados obtenidos para cada análisis, mientras que en el capítulo “Discusión”, se resumen los resultados más destacables. Para terminar, “Conclusiones” recoge el cumplimiento del proyecto a los objetivos planteados y las modificaciones de planificación.

# 3 Estado del arte

## 3.1 Introducción

El 89% de los casos de cáncer gástrico distal (CG) son atribuibles a la infección crónica por *Helicobacter pylori*[9]. El CG es la etapa final de un proceso inflamatorio que perdura durante décadas y que se caracteriza por una secuencia de lesiones conocidas como la cascada de Correa[10]. Esta cascada es un proceso carcinogénico con varias etapas: mucosa gástrica normal o con gastritis no activa (*non-atrophic gastritis*, NAG) que, después de la infección por *H. pylori*, progresa a hacia gastritis crónica activa (*chronic-active gastritis*, CAG), atrofia (AT), metaplasia intestinal (MI), displasia y CG. La AT y la MI se consideran LPCG ya que hacen aumentar progresivamente el riesgo de padecer CG[11]. Los determinantes de la aparición de les LPCG son múltiples y complejos[9]. El conocimiento biológico de les LPCG podría ayudar a reducir la incidencia de malignidad.

*Figura 1. La cascada de Correa*




## 3.2 Problemática

Los estudios transcriptómicos de la cascada de Correa han proporcionado mucha información sobre los genes diferencialmente expresados (GDE) en cada una de las etapas[4,12,13]. Hasta muy recientemente, la mayoría de estos estudios se han realizado sobre biopsias endoscópicas. Esto tiene un inconveniente de interpretación. Si observamos que un gen está desregulado, esta desregulación se puede dar porque:

- el gen está desregulado en algún tipo celular concreto.
- se ha alterado el número de células que expresan ese gen (por muerte celular, proliferación o transdiferenciación).

A lo largo de la cascada de Correa, la mucosa gástrica se va atrofiando por la pérdida de las células parietales y la aparición de un linaje que no aparece en una mucosa gástrica sana. Este linaje tiene fenotipo de epitelio, pero no gástrico, sino similar al del intestino. Esta mucosa intestinal, histológicamente, recibe el nombre de metaplasia intestinal. Aunque en debate sobre su origen[14,15], la metaplasia intestinal podría aparecer por transdiferenciación de las células principales gástricas o por diferenciación de las células madre de la foseta gástrica.

Estos cambios celulares de la mucosa muestran la importancia de establecer las proporciones celulares en los estudios transcriptómicos gástricos. Algunos autores[16], para asegurarse que la expresión diferencial se debe al primer punto, han realizado microdisección de la mucosa gástrica para describir con detalle la expresión diferencial de las áreas metaplásicas. Este tipo de estudios tienen limitaciones, el principal es la estabilidad del RNA[17] pero, además, capturar zonas donde las células parecen microscópicamente uniformes no asegura su homogeneidad.

Existe una complicación adicional. Las biopsias que evalúa el patólogo para el diagnóstico son diferentes de las biopsias utilizadas para los análisis moleculares. Esto es debido a que las muestras que se recogen durante el transcurso de la gastroscopia son muy pequeñas (~5 mg) y a que, las LPCG se muestran generalmente parcheadas sobre la superficie del estómago[11]. Esto genera la incertidumbre de si lo que se ha analizado por transcriptómica es lo que ha diagnosticado el patólogo.

### 3.3 Solución propuesta

Esta problemática en los estudios transcriptómicos gástricos tiene una posible solución computacional llamada deconvolución. La deconvolución es conjunto de métodos matemáticos que permiten inferir las proporciones celulares a partir de datos de expresión génica obtenidos a partir de tejidos. No hemos encontrado en la bibliografía referencias a este tipo de análisis en muestras de pacientes a lo largo de la cascada de Correa.

### 3.4 Deconvolución

La deconvolución se aplica a varias disciplinas. Los primeros en plantearla en el ámbito de la transcriptómica fueron Venet *et al*.[18] hace veinte años y, actualmente, disponemos de numerosas metodologías[6,19,20].

Brevemente, la solución se plantea como un producto de matrices[6]:

$$T_{(tissue)} = C_{(cell\ sign.)} \cdot P_{(proportions)}$$

Donde:

*T* es una matriz que contiene la expresión de *N* genes en *M* muestras (*bulk*). *C* es la matriz de identidades (*signature matrix*), que está constituida por *N* genes de *k* tipos celulares. Por último, *P* contiene las proporciones de los tipos celulares *k* en cada una de las muestras *M*.

O como un modelo de variables latentes[6]:

$$t_{ij} = \sum_{k=1}^{k} c_{ik} \ . p_{ki} + \ e_{ij} \ ; \ \ i = 1 \dots M \ y \ j = 1 \dots N$$

Donde:

$t_{ij}$ es la expresión del gen *i* en la muestra *j* (mezcla observada, *bulk*).

$c_{ij}$ es la expresión media del gen *i* en el tipo celular *k.*

$p_{kj}$ es la proporción del tipo celular *k* enla muestra *j*.

$e_{ij}$ es el error.

*K* = tipos celulares; N= núm. muestras; M= núm. genes.

La aproximación matemática depende de la información de partida disponible. En nuestro caso disponemos de *T* (derivada de microarrays) y de *C* (derivada de experimentos *single-cell*). Con estas premisas, las proporciones celulares se pueden obtener minimizando la suma de cuadrados (*linear least squares*, LLS) de la diferencia entre los valores estimados de $C \cdot P$ y los observados $T$. El inconveniente de esta aproximación es que puede dar como resultado proporciones negativas para algún tipo celular y/o sumas de proporciones mayores de 1. Por ese motivo se utiliza el método *non-negative least squares* (NNLS) con dos limitaciones asignadas: suma de proporciones = 1 y no-negatividad[6]. Esta es la manera que han encontrado los investigadores de CIBERSORTx[21] para solucionar la deconvolución. El rendimiento de NNLS en mezclas de muestras complejas se puede ver afectado por el ruido, por lo que los autores de CIBERSORTx han implementado una serie de filtros. Además, han desarrollado dos métodos de normalización (B y S) que permiten aplicar la deconvolución a varios tipos de datos provenientes de distintas plataformas (microarray, RNA-seq) y muestras (congeladas, FFPE). Estos métodos de corrección se basan en que *T* se puede modelar como una combinación lineal de *C*. Las estimaciones *T**, obtenidas a partir del modelo, sirven para aplicar *ComBat* (un método de corrección por lotes). Con esta estrategia consiguen eliminar las variaciones técnicas entre el *bulk* y la matriz de identidades[21].

### 3.5 Matriz de identidades.

La matriz de identidades, aunque no es imprescindible, aporta un mejor rendimiento al modelo[6]. Estas matrices se pueden derivar de experimentos transcriptómicos de células aisladas por FACS o de líneas celulares que representan tipos celulares[21]. Esta aproximación presenta el problema de que no permite el descubrimiento de nuevos tipos o estados celulares. Las tecnologías de *single-cell RNA-seq* permiten, además de obtener información relevante sobre estadios celulares[22], obtener matrices de identidades a partir de células en suspensión.

### 3.6 *single-cell* RNA-seq

Las secuenciación *single-cell* engloba una serie de técnicas que permiten la obtención de datos multidimensionales de cientos a miles de células individuales. Los datos generalmente son genómicos (genoma, epigenoma, transcriptoma), pero también se actualmente se están desarrollando proteómicos[23].

Brevemente, la captación de células individuales se realiza por medio de micromanipulación (captura láser), microfluidica (droplets, microwells), FACS, entre otros. Cada método tiene su problemática (redimiento, contaminación, inspección de las células) y su elección depende del diseño experimental.

Los investigadores de conjunto de datos que usaremos para la obtención de la matriz de identidades digirieron las biopsias gástricas con colagenasas y las aislaron individualmente con un citómetro de flujo (MoFlo XDP, Beckman coulter)[5]. Un sistema microfluídico (Chromium 10X) se encarga de microencapsular en gotas de aceite: i) células individuales ii) reactivos de retrotranscipción y iii) unas *beads* que contienen cebadores constituidos por: primers de secuenciación Illumina (p5, p7), códigos únicos de 16 nt (10X Barcodes), identificadores únicos de 10 nt (*unique molecular identifiers*, UMIs) y un cebador oligo-dT. Los 10X Barcodes etiquetan células y los UMIs moléculas individuales, lo que permite identificar las células y eliminar artefactos generados durante la amplificación por PCR. Tras la retrotranscripción el aceite es eliminado, los productos obtenidos amplificados por PCR, y las librerías sintetizadas y secuenciadas con los pasos usuales (fragmentación , selección, adición de índices…)[24]. La secuenciación resultará en archivos FASTQ y Cell Ranger (10x Genomics) se encargará del control de calidad, mapeo y la cuantificación. De esta manera se obtiene la matriz de cuentas con genes en filas y células individuales en columnas.

## *3.7* Análisis de experimentos transcriptómicos *single-cell.*

Consta de varios pasos.

### 3.7.1 Control de calidad.

Frecuentemente, las reacciones de secuenciación capturan más de una célula (*doublets*). Estos datos deben filtrarse ya que pueden llevar a errores de interpretación, especialmente si son heterotípicas (las células pertenecen a linajes diferentes). La solución por el momento es computacional. A partir de los datos originales se generan artificialmente miles de *doublets* y después se aplica algún tipo de clasificador como kNN (*k-nearest neighbors*). Aquellas células reales que se agrupen con las generadas, serán etiquetadas como *doublet* (*guilty by association*)[25]. Otro aspecto que se tiene presente es el número de genes detectados (*nFeature*), el número de moléculas de RNA por célula (*nCount*) y el porcentaje de genes mitocondriales (*mito*). Un nFeature bajo puede indicar que esa célula está muriendo, mientras que números altos de nFeature y nCount podrían indicar la presencia de *doublets*. En cuanto al porcentaje de cuentas que mapean en genes mitocondriales cabe decir que su filtrado incluye una parte de interpretación. Un porcentaje alto puede indicar que se ha roto la membrana citoplasmática o que se han activado procesos apoptóticos durante el procesado de la muestra. No obstante, el mismo porcentaje podría tener un significado fisiológico (como las células tubulares renales, con gran actividad mitocondrial) o patológicos (como en células tumorales).

### 3.7.2 Normalización, escalado y selección de genes.

La normalización se usa para eliminar el efecto del tamaño de las librerías. La normalización es el $\log_2$ de la división de las cuentas de cada gen entre las cuentas totales para esa célula y multiplicando por un factor de $10^4$. Tras esto hay una selección de genes. Seurat selecciona los genes más variables y elimina los, teóricamente, menos informativos. Este proceso de selección hace que los resultados finales sean más fácilmente interpretables[26]. Tras esto, los datos son escalados con z-score ($z = (x- \mu) / \sigma$). El escalado se requiere para el siguiente paso.

### 3.7.3 Reducción de la dimensionalidad.

Los data sets de la metodologías -ómicas tienen miles de dimensiones. Se hace necesario reducir su dimensionalidad para la eliminación de variables poco informativas o redundantes que pueden dar problemas como el *curse of dimensionality* (cuantos más *features* se incorporan al modelo, menos distancia hay entre las células). La reducción facilita pues los análisis posteriores (como el *clustering*), aumenta la eficiencia computacional y facilita la visualización en dos dimensiones. Hay varios disponibles, los más frecuentes son PCA, t-SNE, UMAP. Mientras que

el primero se basa en la transformación lineal de los datos, los otros son combinaciones no-lineales[27].

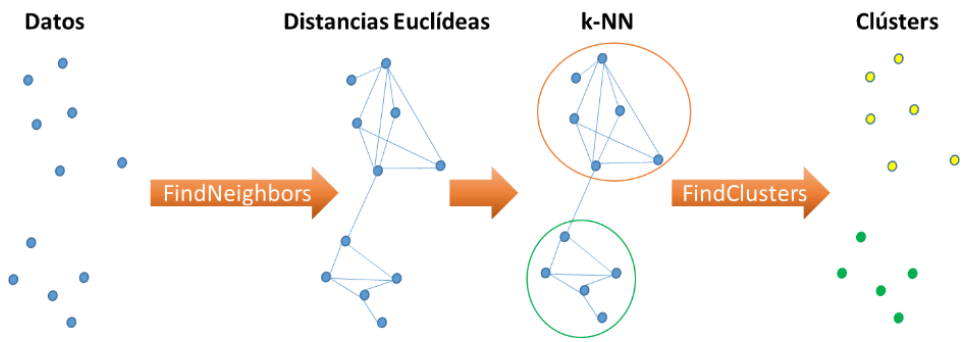
### 3.7.4 Efectos de lotes (*batch effects*).

Los efectos por lotes se producen cuando los efectos que observamos entre grupos se deben a causas no biológicas (técnicas). Estas causas pueden ser temporales, lotes de reactivos, técnica personal, instrumentales, agrupación de casos y de controles, condiciones ambientales… Se han desarrollado múltiples técnicas para corregir los efectos por lotes[28]. Lo ideal sería que el diseño experimental las corrigiera, puesto que también pueden eliminar cierta variación biológica. En este estudio utilizamos Harmony[29]. El propósito es que un *cluster* signifique un tipo celular y no una causa técnica.

### *3.7.5 Clustering.*

Hay multitud de algoritmos de *clustering* para single-cell[30]. El objetivo es el de agrupar células similares que se identificarán y etiquetarán usando marcadores biológicos. Para este estudio usamos las funciones de proporcionadas por Seurat. En una, FindNeighbors(), usando los componentes principales del PCA, se realiza un *clustering* de grafos. Los grafos son redes en las que los nodos están conectados por aristas (Figura 2. Proceso de clustering.). Los nodos son las células y las aristas pueden ser varios tipos de medidas de similitud. Seurat usa como medida de similitud la distancia Euclídea. El *clustering* lo hace con kNN. Con esto se intenta encontrar agrupamientos que contengan nodos muy conectados entre sí (que estén cerca en el espacio Euclídeo), pero que estén poco conectados con otros nodos. Para valorar la similitud también usan el coeficiente de Jaccard. Después, para obtener los *clusters* separados, usan el algoritmo de Louvain. Esto lo hace la función FindClusters(), y el parámetro resolution ayuda a ajustar el número de *clusters* que se obtienen.

*Figura 2. Proceso de clustering.*



Esto lo hacen para que el investigador no tenga que intuir al inicio del *clustering* cuantos tipos celulares aparecerán en el experimento y así facilitar el descubrimiento de nuevos tipos, subtipos o estados celulares.

El *clustering*, sin embargo, es una técnica subjetiva ya que tiene mucho de interpretación, validación y optimización[31] .

### 3.7.6 Anotación.

La anotación consiste en identificar qué tipo celular corresponde a cada *cluster*. Hay, como en los procedimientos anteriores, una variedad de herramientas automáticas para este proceso[32]. Entre ellas están: *i)* las que consultan marcadores en bases de datos, como por ejemplo CellMarker, y *ii)* las que buscan equivalencias con experimentos previos de referencia de *single-cell RNA-seq*. Estos últimos, han sido revisados por expertos y están en bases de datos específicas como el Single Cell Expression Atlas[32]. La principal ventaja es que son capaces de identificar los tipos celulares más comunes y abundantes, pero pueden tener problemas para identificar subtipos.

En el presente estudio decidimos usar la anotación manual, basándonos en los tipos celulares esperados en la bibliografía[5,33,34]. Los tipos celulares se recogen en la Tabla 1. Tipos celulares gastrointestinales

*Tabla 1. Tipos celulares gastrointestinales*

| Grupo celular | Tipo celular | Localización anatómica | Descripción |
|---|---|---|---|
| **Células endocrinas** | G Cells | Antro (algunas duodeno) | Secretoras de gastrina |
| | D Cells | Todo el tracto gastrointestinal | Secretoras de somatostatina |
| | X Cells, MX Cells | Cuerpo | Secretoras de ghrelina |
| | EC Cells | Antro, duodeno, íleon | Células enterocromafinas |
| | ECL Cells | Mucosa gástrica | Células similares a las enterocromafinas. |
| **Células comunes (OX. Y PIL)** | Tuft Cells | | |
| | Pit Mucous Cells | Antro, cuerpo, fundus | Células secretoras de mucosa (= *foveolar, mucous*) |
| | Isthmus Cells | Cuerpo, fundus | Dos subtipos: stem cells y pit cells. |
| | Stem Cells | Antro, cuerpo, fundus | Originan varios tipos celulares |
| **Glándula OXINTICA** | Neck Cells | Cuerpo | Secretoras de mucus. |
| | Parietal Cells | Cuerpo | Secretoras de ácido y factor intrínseco. |
| | Chief Cells | Cuerpo | Secretoras de enzimas zimogénicas (pepsina) . |
| | REG3A+(Antimicrobial) | Cuerpo | Funciones antimicrobianas |
| **Glándula PILÓRICA** | Proliferative Isthmus | Antro | Renovación celular |
| | LYZ+ (Antimicrobial) | Antro | Funciones antimicrobianas |
| **Células INTESTINALES** | Goblet Cells | Intestino delgado | Secretoras de mucus. |
| | Enterocytes | Intestino delgado | Absorción de nutrientes. |
| | Enteroendocrine Cells | Intestino delgado | Células secretoras. |
| **SISTEMA INMUNITARIO** | T Cells | Todo el tracto gastrointestinal | Células T |
| | B Cells | Todo el tracto gastrointestinal | Células B |
| | Macrophages | Todo el tracto gastrointestinal | Macrófagos |
| | Mast Cells | Todo el tracto gastrointestinal | Mastocitos |
| | Neutrophiles | Todo el tracto gastrointestinal | Neutrófilos |
| **Células del ESTROMA** | Smooth Muscle Cells | Todo el tracto gastrointestinal | Células musculares lisas |
| | Endothelial Cells | Todo el tracto gastrointestinal | Células endoteliales |
| | Fibroblasts | Todo el tracto gastrointestinal | Fibroblastos |

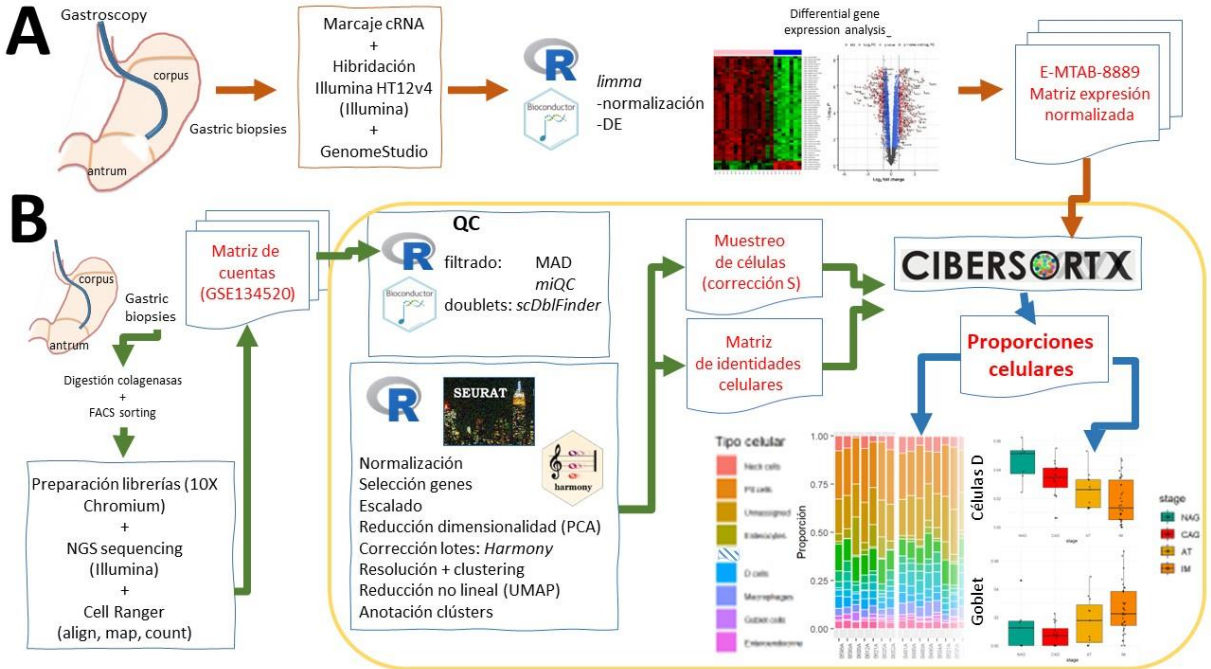
# 4 Metodología

## 4.1 Diseño y datos clínicos:

Las muestras que son el objetivo de la deconvolución son datos de expresión génica de microarrays de Lario S. *et al*[4,35] (Figura 1, panel A). Esta cohorte está constituida por 96 biopsias gástricas (70 antrales y 26 de cuerpo) obtenidas de 78 pacientes dispépticos. A este conjunto de datos lo denominaremos "*bulk expression data*."

Para la obtención de la matriz de identidades celulares hemos usado los datos de Zhang, P. et al[5] (Figura 1, panel B). Este conjunto de datos contiene los datos de expresión génica de célula única de 13 biopsias gástricas obtenidas de 9 pacientes con síntomas dispépticos. Las biopsias se clasifican como NAG (n= 3), CAG (n= 3), IM (n= 6) y cáncer gástrico temprano (n= 1). Los autores no lo mencionan en el artículo, pero a juzgar por las imágenes endoscópicas, las biopsias parecen recogidas de la región gástrica del antro. Las muestras de cuerpo del bulk data no serán analizados.

*Figura 3. Diseño del estudio*



## 4.2 Análisis single-cell RNAseq

Para el análisis *single-cell* partimos de una matriz de cuentas depositada en GEO (GSE134520). Los datos suplementarios ofrecen un archivo comprimido que contiene los 13 archivos que representan las 13 muestras.

Cada uno de esos archivos es un archivo de texto en el que en las columnas están las células, identificadas con sus correspondientes UMIs (*Unique Molecular Identifier*) y en las filas los genes (*features*) cuya expresión se ha determinado. La mayoría de posiciones de la matriz (*counts*) contienen valores cero. Como procesar este tipo de matrices es poco eficiente, lo que se hace es transformarlas en un tipo especial de matrices llamadas matrices dispersas (*sparse matrices*). Las matrices dispersas son las que usa el paquete Seurat[8] para un procesado más eficiente y con menor consumo de memoria.

Usamos el paquete Matrix para convertir los archivos de texto en tres nuevos archivos: i) *features*, ii) *barcodes* (UMIs) y iii) *matrix* (en formato.mtx). Tras esto, creamos el primer objeto Seurat (usando Read10x() y CreateSeuratObject()).

### *4.2.1* Control de calidad: *doublets*

Para la detección de *doublets* se utilizó el paquete scDblFinder[25]. Se identifican mediante simulación, construyendo *doublets* artificiales a partir de los datos originales y eliminando aquellas células que, en un análisis de clústeres, se encuentran cerca de estos *doublets* artificiales. El parámetro dbr, proporción de doublets esperado, se dejó en automático como sugiere la guía de uso para datos originados con 10X.

### 4.2.2 Control de calidad: filtrado calidad.

Usamos métodos adaptativos basados en los datos (*data-driven*) para filtrar las células. Utilizamos dos: *i)* basado en *median absolute deviation* (MAD)[36,37] y, *ii)* basado en *finite mixture model* (FMM)[38] en el que se determina la probabilidad de una célula de ser de baja calidad (*compromised*).

MAD es una medida de variabilidad que se define como:

$$MAD = mediana|x_i - mediana(x)|$$

Este estadístico se usa porque se ve menos afectado por los valores extremos de los datos (es robusto). Se usa para determinar un límite, de n veces MAD, tras el cual los datos son considerados atípicos (*outliers*). Se usó isOutlier() de scater optimizando los parámetros: nmads (2, 2.5, 3) y log (TRUE/FALSE). Algunos investigadores recomiendan usar la transformación logarítmica, ya que los datos de scRNA suelen tener colas largas por la derecha. Se calcularon por muestra (batch = orig.ident). Para el segundo, se ha empleado mixtureModel() del paquete miQC[38] y un *cut-off* de 0.75.

#### 4.2.3 Normalización, escalado y selección de genes.

Para esto se usaron los parámetros por defecto de NormalizeData(), FindVariableFeatures() y ScaleData(). La normalización se usa para eliminar el efecto del tamaño de las librerías. Es el $\log_2$ de la división de las cuentas de cada gen entre las cuentas totales para esa célula y multiplicando por un factor de $10^4$. Seurat selecciona los genes más variables y elimina los, teóricamente, menos informativos. Esto lo hace con FindVariableFeatures(). Tras esto los genes se escalaron a media 0 y desviación standard 1.

#### 4.2.4 Reducción de la dimensionalidad.

Se hizo con RunPCA() y sus parámetros por defecto. El número de dimensiones se escogió por la interpretación del ElbowPlot().

#### 4.2.5 Corrección de efectos por lotes.

Se realizó con el paquete Harmony[29] y su función RunHarmony(). Consideramos cada biopsia como un lote, por lo que corregimos para 'orig.ident'.

#### 4.2.6 Resolución y clustering.

resolution es un parámetro de FindClusters() que se ha de establecer para cada experimento. Valores crecientes de resolution dan lugar a mayor número de clústers. Se examinó para valores de 0.5 a 1.0, en pasos de 0.1. Para cada valor de resolution se obtuvo el número de clusters por la visualización de los UMAP plots (DimPlot(reduction= 'umap')) tras la ejecución de las funciones RunUMAP() y FindNeighbors(), usando las 30 primeras dimensiones del PCA.

#### 4.2.7 Anotación de los clusters.

El etiquetado de los *clusters* se realizó por la revisión visual de varios tipos de gráficos (ViolinPlot(), FeaturePlot()) coloreados con los genes que identifican los tipos celulares gástricos, intestinales e immunitarios[5,33,34].

## 4.3 Deconvolución con CIBERSORTx.

CIBERSORTX es una herramienta de acceso online[39] que requiere de tres tipos de archivo. El primero es la matriz con los datos *bulk*, el segundo es la matriz de identidades con los genes marcadores (*signature matrix*) y la tercera es una matriz para realizar una corrección por lotes.

### 4.3.1 Matriz *bulk*

Los datos de expresión E-MTAB-8889[4], se importaron y filtraron en R. Estos datos de microarrays estaban previamente normalizados por cuantiles. La muestras de cuerpo fueron eliminadas ya que la matriz de identidades celulares se generó con muestras antrales[5].

### 4.3.2 Signature matrix

La tabla de identidades celulares está constituida en las filas por los nombres de los genes marcadores (SYMBOL) y en columnas por la etiqueta de las células. Las celdas contienen la media de expresión del gen *i* para el tipo celular *j*. Seurat proporciona AverageExpression() para este propósito.

### 4.3.3 Matriz *single-cell* para corrección *batch* tipo S.

CIBERSORTx permite hacer una corrección por lotes. En el caso que nos ocupa se requiere del modo *S*, según los tutoriales. El modo S utiliza un muestreo de la matriz de expresión. Se seleccionaron 25 células por cada tipo celular etiquetado. En filas están los nombres de los genes (SYMBOL) y en columnas la etiqueta de las células individuales. Una celda *i, j* contiene las cuentas normalizadas de una única célula para un gen.

Los tres archivos se subieron al servidor de CIBERSORTx y se siguió el siguiente protocolo (Figura Anexo-10): "IMPUTE CELL FRACTIONS"→ "CUSTOM" → seleccionar *batch file* → seleccionar *signature matrix* → "ENABLE BATCH CORRECTION" → "S-MODE" → seleccionar *single-cell reference matrix* → "PERMUTATIONS" → n= 100 → "Run".

El archivo de salida de CIBERSORTX con las proporciones celulares y las significaciones estadísticas se importó a *R*.

## 4.4 Análisis estadístico de las proporciones celulares

Los gráficos QQ y densidad y el test de Shapiro-Wilk se usaron para determinar la normalidad de las distribuciones por grupos dentro de la cascada de Correa. ANOVA o Kruskall-Wallis se utilizaron según corresponda para comparar la proporción de los tipos celulares entre los pasos

de la cascada de Correa. Las comparaciones post hoc por pares entre dos grupos se evaluó mediante Tukey o Wilcoxon rank sum exact test con los p- valores y corrección de pruebas múltiples por Benjamini y Hochberg. La significación se estableció en $p < 0.05$. Se usaron los paquetes stats, cars y ggplot2.

# 5 Resultados

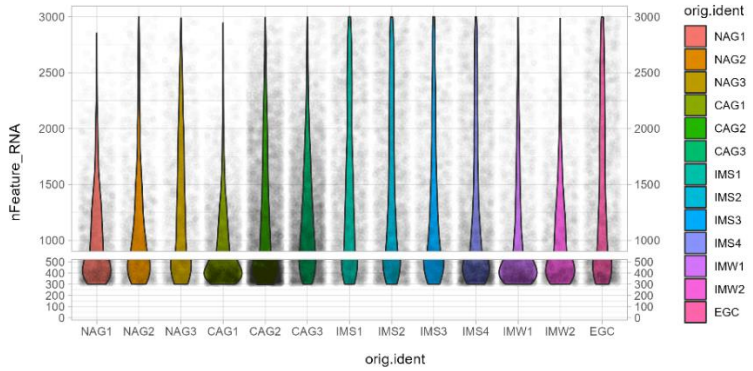
Primero se exponen los resultados del experimento *single-cell* necesarios para obtener la matriz de identidad y de corrección por lotes. Tras esto se recogen los resultados de la deconvolución.

## 5.1 Análisis *single-cell* RNAseq

### 5.1.1 Los datos 'crudos' están prefiltrados.

Con los gráficos de QC nos dimos cuenta que los datos no eran crudos como reportaban los autores, sino que estaban filtrados. Claramente habían eliminado aquellas células con nFeature < 300 (Figura 4)

*Figura 4. Datos prefiltrados por los autores.*




### 5.1.2 Los *doublets* correlacionan con el número total de células.

La Tabla 2 muestra los porcentajes de *doublets* para cada uno de los 13 experimentos. El número de *doublets* correlacionó positivamente con el número de células secuenciadas (Figura 5, r = 0.98, $p<1.9x10^{-9}$).

*Figura 5. Correlación doublets*

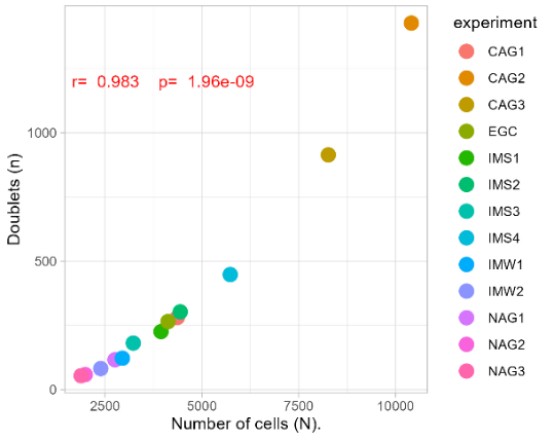



*Tabla 2. Número y porcentaje de doublets*

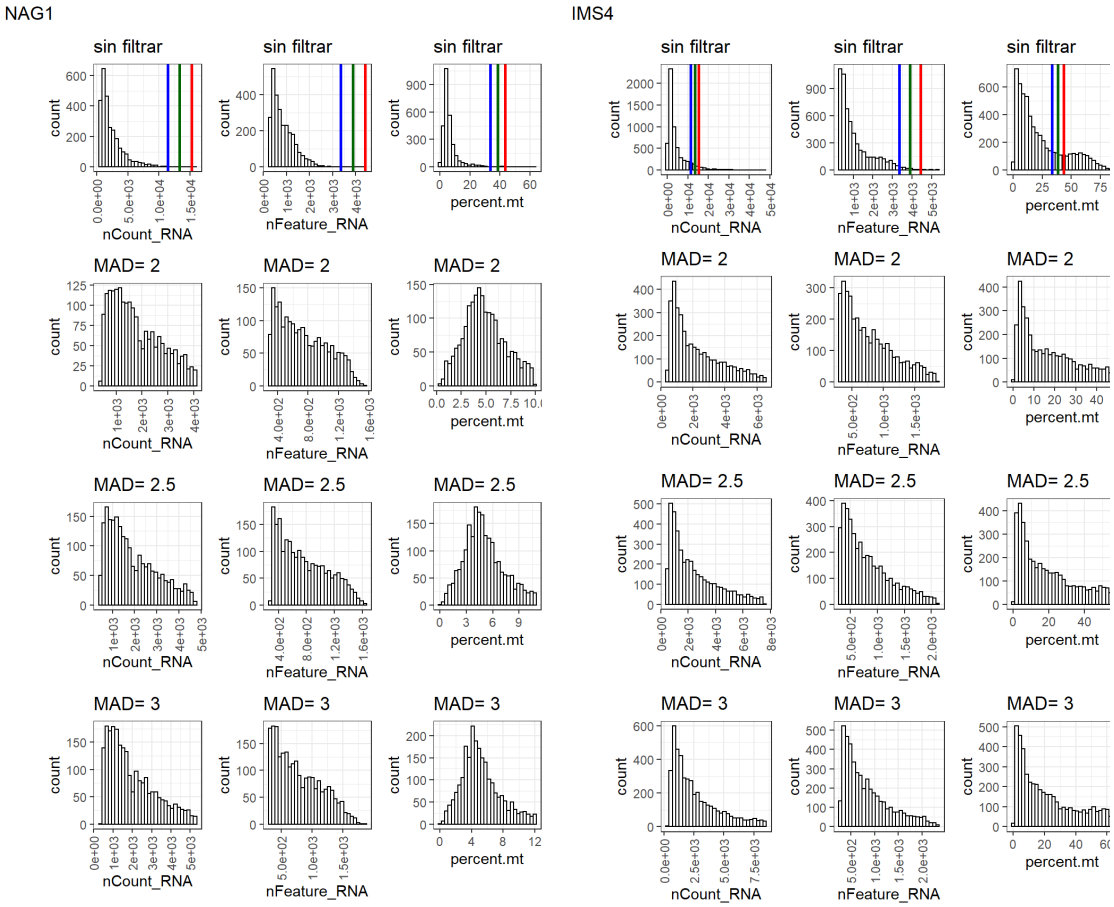
| Experiment | Singlet (N) | Doublet (N) | doublet_percent (%) |
|---|---|---|---|
| **NAG1** | 2656 | 118 | 4.2 |
| **NAG2** | 1937 | 61 | 3.0 |
| **NAG3** | 1815 | 50 | 2.7 |
| **CAG1** | 4090 | 281 | 6.4 |
| **CAG2** | 8972 | 1429 | 13.7 |
| **CAG3** | 7353 | 916 | 11.1 |
| **IMW1** | 2823 | 118 | 4.0 |
| **IMW2** | 2298 | 86 | 3.6 |
| **IMS1** | 3711 | 227 | 5.8 |
| **IMS2** | 4139 | 303 | 6.8 |
| **IMS3** | 3063 | 180 | 5.6 |
| **IMS4** | 5259 | 445 | 7.8 |
| **EGC** | 3848 | 262 | 6.4 |

### 5.1.3 El QC usando MAD elimina células atípicas.

La

Figura 6 muestra los histogramas para dos de los experimentos (NAG1 e IMS4), el resto se pueden ver en el Anexo-3b. El filtrado MAD resultó en distribuciones de varios tipos, algunas son bimodales, otras gausianas (únicamente para %mito) y, la mayoría, sesgadas a la izquierda. Los tres MADs resultan en histogramas similares. Concluimos que, de filtrar por MAD, el valor de nmds= 3 sería el más adecuado ya que preserva un mayor número de células. No se observó efecto del parámetro log en este data set (no se muestra).

*Figura 6. Histogramas tras filtrados nMAD (2, 2.5, 3).*

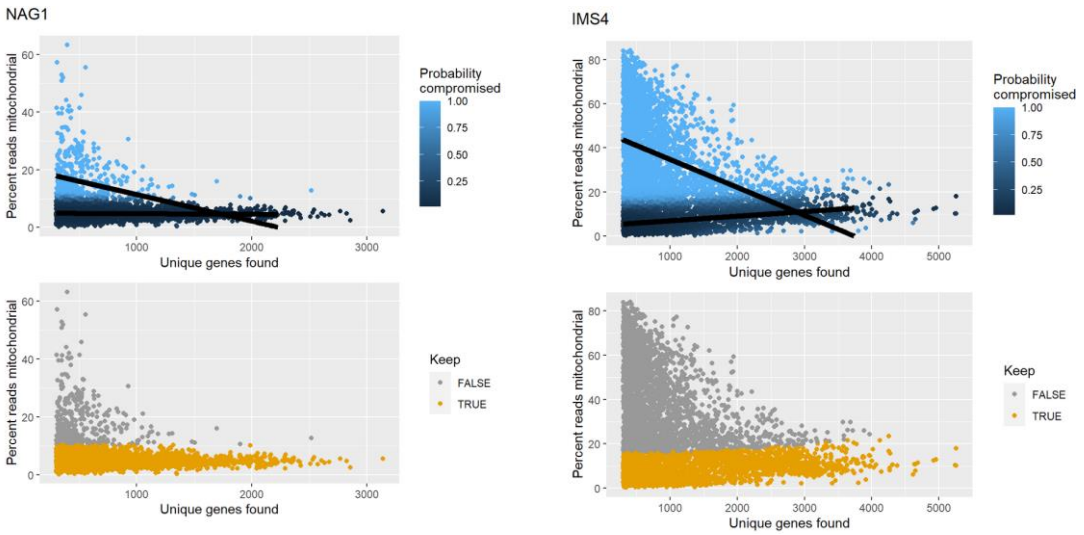



El panel de la izquierda es la muestra NAG1 y el de la derecha IMS4. Las columnas por muestra son las variables nCount, nFeature y percent_mito. Las filas son los histogramas sin filtrar y tras la aplicación de MAD= 2, 2.5, 3. Las barras de verticales de colores en los gráficos sin filtrar indican los límites superiores para MAD= 2 (azul), MAD= 2.5 (verde) y MAD= 3 (rojo).

### 5.1.4 QC basado en miQC.

Se muestra a continuación (Figura 7) dos ejemplos de la salida de miQC, uno con pocas células *compromised* (NAG1, ~13%) y otro con muchas (IMS4, ~47%). Aquellas células con una probabilidad de más de 0.75 de ser de baja calidad son eliminadas (paneles inferiores , puntos grises).

*Figura 7. Salida de 'miQC'.*




El panel de la izquierda es la muestra NAG1 y el de la derecha IMS4. Se muestran los nFeatures en el eje abscisas y el % de genes mitocondriales en el de ordenadas. Las líneas negras del panel superior son las obtenidas por el modelo FMM. El modelo asume que hay dos tipos de células, las intactas (definidas como bajo % de genes mitocondriales y número alto de nFeatures) y las comprometidas (que tienen altos niveles de genes mitocondriales y bajos nFeatures). Con esta base se crea un modelo estadístico probabilístico el cual asigna a cada célula una probabilidad estar comprometida (panel superior, gradación de color azul). Aquellas con probabilidades >0.75 son eliminadas del experimento (puntos grises, panel inferior).

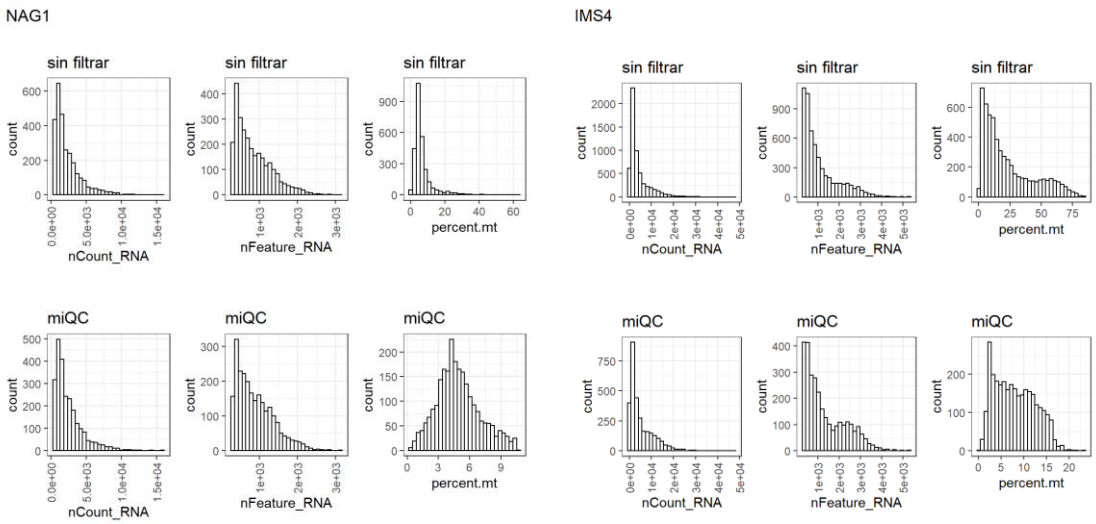
La

Figura 8 muestra de nuevo los histogramas para los experimentos NAG1 e IMS4. El resto se pueden ver en el Anexo-3c. Como con MAD la aplicación de miQC resultó en distribuciones sobre todo sesgadas a la izquierda (para nCount y nFeature) y gausiana para el porcentaje de genes mitocondriales. En general, en nuestro experimento, obtenemos dos tipos de histogramas:

- los que son parecidos a NAG1: NAG1, NAG2, NAG3, CAG1, CAG2, CAG3, IMS3, ECG.

- los similares a IMS4 (con perfiles bimodales): IMS2, IMS4.

*Figura 8. Histogramas tras el filtrado con miQC ($p_{compromised} \leq 0.75$).*

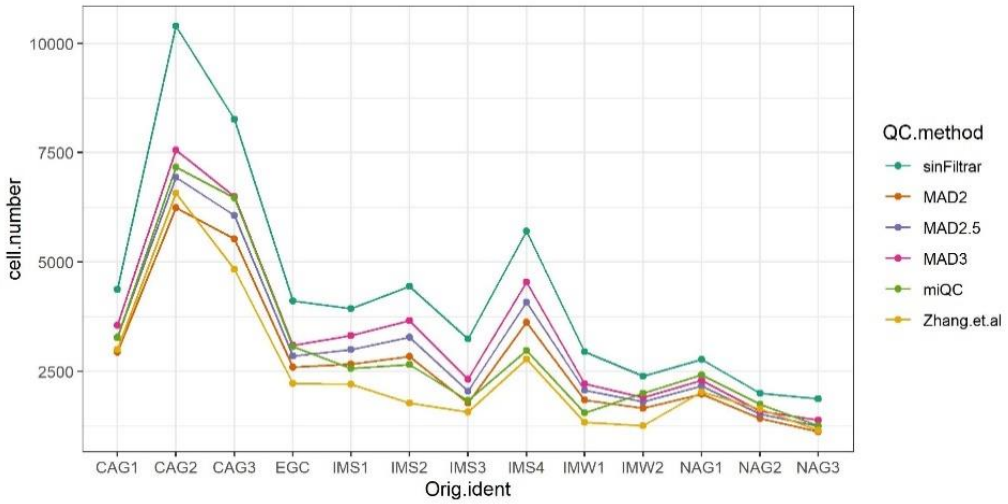



El panel de la izquierda es la muestra NAG1 y el de la derecha IMS4. Las columnas por muestra son las variables *nCount, nFeature* y *percent_mito*. La primera fila de cada panel son los histogramas sin filtrar, la segunda tras la aplicación del modelo FMM con $p_{compromised} \leq 0.75$.

### 5.1.5 Los filtros adaptativos fueron menos restrictivos que filtros de los autores.

En cuanto a los resultados de número de células de los 13 experimentos (Figura 9), vemos que miQC es más restrictivo que MAD para ciertos experimentos (IMS1, IMS2, IMS3, IMS4, IMW1, NAG3), pero menos para otros (IMW2, NAG1, NAG2) y siempre menos restrictivo que los puntos de corte no adaptativos de los autores.

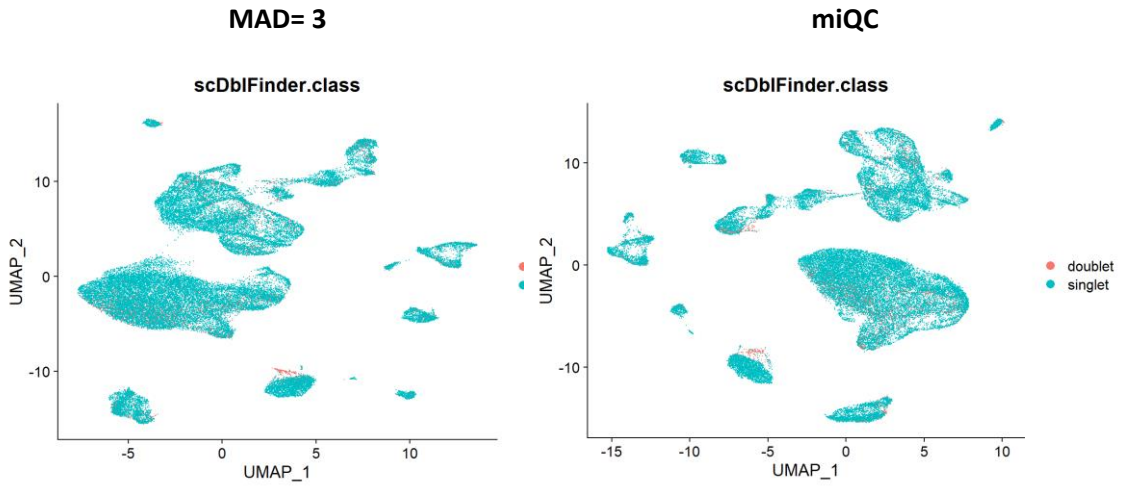
*Figura 9. Comparación QC con MAD, miQC y autores del estudio.*




### 5.1.6 Los filtros adaptativos no consiguen eliminar los *doublets.*

Con tal de comprobar si MAD o miQC descartaban los *doublets* detectados con scDblFinder, elaboramos un UMAP plot. La Figura 10 muestra que estos filtros no son capaces de eliminarlos.

*Figura 10. Gráficos UMAP tras el filtrado con MAD= 3 (izquierda) o miQC (derecha).*




Podemos ver como los doublets (en rojo) se distribuyen por todos los clusters a excepción de una isla más homogénea (abajo en el centro y abajo a la izquierda). Aunque no están pensadas para ello, ninguna de las dos técnicas de filtrado ha conseguido eliminarlos.

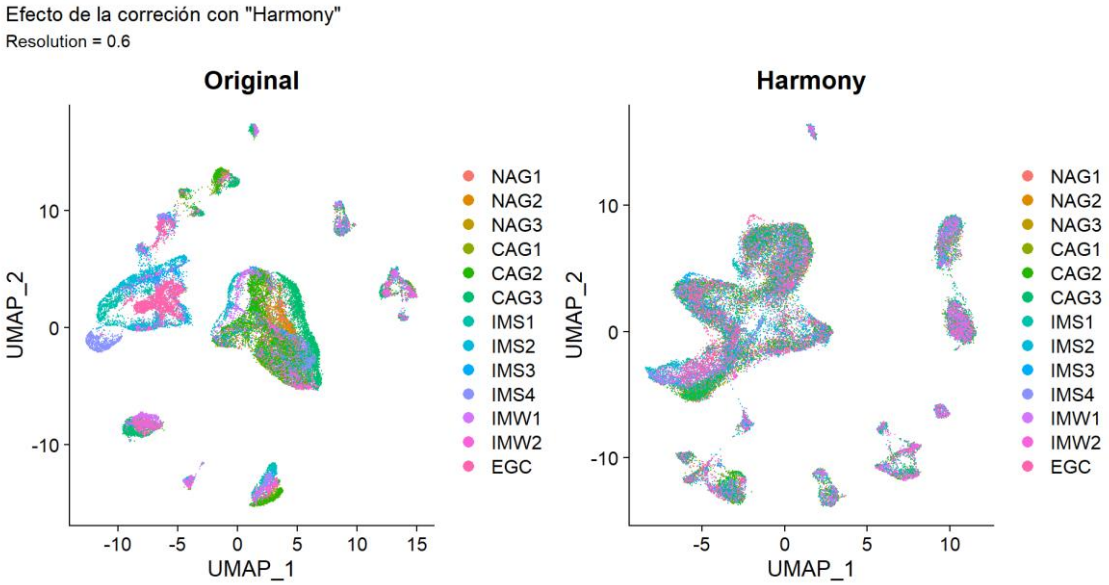
En conjunto los resultados anteriores nos indicaron que el QC que había que seguir era el filtrado con scDbFinder y miQC.

### 5.1.7 Es necesaria la corrección por lotes.

La comparación de los UMAP entre el dataset sin corregir y corregido por muestra (único lote conocido disponible) mostró la necesidad de corregir

Figura 11). Vemos que, tras Harmony, los colores que identifican las muestras se distribuyen mejor por todos los *clústers*. Sin la corrección hay más colores agrupados.

*Figura 11. Corrección por lotes.*

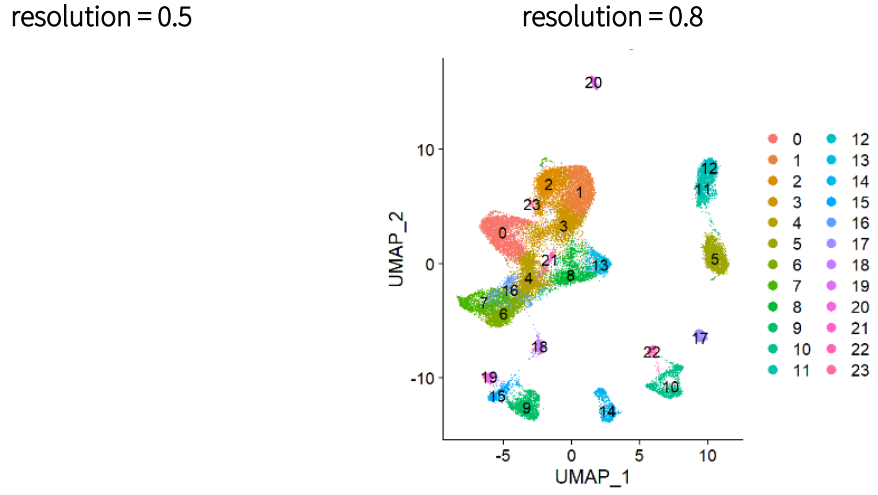


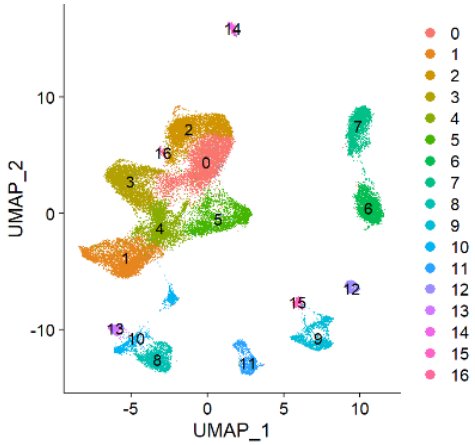

La corrección se realizó con el paquete Harmony y su función RunHarmony(). Se consideró cada biopsia como un lote, por lo que corregimos para 'orig.ident'.

### 5.1.8 Resolution = 0.8 proporciona la granulometría necesaria.

Se probaron seis niveles de resolución. La Figura 12. Efecto de resolution sobre el número de clústers. recoge el UMAP del considerado como válido para nuestro objetivo. El resto se puede encontrar en el Anexo-6.

*Figura 12. Efecto de resolution sobre el número de clústers.*






Este parámetro se ha de optimizar para en cada experimento. Valores crecientes de resolution dan lugar a mayor número de *clusters.* Se determinó para valores de 0.5 a 1.0, en pasos de 0.1. Panel de la izquierda resolution = 0.5, 17 *clusters.* Panel de la derecha: resolution= 0.8, 24 *clusters*.

Decidimos mantener resolution a 0.8, que resulta en 24 *clústers* celulares. Establecemos esta resolución por el número elevado de tipos celulares esperados y porque el *clúster 18* únicamente es identificado a partir de este valor de resolution. El *clúster* 18 se puede etiquetar claramente como células caliciformes (*goblet cells*). Estas células tienen interés en el estudio de las lesiones precursoras.

### 5.1.9 Anotación de los clusters.

La interpretación de los gráficos y la búsqueda bibliográfica permitió elaborar nuestra anotación celular.

*Tabla 3. Etiquetado de los clústers según los datos bibliográficos.*

| Grupo celular | Tipo celular | Marcador[5,33,40] | Seurat *cluster* |
|---|---|---|---|
| **Células endocrinas** | G cells | GAST | 9 |
| | D cells | SST | 15 |
| | X cells, MX cells | GHRL, MLN | 9, 15 (Difuso ) |
| | EC cells | TPH1, DDC, CHGA, REG4 | 21 |
| | ECL cells | CCKBR, CHGA, HDC, P2RY14 | 21 |
| **Células comunes (OX. Y PIL)** | Tuft cells | TRPM5, SH2D6, DCLK1 | Pocas céls. y dispersas. |
| | Pit mucous cells | GKN1, GKN2, MUC5AC, TFF1 | 0, 1, 4 |
| | Isthmus cells | MKI67, STMN1, TOP2A, NMU | 13, 8 |
| | Stem cells | SOX2, CCKBR, OLFM4, LGR5 | 10, 3, 8, 19, 2 |
| **Glándula OXINTICA** | Neck cells | MUC6, TFF2 | 6, 4, 10, 0 |
| | Parietal cells | ATP4A, ATP4B, GIF | No (solo en cuerpo) |
| | Chief cells | PGA3, PGA4, LIPF, PGC | 0 |
| | REG3A+ (antimicrobial) | REG3A, LTF, LCN2 | 10 |
| **Glándula PILÓRICA** | Proliferative isthmus cells | NMU, MKI67, STMN1, TOP2A, BIRC5 | 13, 8 |
| | LYZ+ cells (antimicrobial) | LYZ, PPP1R1B, AQP5 | 6, 10 |
| **Células INTESTINALES** | Goblet cells | MUC2, ITLN1, SPINK4 | 18 |
| | Enterocytes | APOA1, ALPI, FABP1, APOA4 | 7, 16, 4, 21 |
| | Enteroendocrine cells | CHGB, CHGA, TAC1, TPH1, NEUROG3 | 19, 9 |
| **SISTEMA INMUNITARIO** | T cells | CD2, CD3D | 5 |
| | B cells | CD79A | 11, 12 |
| | Macrophages | CSF1R, CD68 | 17 |
| | Mast cells | TPSAB1 | 20 |
| | Neutrophiles | FCGR3A, MMP9, S100A9, ITGB2 | 17 |
| **Células del ESTROMA** | Smooth muscle cells | ACTA2 | 22 |
| | Endothelial cells | VWF, ENG | 14 |
| | Fibroblastos | DCN, PDPN | 10 |

Así:

- la Figura 13 muestra que la expresión de gastrina (GAST) está fuertemente asociada al clúster 9, por lo que podemos asumir que este clúster representa a las células G.
- Lo mismo ocurre con el clúster 18 y los tres marcadores de células caliciformes MUC2, ITLN1, SPINK4 (Figura 14).
- Otros tipos celulares no son tan evidentes, como las células secretoras de moco (pit cells, Figura 15) cuyos marcadores se distribuyen en tres clústers (1,2,6).
- Encontramos que dos tipos celulares, macrófagos y los neutrófilos, comparten un *clúster*, el 17 (Figura 16).
- Uno de los tipos celulares que no se pudo resolver son las *stem cells*. Según la bibliografía, presentan varios marcadores, pero ninguno de ellos parece asociarse claramente a un clúster (Figura 17).
- Por último, hay dos clústers (3, 23) que no presentan marcadores claros y que se han etiquetado como no asignados.

Feature y ViolinPlots adicionales, incluyendo los de QC, se pueden ver en el Anexo-8.

*Figura 13.Identificación del clúster de células D*

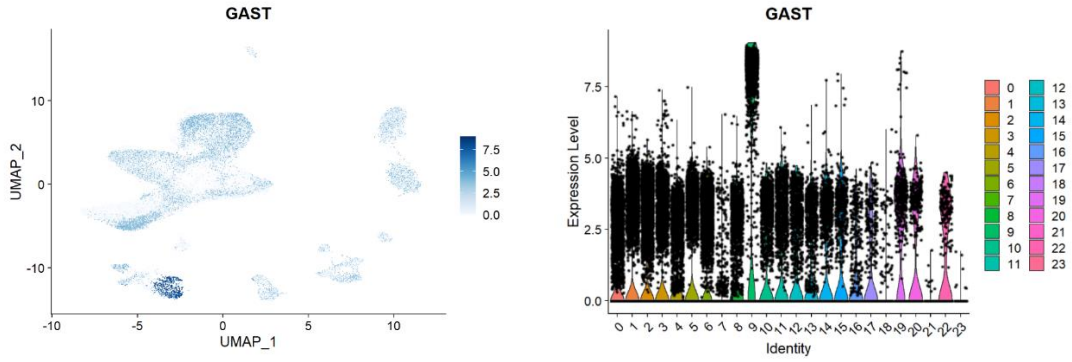


*Figura 14. Identificación del clúster de células caliciformes.*

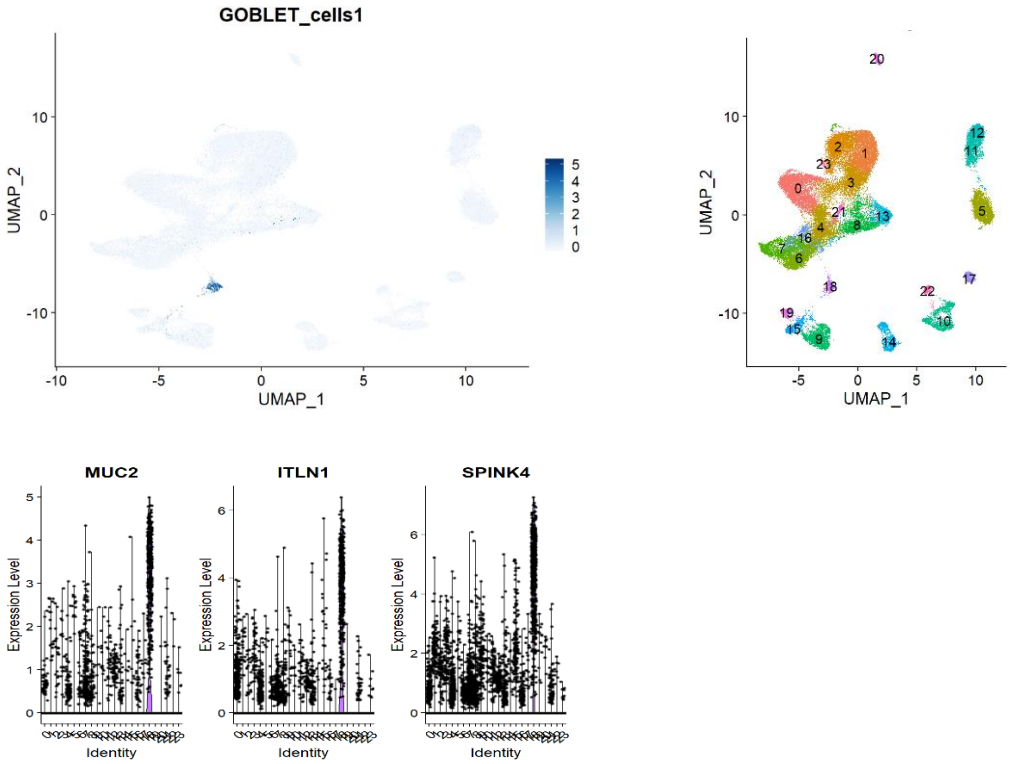



*Figura 15. Identificación del clúster de células secretoras de mucus (pit cells).*

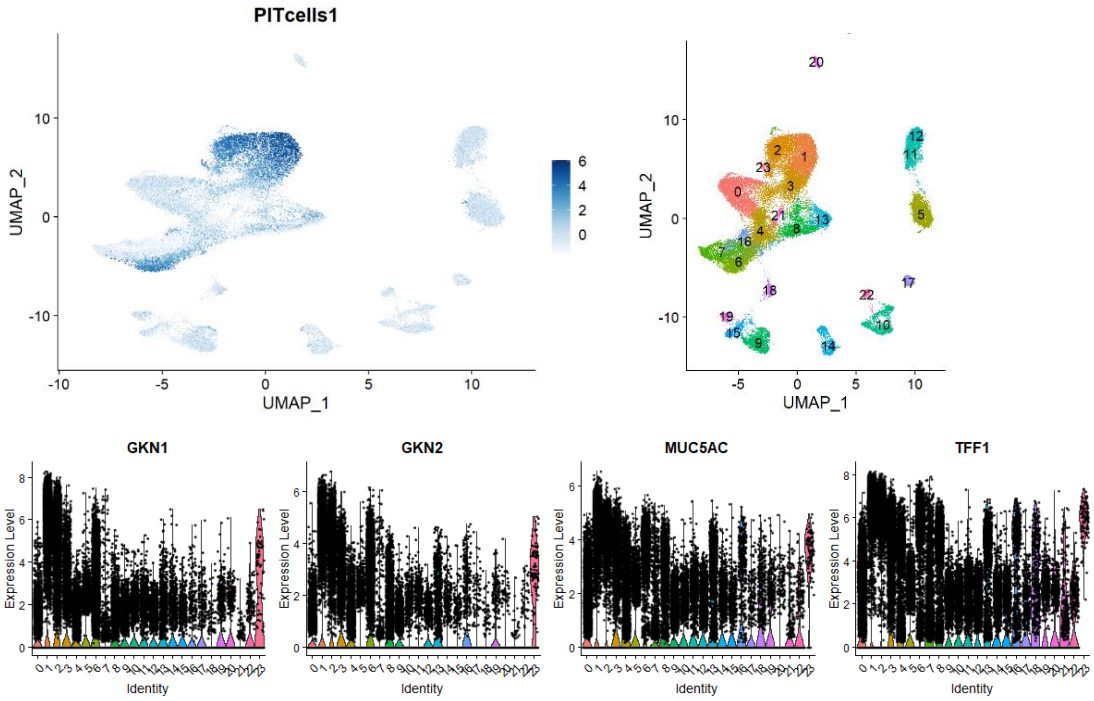


*Figura 16. Macrófagos y neutrófilos*

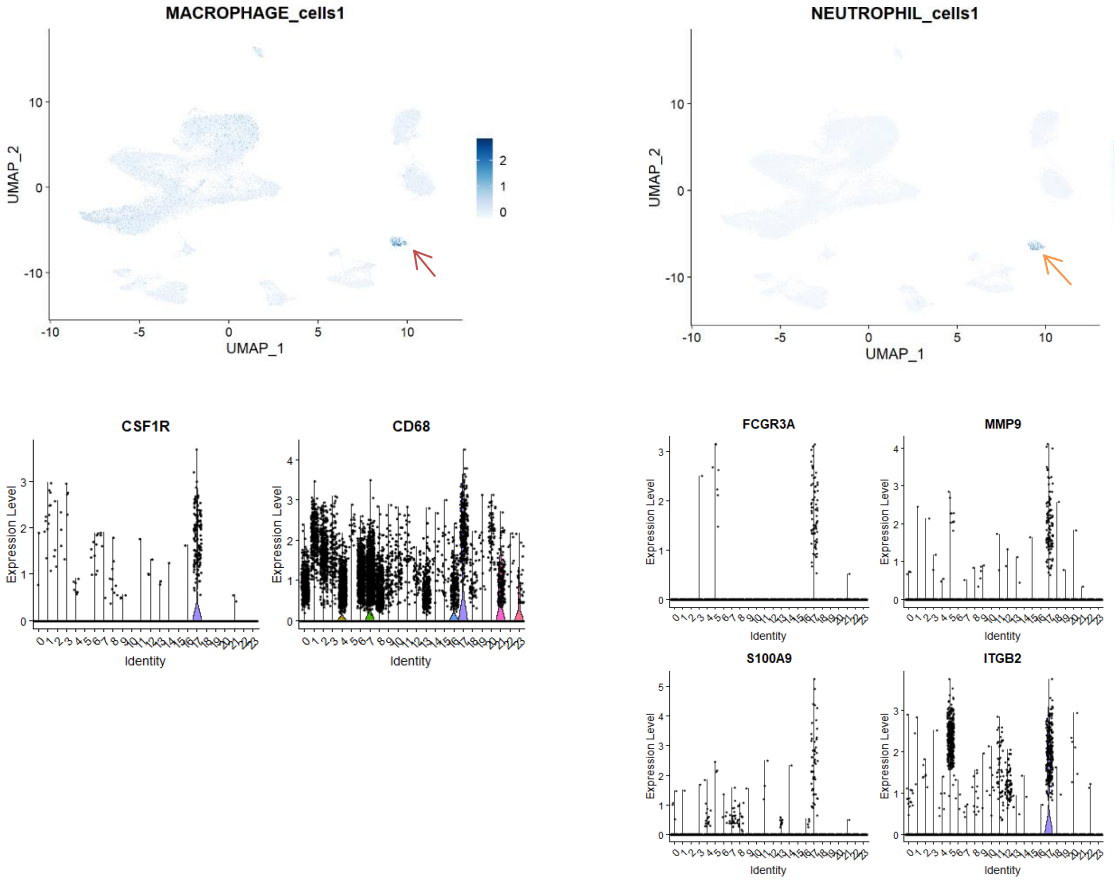
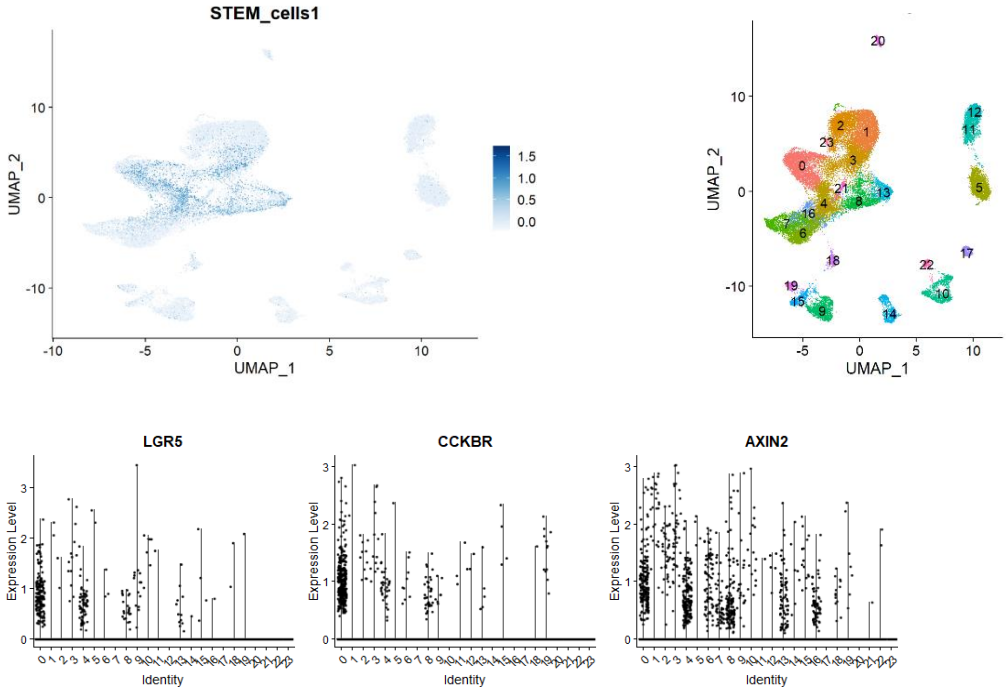
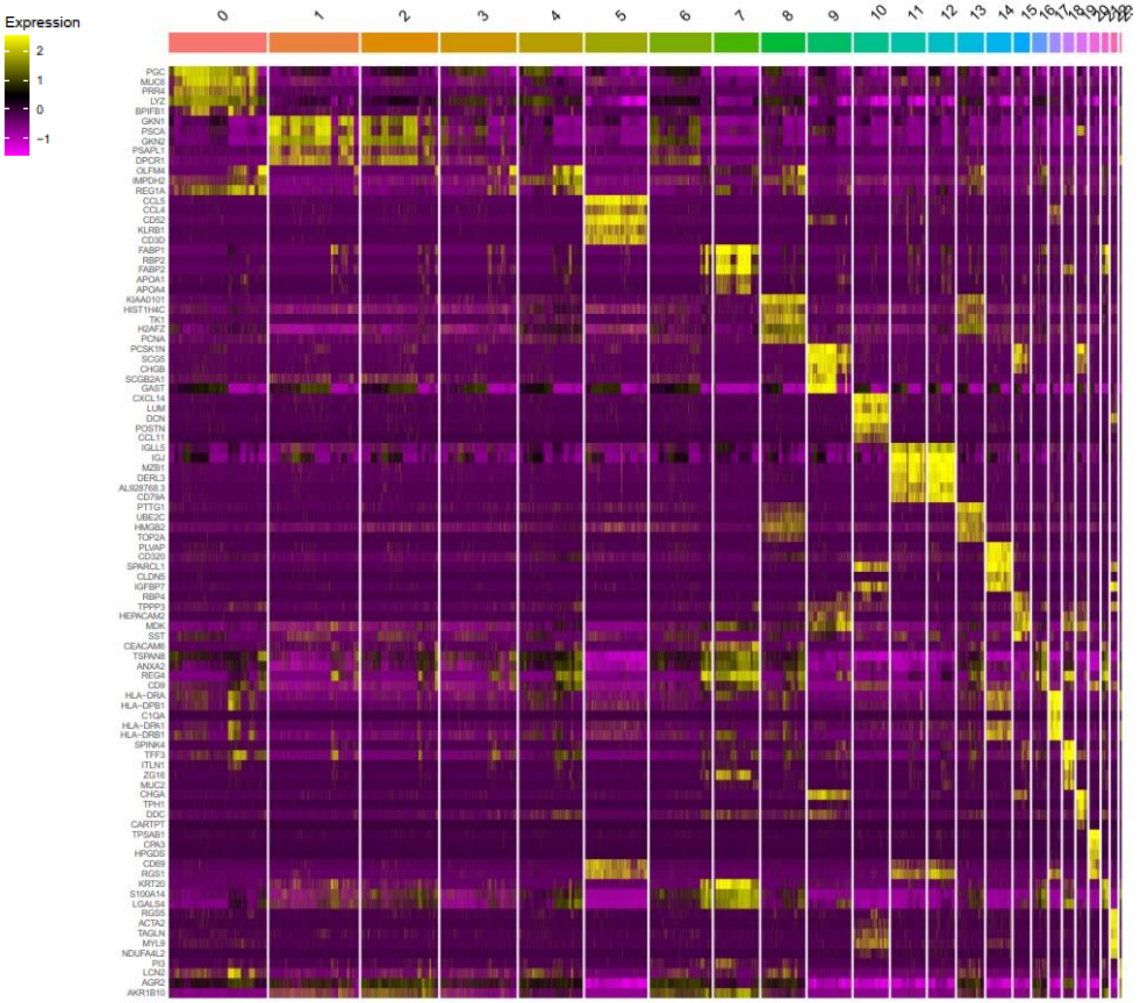



*Figura 17. Stem cells*



La Figura 18. Heatmap de los genes DE por clúster. muestra el *heatmap* de los genes diferencialmente expresados en cada clúster.

*Figura 18. Heatmap de los genes DE por clúster.*




La inspección visual del *heatmap* permitió observar que ciertos tipos celulares estaban distribuidos en varios *clústers*. Como ejemplo, las células secretoras de moco (*pit cells)* aparecieron repartidas en tres *clústers* (1, 2, 6) y las células B en el 11 y 12. Otros tipos, no tan claros, son las *isthmus cells* (8, 13). En el caso de los enterocitos se resolvió fusionar los *clusters* 7, 16, 4, y 21 por los marcadores de enterocitos proximales y distales de referenciados en *the human protein atlas*[41]. Los primeros tres marcadores de cada *cluster*, de células gástricas e intestinales, se muestran en la Tabla 5 y 6. El resto se pueden encontrar en el Anexo-7.

*Tabla 4. Marcadores y clusters: células gástricas.*

| clúster | gene | logFC | pct.1 | pct.2 | p_val_adj |
|---|---|---|---|---|---|
| **0 Chief cells** | | | | | |
| | PGC | 4.62 | 0.908 | 0.627 | 0.00e+00 |
| | MUC6 | 3.93 | 0.797 | 0.246 | 0.00e+00 |
| | PRR4 | 3.29 | 0.552 | 0.054 | 0.00e+00 |
| **1 Pit cells (pit1)** | | | | | |
| | GKN1 | 4.01 | 0.835 | 0.485 | 0.00e+00 |
| | PSCA | 3.10 | 0.744 | 0.350 | 0.00e+00 |
| | GKN2 | 3.09 | 0.854 | 0.393 | 0.00e+00 |
| **2 Pit cells (pit2)** | | | | | |
| | GKN2 | 2.03 | 0.897 | 0.396 | 0.00e+00 |
| | TFF1 | 1.90 | 0.966 | 0.797 | 0.00e+00 |
| | GKN1 | 1.79 | 0.820 | 0.492 | 0.00e+00 |
| **6 Pit cells (pit3)** | | | | | |
| | ALDH3A1 | 1.11 | 0.647 | 0.254 | 0.00e+00 |
| | GPX2 | 1.01 | 0.930 | 0.507 | 0.00e+00 |
| | CYSTM1 | -1.03 | 0.998 | 0.797 | 0.00e+00 |
| **8 Isthmus cells (isth1)** | | | | | |
| | KIAA0101 | 2.37 | 0.834 | 0.069 | 0.00e+00 |
| | HIST1H4C | 2.22 | 0.769 | 0.315 | 0.00e+00 |
| | TK1 | 1.62 | 0.663 | 0.048 | 0.00e+00 |
| **9 G cells** | | | | | |
| | GAST | 6.87 | 0.704 | 0.578 | 1.47e-300 |
| | PCSK1N | 3.66 | 0.797 | 0.066 | 0.00e+00 |
| | SCG5 | 3.39 | 0.851 | 0.044 | 0.00e+00 |
| **13 Isthmus cells (isthm2)** | | | | | |
| | PTTG1 | 2.33 | 0.811 | 0.062 | 0.00e+00 |
| | UBE2C | 2.09 | 0.569 | 0.031 | 0.00e+00 |
| | HIST1H4C | 1.96 | 0.676 | 0.327 | 3.12e-152 |
| **15 D cells** | | | | | |
| | SST | 8.10 | 0.577 | 0.231 | 6.32e-162 |
| | RBP4 | 3.28 | 0.501 | 0.021 | 0.00e+00 |
| | TPPP3 | 2.93 | 0.595 | 0.098 | 0.00e+00 |
| **19 Enteroendocrine cells** | | | | | |
| | CHGA | 5.6 | 0.879 | 0.065 | 0.00e+00 |
| | TPH1 | 4.92 | 0.773 | 0.008 | 0.00e+00 |
| | PCSK1N | 3.90 | 0.780 | 0.094 | 0.00e+00 |

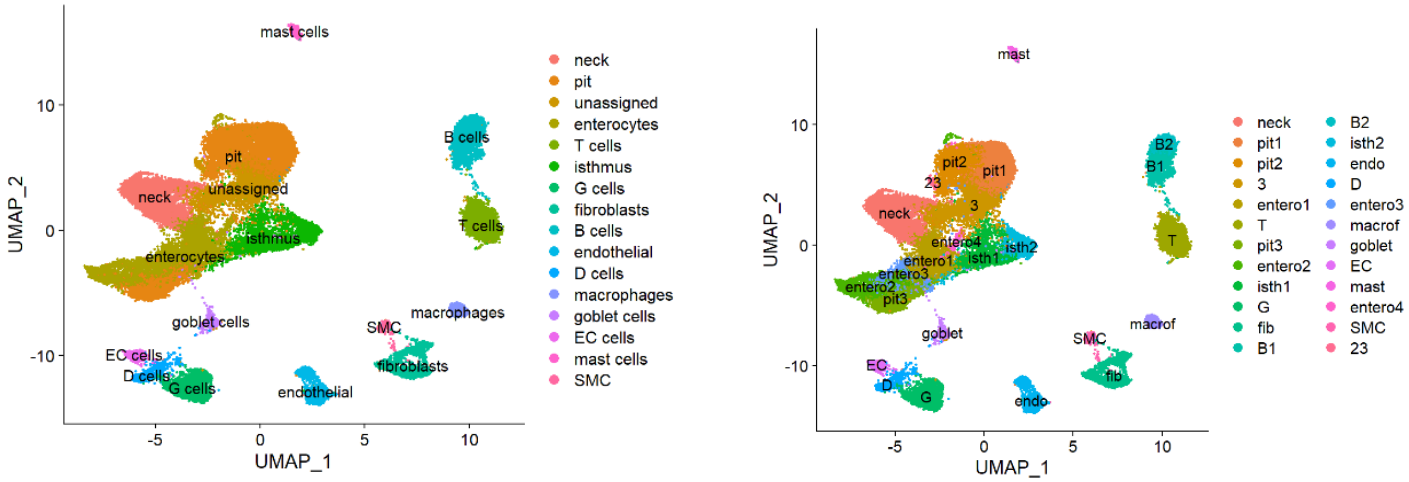
*Tabla 5 Marcadores y clusters: células intestinales.*

| Cluster | gene | logFC | pct.1 | pct.2 | p_val_adj |
|---|---|---|---|---|---|
| **4 (entero1)** | | | | | |
| | OLFM4 | 1.8 | 0.418 | 0.143 | 0.00e+00 |
| | REG1A | 1.2 | 0.554 | 0.342 | 2.05e-82 |
| | RPS2 | 1.0 | 1.000 | 0.941 | 0.00e+00 |
| **7 (entero2)** | | | | | |
| | FABP1 | 5.1 | 0.947 | 0.155 | 0.00e+00 |
| | RBP2 | 4.6 | 0.688 | 0.053 | 0.00e+00 |
| | FABP2 | 4.2 | 0.893 | 0.101 | 0.00e+00 |
| **16 (entero3)** | | | | | |
| | REG4 | 1.8 | 0.650 | 0.298 | 2.01e-110 |
| | TSPAN8 | 1.4 | 0.951 | 0.636 | 1.79e-123 |
| | CEACAM6 | 1.4 | 0.393 | 0.090 | 2.44e-141 |
| **18 Caliciformes (goblet)** | | | | | |
| | SPINK4 | 7.0 | 0.893 | 0.060 | 0.00e+00 |
| | TFF3 | 6.1 | 0.979 | 0.285 | 0.00e+00 |
| | ITLN1 | 5.4 | 0.724 | 0.031 | 0.00e+00 |
| **21 (entero4)** | | | | | |
| | REG4 | 2.6 | 0.992 | 0.299 | 2.52e-179 |
| | FABP1 | 2.5 | 0.962 | 0.189 | 2.13e-240 |
| | S100A14 | 2.2 | 0.989 | 0.476 | 1.14e-139 |

### 5.1.10 Resultado del etiquetado celular.

Decidimos finalizar el etiquetado celular conservando por un lado los 24 *clústers* identificados (Figura 19, panel de la izquierda) y otra uniendo aquellos que son del mismo tipo celular (panel de la derecha).

*Figura 19. Tipos celulares identificados.*

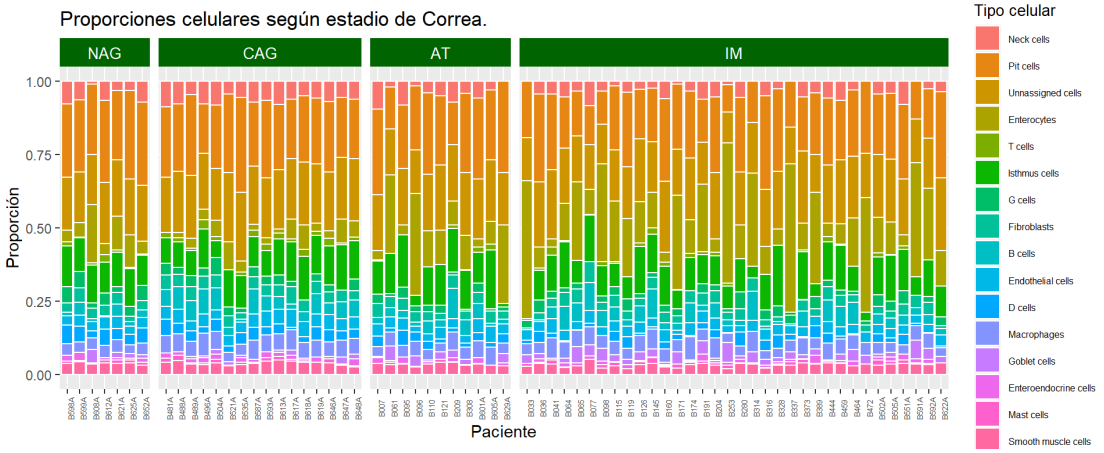



**Sin combinar (24 *clusters*)** **Combinados (16 *clusters*)**

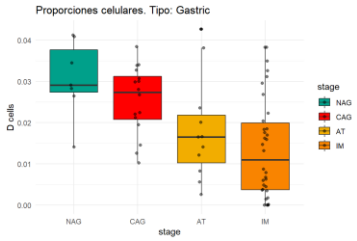
## 5.2 Deconvolución.

Las tras el análisis con Seurat, creamos las matrices de identidades y de corrección para *batch* tipo S para los dos análisis (24 y 16 clusters). Éstos, junto a la matriz de *bulk* data se subieron a los servidores CIBERSORTx. Los análisis estadísticos se encuentran en el Anexo-9.

*Figura 20. Proporciones celulares en la cascada de Correa.*




### Células D




Tests normalidad:

| | QQplots | Shapiro |
|---|---|---|
| NAG | No | Sí |
| CAG | Sí | Sí |
| AT | No | Sí |
| IM | No | No |

Test hipótesis:

D_cells by stage

Kruskal-Wallis chi-2 = 19.845,

df = 3,

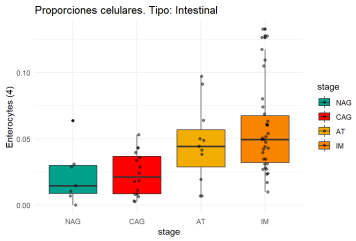
p-value = 0.0001827

Pairwise comparisons using Wilcoxon rank sum exact test

| | NAG | CAG | AT |
|---|---|---|---|
| CAG | 0.0915 | - | - |
| AT | 0.0228 | 0.0915 | - |
| IM | 0.0038 | 0.0056 | 0.1223 |

## P value adjustment method: BH

### Enterocitos




Tests normalidad:

| | QQplots | Shapiro |
|---|---|---|
| NAG | No | No |
| CAG | Sí | Sí |
| AT | No | No |
| IM | No | No |

Test hipótesis:

enterocytes by stage

Kruskal-Wallis chi-squared = 23.815,

df = 3,

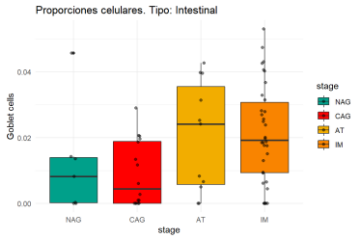
p-value = 2.73e-05

Pairwise comparisons using Wilcoxon rank sum exact test

| | NAG | CAG | AT |
|---|---|---|---|
| CAG | 0.061 | - | - |
| AT | 0.479 | 0.017 | - |
| IM | 0.061 | 2.9e-06 | 0.244 |

# P value adjustment method: BH

### Células caliciformes




Tests normalidad:

| | QQplots | Shapiro |
|---|---|---|
| NAG | No | No |
| CAG | No | No |
| AT | No | No |
| IM | Sí | Sí |

Test hipótesis:

goblet by stage

Kruskal-Wallis chi-squared = 7.9405,

df = 3,

p-value = 0.04726

Pairwise comparisons using Wilcoxon rank sum exact test

| | NAG | CAG | AT |
|---|---|---|---|
| CAG | 0.7329 | - | - |
| AT | 0.4922 | 0.2840 | - |
| IM | 0.2223 | 0.0011 | 0.3440 |

## P value adjustment method: BH

### 5.2.1 La proporción de células D disminuye en la progresión de la cascada de Correa.

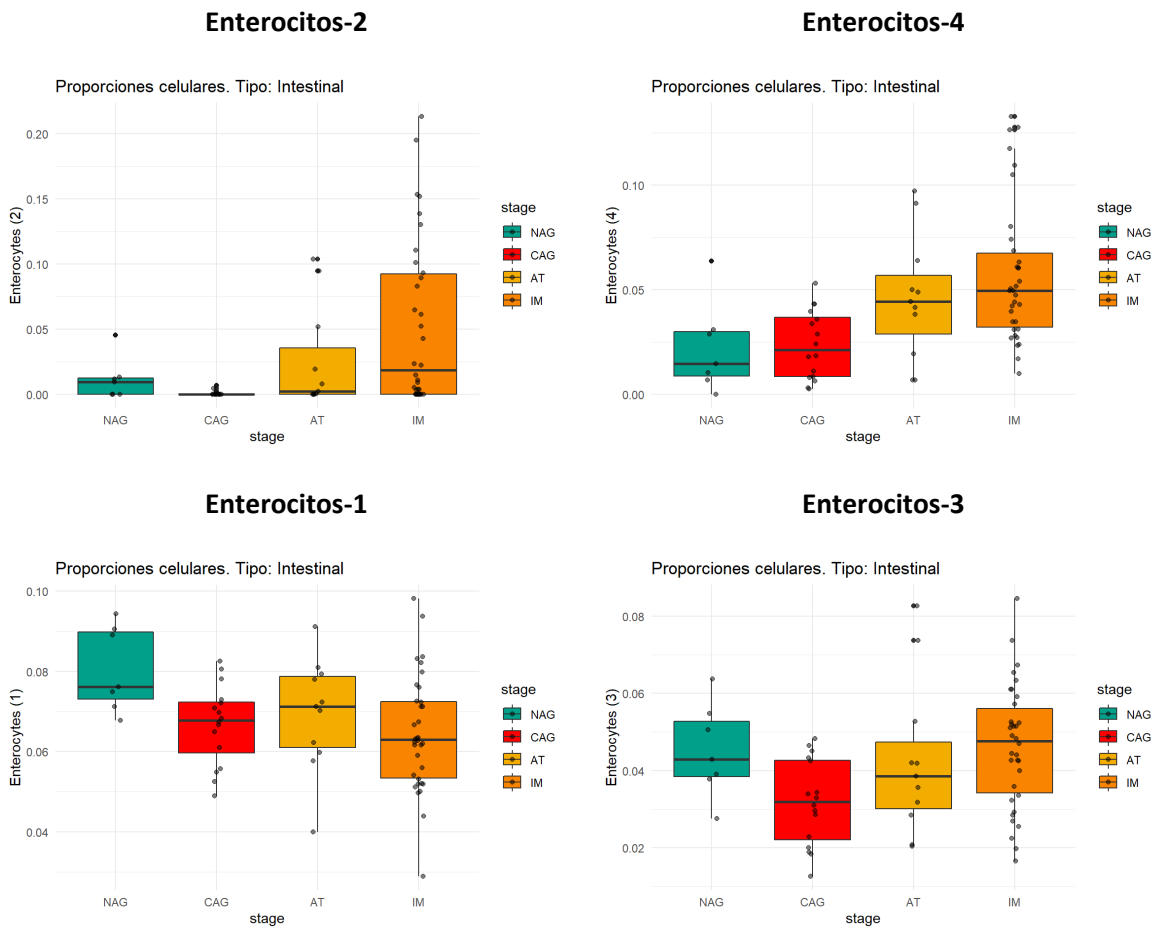
Se muestra a continuación (Figura 20. Proporciones celulares en la cascada de Correa.) el perfil de proporciones celulares (combinadas) para cada estadío de la cascada de Correa. Este perfil es de difícil interpretación. El panel central muestra inferior muestra los diagramas de caja para tres tipos celulares seleccionados. Encontramos diferencias significativas en los tres tipos celulares (células D: $p<0.05$, $\chi^2$= 1.8E-4; enterocitos: p= 2.7E-5, $\chi^2$= 23.8, goblet cels: p= 0.047, $\chi^2$= 7.94, Kruskal-Wallis test). En comparación con NAG, las células D disminuyeron significativamente en las lesiones precursoras (NAG vs AT, p= 0.03; NAG vs IM, p= 0.004; Wilcoxon rank sum exact test). El efecto contrario se observó para los enterocitos, que aumentó su proporción respecto a CAG (CAG vs AT, p= 0.017; CAG vs IM, p= 2.9E-6, Wilcoxon rank sum exact test), pero no a NAG. Solo se observó un aumento significativo de las células caliciformes entre los pacientes con CAG e IM (p= 0.001, Wilcoxon rank sum exact test).

### 5.2.2 Las subpoblaciones 1 a 4 de enterocitos muestran proporciones diferentes en la cascada de Correa.

Los enterocitos es un grupo constituido por 4 clusters que etiquetamos como tal. El análisis de las proporciones por subgrupo muestra diferencias para cada uno de ellos (enterocitos-2: p= 1.5E-3, $\chi^2$= 15.4; enterocitos-4: p= 5.5E-4, $\chi^2$= 17.5; Kruskal-Wallis rank sum test)(enterocitos-1: F= 2.973, p= 0.0382; enterocitos-3: F= 3.353, p= 0.0242, AOV).

La proporción de enterocitos-4 aumentó significativamente en la AT y la IM respecto a CAG (CAG vs AT: p=0.0397; CAG sv IM: p= 0.0012, Wilcoxon rank sum exact test). Los enterocitos-2 aumentaron en la IM respecto a NAG y CAG (NAG vs IM: p=0.0178; CAG sv IM: p= 0.0007, Wilcoxon rank sum exact test). La proporción de enterocitos-1 disminuyó significativamente respecto a NAG (NAG vs IM: p= 0.0232. Tukey) y los enterocitos-3 aumentaron en CAG vs IM (p= 0.0141, Tukey). Los boxplots se recogen en la Figura 21.

*Figura 21. Subpoblaciones de enterocitos.*



# 6 Discusión

Las lesiones precursoras del cáncer gástrico, especialmente la metaplasia intestinal, son un factor de riesgo para la aparición del cáncer gástrico. Como hemos visto, estas lesiones se caracterizan por la sustitución, en el tejido gástrico, de células secretoras por células con fenotipo intestinal, de función absortiva. Existe en la literatura debate sobre qué tipo celular gástrico es origen de estas células con fenotipo intestinal y sobre los determinantes de su aparición y progresión. La gran variedad de tipos celulares presentes en la mucosa gástrica de los pacientes con lesiones precursoras hace que los estudios -omicos se interpreten de forma parcial, ya que no se puede atribuir a un tipo celular la alteración en la expresión de determinados genes.

Aquí hemos comprobado, usando únicamente datos de expresión génica, que efectivamente durante la progresión de la cascada de Correa existe un aumento de células metaplásicas intestinales, concretamente enterocitos y células caliciformes (Figura 20). Además, ajustando parámetros de granulometría, hemos visto que los enterocitos se pueden asociar a cuatro *clusters* con marcadores diferentes (Tabla 5). El significado de estos resultados es incierto. Existen al menos dos tipos de metaplasia intestinal, la completa y la incompleta[42]. La extensión de la metaplasia incompleta justificaría el seguimiento de estos pacientes. Ambos tipos se diferencian en la expresión de ciertas mucinas y en su similitud a al epitelio intestinal proximal o distal. Nuestros subtipos también podrían ser células en distintas fases del ciclo celular. Además, hemos visto que las células D, secretoras de somatostatina, desaparecen durante la progresión de la cascada de Correa (Figura 20). Existe escasa bibliografía al respecto, pero este resultado estaría en concordancia[43,44]. Se sabe además que una de los efectos de la infección por *H. pylori* es la disminución de la expresión de somatostatina y un aumento de gastrina que llevan a una hipersecreción ácida por parte del cuerpo gástrico[45].

Este análisis no está exento de limitaciones. Uno de ellos refiere a la deconvolución. Hasta donde he podido llegar, las técnicas de deconvolución asignan siempre un tipo celular. Esto puede ser una limitación ya que, en nuestro caso, la matriz de identidades se ha basado en 13 pacientes que podrían no ser buenos representantes de los pasos en la cascada de Correa. Otra limitación refiere al análisis *single-cell* para obtener la matriz de identidades. En la mayoría de pasos (QC, normalización, resolución…) hay que tomar decisiones que van a afectar a los *clusters* que vamos

a obtener. Una de las decisiones importantes es la interpretación asignación de *clusters*. Aquí hemos usado en parte la tabla de marcadores publicada por Zhang *et al*.[5] y por otros[33], pero no hemos intentado comparar nuestros tipos celulares con los obtenidos por los autores del artículo. Hemos preferido hacerlo a ciegas. El estudio comparativo no era un objetivo del presente proyecto. Por último, las biopsias analizadas por Zhang *et al*.[5] son antrales por lo que hay una parte de tipos celulares (concretamente las células parietales, cuya pérdida es la responsable de la atrofia gástrica) que no aparecen en la matriz de identidades. Este punto lo hemos podido comprobar por la escasa presencia de marcadores de estas células (ATP4A, ATP4B, GIF). Relacionado con este punto sería muy interesante analizar qué marcadores comparten y difieren las células metaplasicas antrales y corporales.

# 7 Conclusiones

## 7.1 Conclusiones

*Lecciones:*

He aprendido que los análisis *single-cell* tienen un componente subjetivo importante. La toma de decisiones en cada paso puede alterar significativamente los resultados de las matrices de identidades y, consecuentemente, del proceso de deconvolución.

La metodología utilizada me ha permitido aplicar conocimientos adquiridos a lo largo del máster, como el análisis multivariante, la reducción de dimensionalidad, diseño experimental, entre otros.

*Objetivos:*

Tanto el objetivo general como los objetivos específicos se han alcanzado. Cabría realizar varias acciones para confirmar estos resultados (se recogen en la siguiente sección).

## 7.2 Líneas de futuro

Hay varios puntos:

- Los análisis *single-cell* presentan muchos puntos de optimización de los pasos. Sería conveniente valorar al menos alguno de ellos. Por ejemplo, para el control de calidad, el *clustering*, la normalización hay distintas maneras de abordarlo. De hecho se podría plantear con pipeComp[36], un *framework* para la comparación de

pipelines que van desde la obtención de la matriz de cuentas hasta el etiquetado de los *clusters*.

- Un punto importante es la asignación de *clusters*. Sería recomendable reevaluarlos e interpretarlos con ayuda de patólogos especializados en el sistema digestivo.
- Sería recomendable analizar estados celulares como la división celular. El epitelio gástrico está en continua renovación.
- También existen varios métodos de deconvolución, al menos dos se podrían evaluar y comparar con CIBERSORTx. Uno, AutoGenes[46], se probó sin éxito (por mis limitaciones en Python). El otro, SCDC[47], escrito en R, no hubo tiempo. Ambos autores fueron contactados para que nos confirmaran que sus métodos eran compatibles con los datos de microarrays.
- Por falta de tiempo no se ha explicado una de las funciones de CIBERSORTx, que es que la expresión diferencial se puede corregir por las proporciones celulares.
- Sería conveniente, usando las matrices obtenidas en este trabajo, analizar las proporciones de otros estudios con pacientes con LPCG, como el de Nookaev *et al*.[12] y comprobar la estabilidad de nuestras proporciones.

### 7.3 Seguimiento de la planificación

La planificación se ha visto modificada temporalmente respecto a la idea inicial. Esta modificación no ha requerido de cambios en las tareas. Ciertas tareas, que son críticas en los análisis *single cell,* se han alargado. Estas tareas son el control de calidad y el etiquetado manual de *clusters*. Además, el análisis single-cell requiere poder computacional. Sólo el objeto Seurat requería de 1.8Gb de memoria. CIBERSORTx, aunque de acceso web, requería además de varias horas de procesado.

Con todo se ha podido completar el análisis a tiempo.

# 8 Glosario

**Metaplasia**: es un cambio en el fenotipo celular. Una célula se transforma por trasndiferenciación en otra. En el estómago la metaplasia intestinal indica que hay células gástricas (típicamente secretoras) que se han diferenciado a intestinales (típicamente absortivas). En la metaplasia gástrica encontramos céluals gástricas en tejido intestinal.

**Atrofia gástrica**: disminución de parte de la función secretora del estómago por perdida de células.

**Actividad**: se refiere a la infiltración de células polimorfonucleares (neutrófilos) en el tejido gástrico. Es un marcador de infección.

Acrónimos:

AT: atrofia LLS: linear least squares.

CAG: Chronic-active gastritis

IM: metaplasia intestinal.

MAD: median absolute deviation

NAG: Non-active gastritis. Gastritis no activa.

NNMF: non-negative matrix factorization

PCA: principal component analysis.

SVR: support vector regression.

# 9 Bibliografía


(1) Tang, F.; Barbacioru, C.; Wang, Y.; Nordman, E.; Lee, C.; Xu, N.; Wang, X.; Bodeau, J.; Tuch, B. B.; Siddiqui, A.; Lao, K.; Surani, M. A. MRNA-Seq Whole-Transcriptome Analysis of a Single Cell. *Nat. Methods* **2009**, *6* (5), 377–382. https://doi.org/10.1038/nmeth.1315.

(2) Yu, T.; Scolnick, J. Complex Biological Questions Being Addressed Using Single Cell Sequencing Technologies. *SLAS Technol.* **2021**, S2472-6303(21)00013-3. https://doi.org/10.1016/j.slast.2021.10.013.

(3) Curtius, K.; Wright, N. A.; Graham, T. A. Evolution of Premalignant Disease. *Cold Spring Harb. Perspect. Med.* **2017**, *7* (12), a026542. https://doi.org/10.1101/cshperspect.a026542.

(4) Lario, S.; Ramírez-Lázaro, M. J.; González-Lahera, A.; Lavín, J. L.; Vila-Casadesús, M.; Quílez, M. E.; Brunet-Vega, A.; Lozano, J. J.; Aransay, A. M.; Calvet, X. Cross-Sectional Study of Human Coding- and Non-Coding RNAs in Progressive Stages of Helicobacter Pylori Infection. *Sci. Data* **2020**, *7* (1), 296. https://doi.org/10.1038/s41597-020-00636-6.

(5) Zhang, P.; Yang, M.; Zhang, Y.; Xiao, S.; Lai, X.; Tan, A.; Du, S.; Li, S. Dissecting the Single-Cell Transcriptome Network Underlying Gastric Premalignant Lesions and Early Gastric Cancer. *Cell Rep.* **2019**, *27* (6), 1934-1947.e5. https://doi.org/10.1016/j.celrep.2019.04.052.

(6) Avila Cobos, F.; Vandesompele, J.; Mestdagh, P.; De Preter, K. Computational Deconvolution of Transcriptomics Data from Mixed Cell Populations. *Bioinforma. Oxf. Engl.* **2018**, *34* (11), 1969–1979. https://doi.org/10.1093/bioinformatics/bty019.

(7) Zhang, M.; Hu, S.; Min, M.; Ni, Y.; Lu, Z.; Sun, X.; Wu, J.; Liu, B.; Ying, X.; Liu, Y. Dissecting Transcriptional Heterogeneity in Primary Gastric Adenocarcinoma by Single Cell RNA Sequencing. *Gut* **2021**, *70* (3), 464–475. https://doi.org/10.1136/gutjnl-2019-320368.

(8) *Seurat - Guided Clustering Tutorial*. https://satijalab.org/seurat/articles/pbmc3k_tutorial.html (accessed 2022-03-02).

(9) Thrift, A. P.; Nguyen, T. H. Gastric Cancer Epidemiology. *Gastrointest. Endosc. Clin. N. Am.* **2021**, *31* (3), 425–439. https://doi.org/10.1016/j.giec.2021.03.001.

(10) Correa, P.; Haenszel, W.; Cuello, C.; Tannenbaum, S.; Archer, M. A Model for Gastric Cancer Epidemiology. *Lancet Lond. Engl.* **1975**, *2* (7924), 58–60. https://doi.org/10.1016/s0140-6736(75)90498-5.

(11) Shah, S. C.; Piazuelo, M. B.; Kuipers, E. J.; Li, D. AGA Clinical Practice Update on the Diagnosis and Management of Atrophic Gastritis: Expert Review. *Gastroenterology* **2021**, *161* (4), 1325-1332.e7. https://doi.org/10.1053/j.gastro.2021.06.078.

(12) Nookaew, I.; Thorell, K.; Worah, K.; Wang, S.; Hibberd, M. L.; Sjövall, H.; Pettersson, S.; Nielsen, J.; Lundin, S. B. Transcriptome Signatures in Helicobacter Pylori-Infected Mucosa Identifies Acidic Mammalian Chitinase Loss as a Corpus Atrophy Marker. *BMC Med. Genomics* **2013**, *6*, 41. https://doi.org/10.1186/1755-8794-6-41.

(13) Companioni, O.; Sanz-Anquela, J. M.; Pardo, M. L.; Puigdecanet, E.; Nonell, L.; García, N.; Parra Blanco, V.; López, C.; Andreu, V.; Cuatrecasas, M.; Garmendia, M.; Gisbert, J. P.; Gonzalez, C. A.; Sala, N. Gene Expression Study and Pathway Analysis of Histological Subtypes of Intestinal Metaplasia That Progress to Gastric Cancer. *PloS One* **2017**, *12* (4), e0176043. https://doi.org/10.1371/journal.pone.0176043.

(14) Meyer, A. R.; Goldenring, J. R. Injury, Repair, Inflammation and Metaplasia in the Stomach. *J. Physiol.* **2018**, *596* (17), 3861–3867. https://doi.org/10.1113/JP275512.

(15) Kinoshita, H.; Hayakawa, Y.; Niu, Z.; Konishi, M.; Hata, M.; Tsuboi, M.; Hayata, Y.; Hikiba, Y.; Ihara, S.; Nakagawa, H.; Hirata, Y.; Wang, T. C.; Koike, K. Mature Gastric Chief Cells Are Not Required for the Development of Metaplasia. *Am. J. Physiol. Gastrointest. Liver Physiol.* **2018**, *314* (5), G583–G596. https://doi.org/10.1152/ajpgi.00351.2017.

(16) Lee, H.-J.; Nam, K. T.; Park, H. S.; Kim, M. A.; Lafleur, B. J.; Aburatani, H.; Yang, H.-K.; Kim, W. H.; Goldenring, J. R. Gene Expression Profiling of Metaplastic Lineages Identifies CDH17 as a Prognostic Marker in Early Stage Gastric Cancer. *Gastroenterology* **2010**, *139* (1), 213-225.e3. https://doi.org/10.1053/j.gastro.2010.04.008.

(17) Wang, H.; Owens, J. D.; Shih, J. H.; Li, M.-C.; Bonner, R. F.; Mushinski, J. F. Histological Staining Methods Preparatory to Laser Capture Microdissection Significantly Affect the Integrity of the Cellular RNA. *BMC Genomics* **2006**, *7*, 97. https://doi.org/10.1186/1471-2164-7-97.

(18) Venet, D.; Pecasse, F.; Maenhaut, C.; Bersini, H. Separation of Samples into Their Constituents Using Gene Expression Data. *Bioinforma. Oxf. Engl.* **2001**, *17 Suppl 1*, S279-287. https://doi.org/10.1093/bioinformatics/17.suppl_1.s279.

(19) Avila Cobos, F.; Alquicira-Hernandez, J.; Powell, J. E.; Mestdagh, P.; De Preter, K. Benchmarking of Cell Type Deconvolution Pipelines for Transcriptomics Data. *Nat. Commun.* **2020**, *11*, 5650. https://doi.org/10.1038/s41467-020-19015-1.

(20) Jin, H.; Liu, Z. A Benchmark for RNA-Seq Deconvolution Analysis under Dynamic Testing Environments. *Genome Biol.* **2021**, *22* (1), 102. https://doi.org/10.1186/s13059-021-02290-6.

(21) Newman, A. M.; Steen, C. B.; Liu, C. L.; Gentles, A. J.; Chaudhuri, A. A.; Scherer, F.; Khodadoust, M. S.; Esfahani, M. S.; Luca, B. A.; Steiner, D.; Diehn, M.; Alizadeh, A. A. Determining Cell Type Abundance and Expression from Bulk Tissues with Digital Cytometry. *Nat. Biotechnol.* **2019**, *37* (7), 773–782. https://doi.org/10.1038/s41587-019-0114-2.

(22) Wagner, A.; Regev, A.; Yosef, N. Revealing the Vectors of Cellular Identity with Single-Cell Genomics. *Nat. Biotechnol.* **2016**, *34* (11), 1145–1160. https://doi.org/10.1038/nbt.3711.

(23) Ogbeide, S.; Giannese, F.; Mincarelli, L.; Macaulay, I. C. Into the Multiverse: Advances in Single-Cell Multiomic Profiling. *Trends Genet. TIG* **2022**, S0168-9525(22)00077-4. https://doi.org/10.1016/j.tig.2022.03.015.

(24) *Chromium Single Cell 3' Reagent Kits User Guide (v2 Chemistry) -User Guide -Library Prep -Single Cell Gene Expression -Official 10x Genomics Support*. https://support.10xgenomics.com/single-cell-gene-expression/library-prep/doc/user-guide-chromium-single-cell-3-reagent-kits-user-guide-v2-chemistry (accessed 2022-05-23).

(25) Germain, P.-L.; Lun, A.; Macnair, W.; Robinson, M. D. Doublet Identification in Single-Cell Sequencing Data Using ScDblFinder. F1000Research September 28, 2021. https://doi.org/10.12688/f1000research.73600.1.

(26) Stuart, T.; Butler, A.; Hoffman, P.; Hafemeister, C.; Papalexi, E.; Mauck, W. M.; Hao, Y.; Stoeckius, M.; Smibert, P.; Satija, R. Comprehensive Integration of Single-Cell Data. *Cell* **2019**, *177* (7), 1888-1902.e21. https://doi.org/10.1016/j.cell.2019.05.031.

(27) Linderman, G. C. Dimensionality Reduction of Single-Cell RNA-Seq Data. *Methods Mol. Biol. Clifton NJ* **2021**, *2284*, 331–342. https://doi.org/10.1007/978-1-0716-1307-8_18.

(28) Tran, H. T. N.; Ang, K. S.; Chevrier, M.; Zhang, X.; Lee, N. Y. S.; Goh, M.; Chen, J. A Benchmark of Batch-Effect Correction Methods for Single-Cell RNA Sequencing Data. *Genome Biol.* **2020**, *21* (1), 12. https://doi.org/10.1186/s13059-019-1850-9.

(29) Korsunsky, I.; Millard, N.; Fan, J.; Slowikowski, K.; Zhang, F.; Wei, K.; Baglaenko, Y.; Brenner, M.; Loh, P.; Raychaudhuri, S. Fast, Sensitive, and Accurate Integration of Single Cell Data with Harmony. *Nat. Methods* **2019**, *16* (12), 1289–1296. https://doi.org/10.1038/s41592-019-0619-0.

(30) Qi, R.; Ma, A.; Ma, Q.; Zou, Q. Clustering and Classification Methods for Single-Cell RNA-Sequencing Data. *Brief. Bioinform.* **2020**, *21* (4), 1196–1208. https://doi.org/10.1093/bib/bbz062.

(31) Chipster Tutorials. *7. Clustering of ScRNA-Seq Data*; 2019.

(32) Clarke, Z. A.; Andrews, T. S.; Atif, J.; Pouyabahar, D.; Innes, B. T.; MacParland, S. A.; Bader, G. D. Tutorial: Guidelines for Annotating Single-Cell Transcriptomic Maps Using Automated and Manual Methods. *Nat. Protoc.* **2021**, *16* (6), 2749–2764. https://doi.org/10.1038/s41596-021-00534-0.

(33) Busslinger, G. A.; Weusten, B. L. A.; Bogte, A.; Begthel, H.; Brosens, L. A. A.; Clevers, H. Human Gastrointestinal Epithelia of the Esophagus, Stomach, and Duodenum Resolved at Single-Cell Resolution. *Cell Rep.* **2021**, *34* (10), 108819. https://doi.org/10.1016/j.celrep.2021.108819.

(34) Xiao, S.; Zhou, L. Gastric Stem Cells: Physiological and Pathological Perspectives. *Front. Cell Dev. Biol.* **2020**, *8*, 571536. https://doi.org/10.3389/fcell.2020.571536.

(35) Lario, S.; Ramírez-Lázaro MJ, M. J.; González-Lahera, A.; Lavín, J. L.; Vila-Casadesus, M.; Quilez, M. E.; Brunet-Vega, A.; Lozano, J. J.; Aransay, A. M.; Calvet, X. MRNA Microarray Data from Patients with Non-Active Gastritis, Chronic Active Gastritis and Precursor Lesions of Gastric Cancer (E-MTAB-8889). *ArrayExpress*. 2020, p https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-8889/.

(36) Germain, P.-L.; Sonrel, A.; Robinson, M. D. PipeComp, a General Framework for the Evaluation of Computational Pipelines, Reveals Performant Single Cell RNA-Seq Preprocessing Tools. *Genome Biol.* **2020**, *21* (1), 227. https://doi.org/10.1186/s13059-020-02136-7.

(37) Andrews, T. S.; Kiselev, V. Y.; McCarthy, D.; Hemberg, M. Tutorial: Guidelines for the Computational Analysis of Single-Cell RNA Sequencing Data. *Nat. Protoc.* **2021**, *16* (1), 1–9. https://doi.org/10.1038/s41596-020-00409-w.

(38) Hippen, A. A.; Falco, M. M.; Weber, L. M.; Erkan, E. P.; Zhang, K.; Doherty, J. A.; Vähärautio, A.; Greene, C. S.; Hicks, S. C. MiQC: An Adaptive Probabilistic Framework for Quality Control of Single-Cell RNA-Sequencing Data. *PLoS Comput. Biol.* **2021**, *17* (8), e1009290. https://doi.org/10.1371/journal.pcbi.1009290.

(39) Steen, C. B.; Liu, C. L.; Alizadeh, A. A.; Newman, A. M. Profiling Cell Type Abundance and Expression in Bulk Tissues with CIBERSORTx. *Methods Mol. Biol. Clifton NJ* **2020**, *2117*, 135–157. https://doi.org/10.1007/978-1-0716-0301-7_7.

(40) Sáenz, J. B.; Mills, J. C. Acid and the Basis for Cellular Plasticity and Reprogramming in Gastric Repair and Cancer. *Nat. Rev. Gastroenterol. Hepatol.* **2018**, *15* (5), 257–273. https://doi.org/10.1038/nrgastro.2018.5.

(41) *The Human Protein Atlas*. https://www.proteinatlas.org/ (accessed 2022-05-26).

(42) Correa, P.; Piazuelo, M. B.; Wilson, K. T. Pathology of Gastric Intestinal Metaplasia: Clinical Implications. *Am. J. Gastroenterol.* **2010**, *105* (3), 493–498. https://doi.org/10.1038/ajg.2009.728.

(43) Torres, A. J.; Ortega, L.; Blanco, J.; Fernandez-Durango, R.; Hernandez, F.; Suarez, A.; Cuberes, R.; Sanz, J.; Balibrea, J. L. Antral Gastrin-Producing G-Cells and Somatostatin-Producing D-Cells in Peptic Ulcer. *Virchows Arch. A Pathol. Anat. Histopathol.* **1986**, *410* (3), 165–171. https://doi.org/10.1007/BF00710821.

(44) Liu, Y.; Vosmaer, G. D. C.; Tytgat, G. N. J.; Xiao, S.-D.; Ten Kate, F. J. W. Gastrin (G) Cells and Somatostatin (D) Cells in Patients with Dyspeptic Symptoms: Helicobacter Pylori Associated and Non-Associated Gastritis. *J. Clin. Pathol.* **2005**, *58* (9), 927–931. https://doi.org/10.1136/jcp.2003.010710.

(45) Calam, J. Helicobacter Pylori Modulation of Gastric Acid. *Yale J. Biol. Med.* **1999**, *72* (2–3), 195–202.

(46) Aliee, H.; Theis, F. J. AutoGeneS: Automatic Gene Selection Using Multi-Objective Optimization for RNA-Seq Deconvolution. *Cell Syst.* **2021**, *12* (7), 706-715.e4. https://doi.org/10.1016/j.cels.2021.05.006.

(47) Dong, M.; Thennavan, A.; Urrutia, E.; Li, Y.; Perou, C. M.; Zou, F.; Jiang, Y. SCDC: Bulk Gene Expression Deconvolution by Multiple Single-Cell RNA Sequencing References. *Brief. Bioinform.* **2021**, *22* (1), 416–427. https://doi.org/10.1093/bib/bbz166.

# 10 Anexos

**Anexo-1.** Cronograma.

**Anexo-2**. Rmarkdown Importación datos-código R.

**Anexo-3a**. Control de calidad-código R.

**Anexo-3b**. Resultados (plots) filtrado MAD.

**Anexo-3c.** Resultados (plots) filtrado miQC.

**Anexo-4.** Normalización, UMAP para comprobar Control de calidad.

**Anexo-5**. QC definitivo-código R.

**Anexo-6**. Resultados (plots) Resolution y Batch effects-código R.

**Anexo-7.** Tablas FC inmunitario-estroma-unassigned**.**

**Anexo-8**. Rmarkdown-FindMarkers (violin-featurePlots).

**Anexo-9.** Rmarkown-Estadística descriptiva, AOV y Kruskal- código R.

**Anexo-10.** Entrada y salida de CIBERSORTx (captura).