Técnicas de Reducción de Dimensionalidad para el Estudio de Perfiles de Expresión Genética

Romina Yalovetzky

Tesis de Licenciatura en Ciencias Físicas

Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

Junio 2019

- TEMA: Análisis de paisajes transcripcionales de célula única para el estudio de procesos de desarrollo y diferenciación celular con aplicaciones a neurogénesis adulta.
- ALUMNO: Romina Yalovetzky.
- LU: 124/14.
- DIRECTOR DEL TRABAJO: Dr. Ariel Chernomoretz.
- FECHA DE INICIACIÓN: Septiembre 2018.
- FECHA DE FINALIZACIÓN: Junio 2019.
- FECHA DE EXAMEN:
- INFORME FINAL APROBADO POR:

Autor	Jurado
Director	Jurado
Profesor de Tesis de Licenciatura	Jurado

Nada en este mundo debe ser temido, debe ser solamente comprendido. Ahora es el momento de comprender más, para temer menos.

Marie Curie

UNIVERSIDAD DE BUENOS AIRES

Resumen

Facultad de Ciencias Exactas y Naturales

Departamento de Física

Licenciatura en Ciencias Físicas

Los experimentos de secuenciación de célula única permiten ganar conocimiento sobre la biología del desarrollo a partir del acceso cuantitativo a perfiles transcripcionales a nivel de célula única. A diferencia de otras técnicas de secuenciación, esta conlleva un análisis de volumenes de datos sin precedentes. También, esta alta resolución a nivel genómico sufre una limitación, inherente a la técnica, que es una ineficiente captura del RNAm que se representa en el reporte de una alta tasa de expresiones nulas (eventos llamados *dropouts*).

En esta tesis estudiamos técnicas basadas en teoría de redes complejas y en minería de datos para reducir la cantidad de datos y limitarnos a aquellos que son realmente relevantes para la obtención de conclusiones sobre los procesos biológicos. En particular, trabajamos con una técnica de reducción de dimensionalidad en la que se seleccionan genes de acuerdo a la variabilidad de los niveles de expresión sobre vecindades de células definidas en un grafo que modela el espacio de estados celulares.

Cuantificamos la efectividad de esta técnica utilizando dos modelos de datos típicos extraidos en estos experimentos. El primero de estos simula la evolución dinámica de la expresión de conjuntos de genes que evolucionan colectivamente por estar asociados a procesos biológicos. Complejizamos esta dinámica biológica simulando genes que están involucrados en más de un proceso en un segundo modelo. A su vez trabajamos sobre datos experimentales que se corresponden con un trabajo en el que estudiaron la relación entre el desarollo y la neurogénesis en el giro dentado del hipocampo en ratones para desarrollar nuevos criterios de selección de genes que estudiamos en profundidad.

Por otro lado, con el objetivo de reducir los efectos de la falta de información por la ineficiente captura del RNAm (*dropouts*), estudiamos una técnica de imputación de expresión de genes en células. Analizamos el funcionamiento de esta técnica y también cuantificamos su efectividad trabajando con el modelo más complejo sobre el cual simulamos los eventos *dropout*.

Agradecimientos

Quisiera comenzar agradeciendo a mis padres por apoyar incondicionalmente todos mis sueños y proyectos. Su ayuda ha hecho posible que pueda haber alcanzado este objetivo tan importante en mi vida.

Gracias Dr. Ariel Chernomoretz, porque sin su ayuda y dedicación no hubiera sido posible esta tesis. Quiero agradecerle por abrirme las puertas de su laboratorio y por todo lo que aprendí.

Agadezco a los profesores de la Facultad de Cs. Exactas y Naturales de la UBA. Especialmente al Dr. Mitnik, al Dr. Otero y Garzón y al Dr. Marzocca que me acompañaron y enseñaron tanto.

Y finalmente, gracias a mis amigos de la facultad, en especial a Sol y Mailen, por haber transitado juntos esta hermosa carrera y por haber hecho que sean de los mejores años de mi vida. Gracias a Felipe por haberme acompañado con tanto amor en todo este proceso.

Índice general

Capítulo 1

Introducción

A lo largo de su desarrollo, una célula necesita realiza numerosos procesos biológicos que son llevados a cabo por macromoléculas con funciones específicas, generalmente proteínas. Cada segmento de ADN que contiene la información necesaria para sintetizar alguna proteína se denomina gen. Cuando una célula necesita generar una cierta proteína, el gen que contiene la información para codificarla recibe una orden para ser expresado. Es decir, la célula copia el segmento de ADN correspondiente a ese gen, en un segmento de ARN mensajero (ARNm) que es un ácido ribonucléico de cadena simple que llevará la información contenida en el gen fuera del núcleo para que la proteína pueda ser sintetizada. Este proceso de copiado del ADN a ARNm se conoce como transcripción y es regulado por proteínas especiales, llamadas factores de transcripción. Las mismas tienen la capacidad de ligarse a regiones del ADN, promoviendo, iniciando o suprimiendo la transcripción.

Una vez que el ARNm es exportado del núcleo al citoplasma, la información contenida en él debe ser codificada en una secuencia de aminoácidos para que la proteína correspondiente sea sintetizada. Este proceso se llama traducción y tiene lugar en complejos proteicos llamados ribosomas. Este proceso completo es lo que se conoce como Dogma Central de la Biología Molecular.

La cantidad de copias que se realizan de cada gen, es decir la cantidad de ARNm que se produce en el proceso de transcripción para un dado gen, se denomina nivel de expresión y el mismo depende de la cantidad de proteínas que la célula necesite producir en cada momento. El conjunto de números que indica el nivel de expresión de todos los genes activos en la célula se conoce como perfil de expresión transcripcional de esa célula.

La idea que subyace en el análisis de perfiles expresión génica es que el estado biológico de una célula está representado por su perfil de expresión transcripcional, el cual se refleja en la concentración de moléculas de ARNm que están siendo exportadas del núcleo.

Gracias a la tecnología que disponemos hoy en experimentos de secuenciación de célula única (sc-RNASeq, por *single cell RNA sequencing*, en inglés) es posible observar simultáneamente los niveles de expresión de miles de genes que tiene lugar dentro de decenas/miles de células, de a una por vez (Eberwine et al., 2014) (Nawy, 2013). Así, los datos scRNASeq nos permiten monitorear estados y cambios transcripcionales con un detalle sin precedentes que ha permitido ganar conocimiento sobre múltiples aspectos moleculares de la biología del desarrollo.

Al mismo tiempo, ese nivel de detalle es el que conlleva múltiples desafíos relacionados con el volumen y complejidad de los datos relevados. Técnicas específicas para normalizar y pre-procesar datos, métodos de visualización y reconocimiento de estructuras en espacios de alta dimensionalidad actualmente se encuentran en plena evolución y desarrollo (Rostom et al., 2017) (Rizvi et al., 2017).

En los experimentos scRNASeq se obtienen "fotografías" del desarrollo biológico del tejido estudiado a partir de medir el estado transcripcional para el orden de 30000 genes en alrededor de miles o decenas de miles de células diferentes. Sin embargo, se asume como hipótesis de trabajo que no todos los genes sobre los cuales se recaba información se expresan independientemente sino que su expresión se coordina para llevar a cabo funciones biológicas de mayor escala dentro de la célula. Esto conduce a la hipótesis de que el estado celular y la biológia plausible y relevante, puede ser descripta en realidad sobre una variedad de menor dimensión embebida en el *espacio de estados celulares*. Esta es la *Variedad Biológicamente Relevante* a la cual nos refireremos a lo largo de la tesis como VBR.

Para poder combinar estas "fotografías" en una imagen coherente necesitamos un "reloj interno" que diga, para cada célula, dónde se encuentra en el proceso que caracteriza al tejido. Una forma de inferir este reloj es reconstruyendo una trayectoria sobre la VBR que involucren transiciones entre células cercanas, desde un estadío celular inicial.

Otro desafío que trae la tecnología scRNASeq viene por una limitación técnica inherente que es la ineficiente captura del RNAm que lleva a reportar valores nulos, conocidos como *dropouts*. Esto resulta en que se suelan reportar experimentos en los que un gran número (que a veces supera el 90 %) de datos presentan valores nulos, sólo por razones técnicas. Obviamente, esto dificulta aún más nuestra capacidad de inferir la superficie de menor dimenesionalidad por la carencia de información.

En esta tesis modelamos a la VBR como un grafo cuyos nodos representan muestras de célula única cuyas vecindades se construyen a partir de las distancias euclideanas entre los respectivos transcriptomas medidos en las muestras. Al ser la VBR una variedad de células en un espacio de menor dimensión, consideraremos una cantidad menor de genes para su modelado. Para ello, existen varios criterios de selección. En particular, estudiaremos uno llamado Slicer (Welch et al., 2016) que parte de la siguiente premisa. Los genes involucrados en el modelado de la VBR deberán presentar relativamente alta variabilidad a lo largo de los experimentos, de manera

de ser informativos, pero al mismo tiempo dicha variabilidad tiene que ser compatible con la topología de la VBR que modelamos como el grafo.

La tesis se encuentra organizada de la siguente forma. En los Cap.2 y Cap.3 se utilizan datos sintéticos para modelar la dinámica de genes y sobre estos se estudia la técnica de selección de genes de Slicer. Diremos que la técnica es efectiva si, una vez aplicada, sobre el grafo de células (la aproximación a la VBR) podemos reconocer trayectorias entre células cercanas. Primero se trabaja sobre un modelo simple (Welch et al., 2016) en el Cap.2 y luego, se presenta un modelo más robusto en un sentido biológico construido por nosotros a partir un conocimiento de a qué procesos biológicos están asociados los genes en el Cap.3.

En el Cap.4 estudiamos la selección de genes por el mismo criterio sobre datos que se obtuvieron de un experimento realizado por Linnarson Lab (Hochgerner et al., 2018) en el contexto de un estudio del desarrollo de células del giro dentado del hipocampo en ratones para entender cómo es la neurogénesis en cada estadío. Por la fracción alta de expresiones nulas (90 %), i.e. *dropouts*, el criterio de selección de genes de Slicer no resulta ser satisfactorio. En consecuencia, proponemos tres nuevos criterios de selección que son estudiados en detalle.

Por otro lado, en el Cap.5 nos proponemos atacar el problema de la tasa alta de *dropouts* que encontramos en los datos experimentales. Para ello estudiamos un método de imputación llamado Scimpute (Li and Jessica Li, 2018). Cuantificamos su performance aplicando el algoritmo sobre el modelo más robusto explicado en el Cap.3 sobre el cual simulamos los eventos *dropouts*. Finalmente se presentan las principales conclusiones del trabajo y las perspectivas futuras.

Capítulo 2

Reconstrucción de variedades de baja dimensión. Modelo sintético

El enfoque que abordaremos en la tesis implica representar a cada célula, cuyo transcriptoma fue medido, proyectada en un espacio de perfiles de expresión génica de alta dimensionalidad (dim $\sim o(10000)$). En este capítulo utilizaremos un modelo sintético de los datos recopilados en un experimento scRNASeq para comenzar a analizar metodologías que permitan reducir la dimensionalidad en este tipo de descripciones.

2.1. Espacio de estados celulares y variedad biológicamente relevante

Una de las técnicas más recientes para interrogar el estado celular es conocida como secuenciación de célula única (scRNAseq). La misma permite examinar la información asociada al conjunto de moléculas de ARN (llamado transcriptoma) con resolución de célula individual (Tang et al., 2009). La gran ventaja de utilizar esta tecnología es el acceso a información sobre heterogeneidad celular, que hasta este momento se perdía al utilizar otros métodos que muestrean el transcriptoma a nivel tejido, por lo que promediaban sobre un conjunto de células el nivel de expresión relevado (metodologías tipo *bulk*, en inglés). El acceso a información a nivel de célula única es muy relevante para estudiar la diversidad celular, entender procesos de desarrollo y diferenciación, a la vez que para estudiar las redes de regulación génica que coordinan dichos procesos (Hochgerner et al., 2018).

Al generar información de células individuales, en vez de un valor medio sobre un ensamble celular (como en experimentos *bulk*) esta tecnología produce un volumen muy grande de datos complejos de alta dimensionalidad. Como señalamos en Cap.1, típicamente se recaba información sobre el estado transcripcional para del orden de 30000 genes en alrededor de miles o decenas

de miles de células diferentes. Sin embargo, es razonable asumir como hipótesis de trabajo que no todos los 30000 genes se expresan independientemente, sino que su expresión se coordina (se *ordena*) para llevar a cabo funciones biológicas de mayor escala de organización dentro de la célula. Esto conduce a la hipótesis de que el estado celular puede ser descripto en realidad sobre una variedad de células de mucha menor dimensión, que involucre explícitamente a una menor cantidad de genes relevantes, la VBR.

En experimentos scRNASeq se mide la expresión del transcriptoma de células. A partir de estos datos se contruye lo que se conoce como *matriz de expresión* E donde las columnas representan una muestra scRNASeq (una célula) y las filas los genes de forma que cada elemento de la matriz E_{ij} representa el nivel de expresión del gen i-ésimo en la célula j-ésima medida.

Cabe destacar que con esta tecnología se mide el transcriptoma de la célula de forma invasiva por lo que no es posible seguir la evolución de cada célula a lo largo del tiempo del desarrollo biológico, i.e. en varios experimentos consecutivos. Por le contrario, en un mismo experimento, son relevadas poblaciones de células que podrían incluir eventualmente muestreos de tipos o estados celulares de diferentes estadíos. De esta forma, es posible eventualmente detectar procesos y evoluciones temporales de manera indirecta, reconstruyendo trayectorias que ocurren a lo largo de un parámetro denominado *pseudo-tiempo* que intenta modelar el tiempo de evolución biológico a partir de la distribución del ensamble de células profiladas en el experimento.

En la práctica, los perfiles del transcritoma relevados en muestras de célula única se pueden considerar como puntos en un *espacio de estados celulares* donde se define la distancia entre dos puntos como la distancia euclideana entre los transcriptomas (las expresiones de los genes). En este espacio dos células son cercanas si presentan perfiles de expresión similares (como es de esperar por ejemplo, en un mismo tipo celular en dos estados de desarrollo cercanos). En este contexto, una manera de reconstruir *trayectorias de desarrollo* implica reconocer caminos sobre el *espacio de estados celulares* que involucren transiciones entre células cercanas, desde un estadío celular inicial.

Si lo que tiene sentido físico es, en vez del *espacio de estados celulares* completo, la VBR sobre la que se dispone el ensamble de estados celulares, las métricas y trayectorias biológicamente relevantes deben ser geodésicas sobre la misma que, localmente, pueden ser aproximadas como la distancia euclideana entre células vecinas. Esto correspondería a aproximaciones locales utilizando el plano tangente a la variedad.

La aproximación a la VBR empleada en esta tesis consiste en representar a esta variedad como un grafo de nodos, asociados a celulas muestreadas, en donde la distancia entre los nodos viene dada por similitud de los transcriptomas medidos. En la Fig.2.1 observamos un esquema de este modelo donde las células son representadas como nodos enlazados con las células cercanas definiendo así vecindades. Vemos en la figura que las métricas locales lineales pueden extenderse a escalas globales a partir de solapamientos parciales de las mismas, de la misma manera que los planos tangentes (en parte superior) se solapan para recubrir y describir la variedad no-lineal de la VBR que se muestra en la parte inferior.

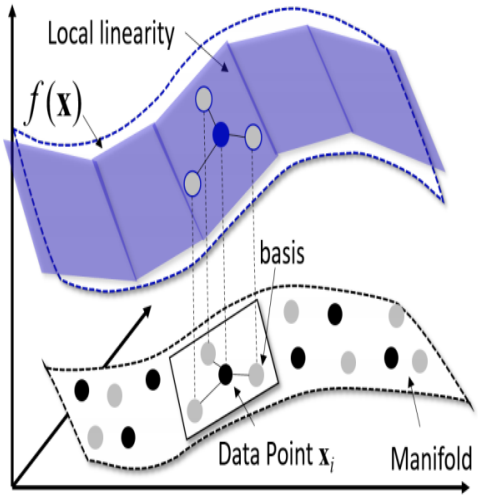


FIGURA 2.1: Esquema de la aproximación para describir el *espacio de estados celulares* como puntos sobre un sábana que representa a la VBR. Se ve que este es una sucesión de planos indicados en azul en la parte superior. Estos contienen a las células como nodos que se enlazan según las vecindades. Figura extraida (Science, 2018).

La VBR nos permite hacer el reconocimiento de vecindades de células similares a partir de la distancia euclideana entre ellas. Así es posible trazar caminos que involucran transiciones entre las vecindades y reconocer trayectorias descriptas por una variable paramétrica que denominaremos *pseudo-tiempo*. En Fig.2.2 se esquematiza la VBR como un espacio de estados de menor dimensionalidad en donde el color de cada célula indica su tipo. Nótese que no se enlazan los puntos como nodos en el grafo para permitir la correcta visualización.

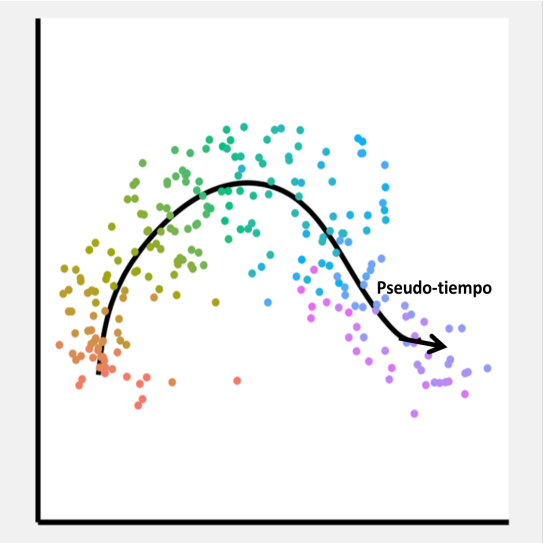
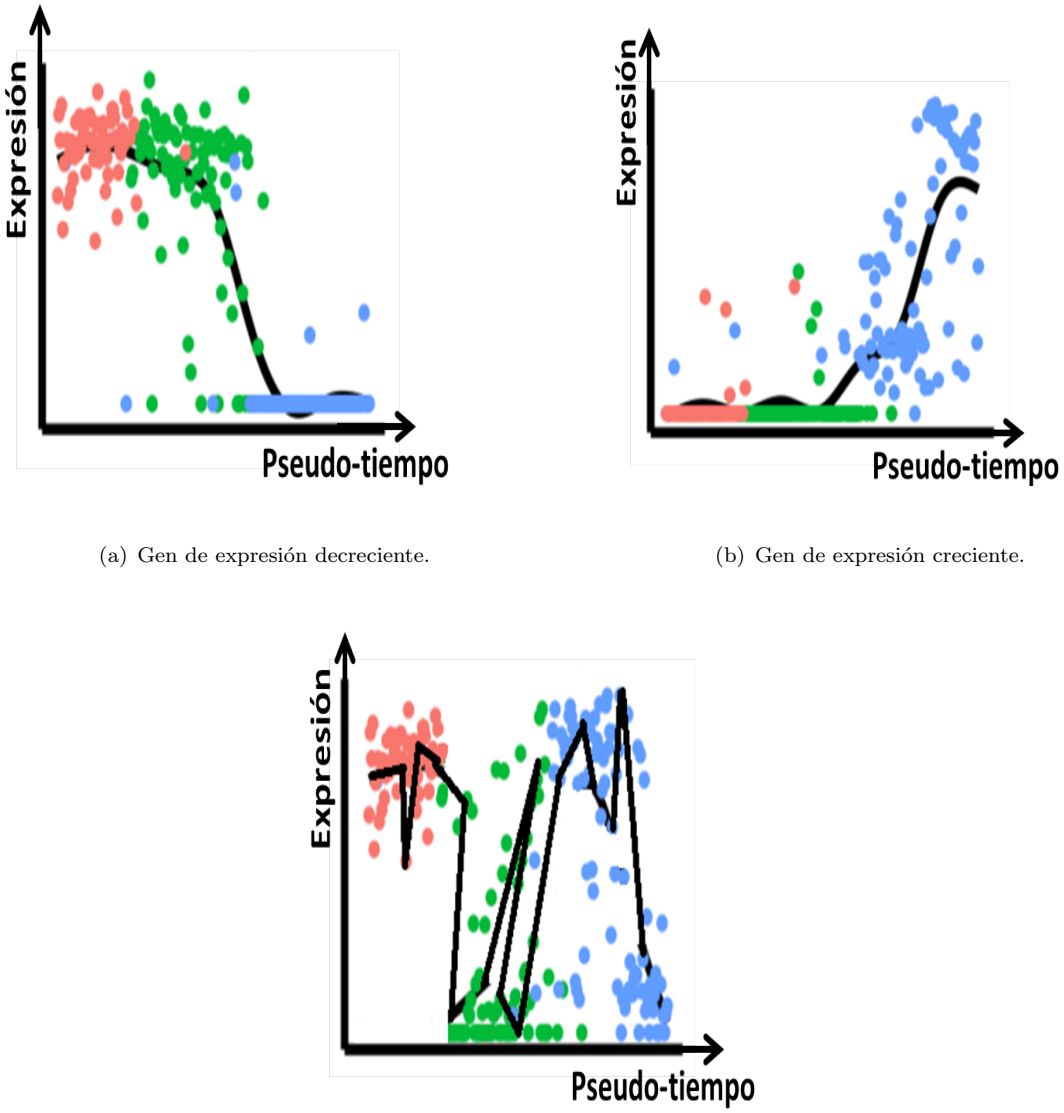


FIGURA 2.2: Esquema del modelo de la VBR en el cual las muestras scRNASeq se representan como puntos cuyo color indica el tipo celular. Las distancias euclideanas entre los puntos permite reconocer vecinades que se condicen con los tipos celulares y así se infiere una trayectoria parametrizada por el *pseudo-tiempo*. Figura extraida (Cannoodt et al., 2016) y fue editada.

Observamos que la cercanía (en distancia euclideana) entre las células permite el armado de vecindades que se condicen con los tipos celulares presentes en el desarrollo biológico. Es así como se infiere una trayectoria de estados celulares (indicada en línea negra) parametrizada por el parámetro *pseudo-tiempo*.

El reconocimiento de las vecindades de las células y la reconstrucción de las trayectorias sobre la VBR no lo llevaremos a cabo considerando la totalidad de genes relevados, sino de un conjunto particularmente informativo de genes. Los mismos, deberán presentar relativamente alta variabilidad a lo largo de los experimentos, de manera de ser informativos, pero al mismo tiempo dicha variabilidad tiene que ser compatible con la topología de la trayectoria inferida sobre la VBR. Veamos en Fig.2.3, de forma esquemática, las expresiones de algunos genes sobre la trayectoria de células inferida en Fig.2.2.



(a) Gen de expresión decreciente.

(b) Gen de expresión creciente.

(c) Gen de expresión aleatoria.

FIGURA 2.3: Esquemas de la evolución de la expresión de tres tipos de genes sobre la trayectoria inferida sobre la VBR parametrizada por el *pseudo-tiempo* (esquematizaado en Fig.2.2). Figura extraida (Cannoodt et al., 2016) y editada.

En los paneles A y B se observan genes que presentan evoluciones de sus niveles de expresión suaves respecto al parámetro *pseudo-tiempo* compatibles con la topología inferida. Reconocemos en el panel A expresiones altas del gen en ciertos tipos de células, las representadas por color salmón y verde, mientras que para las del tipo de color celeste, el gen no se expresa significativamente. Se trata de un gen con expresión creciente en la trayectoria inferida. Por el contrario, en el panel B se esquematiza el comportamiento inverso correspondiente a un gen de expresión decreciente.

Afirmamos que estos genes son los informativos de la biología del desarrollo y que presentan variabilidades locales relativamente bajas en vecindades de células a lo largo de las trayectorias reconstruidas. Por el contrario, observamos en el panel C una evolución que no es compatible con la topología inferida pues la alta variabilidad en la expresión en las vecindades (representadas por los colores) no permite asociar un comportamiento de evolución con los tipos celulares de la trayectoria reconstruida.

En consecuencia, una forma de reconocer genes de relevancia biológica en el desarrollo es cuantificando la variación de la expresión de estos localmente, es decir en las vecindades de las células. Aquellos genes informativos presentarán una variabilidad local menor que la global y serán estos los que participarán en la reconstrucción de la VBR.

Para cuantificar la variación sobre las vecindades definimos una varianza local que, análogamente a como se define la varianza global (σ^2), es un promedio de las variabilidad en cada vecindad que llamaremos varianza entre vecinos (S^2). De esta forma, seleccionaremos para la construcción de la VBR a aquellos genes que cumplen $S^2 < \sigma^2$.

Este criterio presentado de selección de genes es conocido como la técnica de reducción de dimensionalidad de Slicer (Welch et al., 2016). En este capítulo estudiaremos en particular esta técnica sobre un modelo que simula los resultados que se obtienen de un experimento scRNASeq.

2.2. Modelo sintético de expresión celular

Dentro de las células los genes se orquestan para llevar adelante diferentes proceso biológicos, como por ejemplo la neurogénesis. Siguiendo el enfoque introducido en Welch et al. (2016) consideraremos la existencia de cinco procesos biológicos que evolucionan temporalmente de acuerdo a las ec.2.1 donde t es un vector que representa al tiempo de la simulación y c_i es una valor de una función normalmente distribuida en el 0 con una desviación estándar (**std**) que se adiciona para cada t.

$$\begin{aligned} f_1 &= (5cos(t/5) + 8) + c_1 \\ f_2 &= (5sin(t/5) + 8) + c_2 \\ f_3 &= \sqrt{t} + c_3 \\ f_4 &= \tfrac{1}{2}(\tfrac{t}{20})^2 + c_4 \\ f_5 &= \tfrac{1}{4}(16 - \tfrac{t}{20})^2 + c_5 \end{aligned} \tag{2.1}$$

De esta forma vamos a simular los datos que se obtienen de un experimento scRNASeq modelando la evolución de los genes que están agrupados en grupos asociados a ciertos procesos biológicos.

Si se piensa que en un dado tiempo el tejido estudiado se encuentra realizando un cierto proceso biológico, los genes asociados a este, llamemoslos N_i, tendrán niveles altos de expresión. Al

evolucionar el tiempo, ahora algunos procesos comienzan a atenuarse mientras que otros nuevos comienzan a ocurrir caracterizados por los genes N_j. De esta forma, los genes N_i comienzan a tomar niveles de expresión más bajos mientras que los N_j, más altos.

De esta manera, construimos una matriz de 80 muestras scRNASeq y 500 genes (100 para cada grupo regulatorio). El perfil de expresión de cada gen en la célula j-ésima se corresponde con el valor que toma la función del grupo al que pertenece a tiempo t_j, e incluye un ruido gaussiano aditivo centrado en 0 caracterizado por una desviación estándar **std**. El tiempo se representa con el vector t que es una sucesión equiespaciada de L valores entre 0 y 80 y como modelamos 80 células, $L = 80$.

Adicionalmente, es de esperar que no todos los genes presentan una evolución temporal coordianda. Para simular esto, a un porcentaje de los genes, representados por **porcen**, les permutamos el orden de los perfiles de expresión así no muestran una dependencia explícita con el tiempo.

En resumen, simulamos una *matriz de expresión* de un experimento scRNASeq parametrizada por dos parámetros **E(std, porcen)**:

- **Std**. La desviación estándar de la función normalmente distribuida que se le suma a la expresión de los genes como un ruido gaussiano.
- **Porcen**. El porcentaje de genes total (misma en cada grupo regulatorio) de genes a los que se les ha eliminado una dependencia explícita con el tiempo.

A partir de la *matriz de expresión* **E(std, porcen)** vamos a construir un grafo de k-primeros vecinos, donde cada nodo representa un perfil transcripcional extraido de una muestra de célula única. Cada nodo se enlazará con sus k vecinos más proximos. Es bueno notar que la red así construida es una buena aproximación a la VBR, ya que induce una representación global que respeta relaciones de vecindades locales observadas en el espacio original (ver Fig.2.1 y Secc.2.1).

En la práctica consideramos a cada perfil como un vector (i.e. columnas de E). El perfil j-ésimo lo designamos $E_{i,j}, i = 1, 2, ..., N$ (N total de genes) donde cada uno de sus elementos representa la expresión de cada uno de los genes. La distancia euclideana $d(E_{i,j}, E_{i,k})$ entre dos perfiles j-ésimo y k-ésimo se encuentra definida como:

$$d(E_{i,j}, E_{i,k}) = \sqrt{\sum_i (E_{i,j} - E_{i,k})^2} \tag{2.2}$$

Para armar la matriz de adyacencia de la red calculamos para cada nodo (perfil transcriptómico) la distancia euclideana con el resto de los nodos, y enlazamos a cada nodo con los k-nodos vecinos que presentan la menor distancia para con él. En el proceso de armado de la red, elegimos el menor valor de k tal que la estructura del grafo quede completamente conexa. Comenzamos

caracterizando a la red que se corresponde con la matriz E con una baja fracción de genes sin dinámica (**porcen** = 0.1) y un bajo ruido en la expresión de los genes (**std** = 0.1) y $k = 3$ (tal que queda el grafo conectado).

La Fig.2.4 muestra un subgrafo de sólo 10 nodos de la red completa donde cada nodo representa a una célula que ha sido etiquetada de acuerdo al tiempo de la simulación. Nótese que es posible reconocer una trayectoria dinámica en la reconstrucción provista por la red. Esta va desde el nodo etiquetado como 1 hasta el nodo 10 que se asemeja al orden dado por el tiempo de la simulación. Sin embargo, este orden no es nodo a nodo tal como se simuló.

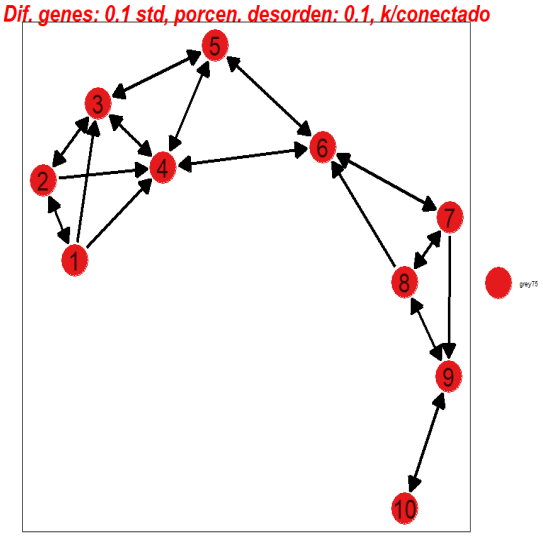


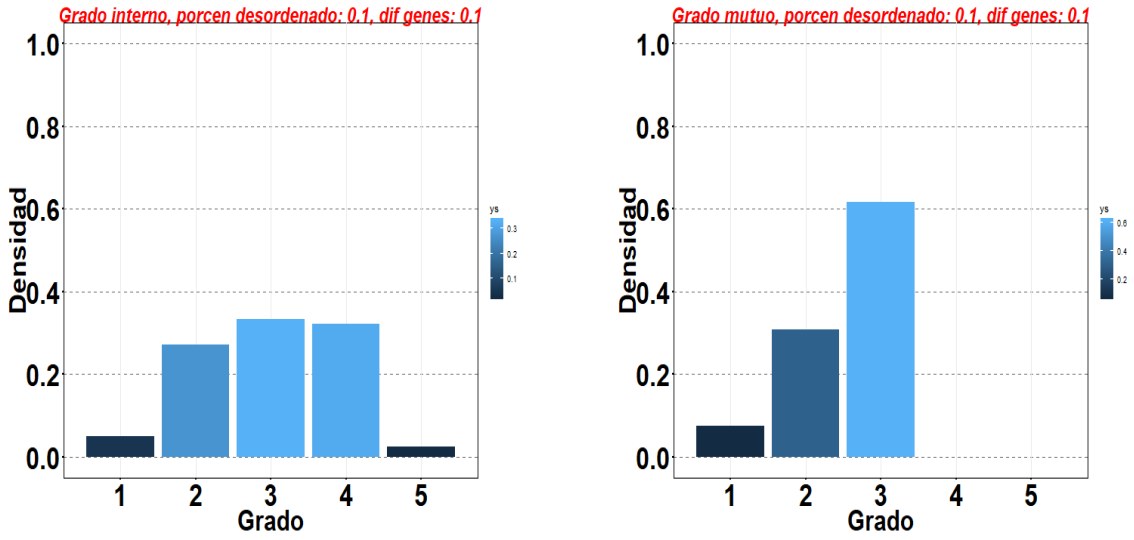
FIGURA 2.4: Subred de 10 nodos de la red total para baja fracción genes sin dinámica (**porcen** = 0.1) y una bajo ruido en expresión (**std** = 0.1). Cada nodo representa un perfil del transcriptoma de una célula muestreada.

Se muestran algunas caracteristicas del grafo completo en Tabla. 2.1. Al tratarse de una red dirigida, se definen tres tipos de grado: grado interno, grado externo y grado mutuo. El externo se encuentra fijo en 3 pues $k = 3$. Sin embargo, se da una distribución para el grado interno y el mutuo.

Nodos	**Enlaces**	**Grado interno medio**	**Grado externo medio**
80	243	3	3

TABLA 2.1: Características de la red de muestras de baja fracción de genes sin dinámica (**porcen** = 0.1) y de bajo ruido en expresión (**std** = 0.1). La cantidad de vecinos es $k = 3$, se encontró que es el valor tal que el grafo queda conectado.

Graficamos la distribución del grado interno y del mutuo para la red (Fig. 2.5). Vemos que los máximos de densidad para ambas distribuciones se da para 3 vecinos. La diferencia que se observa es que mientras que el grado mutuo, como vemos en el panel A, llega a tomar hasta grado 3, vemos en el panel B que la red presenta densidad no nula hasta un grado interno de 5.



(a) Grado interno.

(b) Grado mutuo.

FIGURA 2.5: Distribuciones de grado para la red con una baja fracción de genes sin orden temporal (0.1) y un bajo ruido en expresión $std = 0.1$.

Concluimos que la red construida es una buena aproximación a la VBR sobre la cual se puede reconocer la existencia de una trayectoria de la evolución dinámica. Dado que los índices de los nodos se corresponden con la evolución en el tiempo modelada, esperamos que el nodo j-ésimo tenga como vecino al nodo (j-1)-ésimo y al (j+1)-ésimo de forma de ordenarse de forma creciente. Aunque se reconoce la existencia de una trayectoria en Fig.2.4, no se trata del orden estricto que esperamos.

Como ya explicamos, la reconstrucción de trayectorias sobre la VBR depende de los genes seleccionados. Esto nos lleva a adoptar un criterio para seleccionar genes tal que si construimos el grafo, considerando sólo las expresiones en estos genes, podamos reconocer el orden de evolución estricto que modelamos nodo a nodo a diferencia del orden que se observó en Fig.2.4.

El criterio para la selección de genes que estudiamos (presentado en Secc.2.1) se basa en la idea de que aquellos genes que participarán en la reconstrucción de la VBR serán aquellos que presenten una variabilidad compatible con la topología de la misma. Dado que aproximamos esta variedad como un grafo cuyos nodos son las muestras scRNASeq, su topología vendrá dada por los enlaces entre nodos que definen vecindades locales. De esta forma, consideraremos que un variabilidad es compatible con la topología del grafo, toda vez que la varianza local (S^2), definida sobre las vecindades de los nodos del grafo, sea menor que la varianza global (σ^2), presentada en la totalidad de la población. De esta manera, afirmaremos que el gen g es seleccionado si cumple: $S_g^2 < \sigma_g^2$.

La varianza local del gen g, S_g, se define según ec.2.3. Veamos la primera sumatoria, se compara la expresión del gen g en la célula i con la expresión del mismo en las k_c vecinas de i (las $N(i, j)$,

$j = 1, 2..., k_c$). Para extender el cálculo a toda la población de células, sumamos sobre el índice i hasta las M células y promediamos dividiendo por $Mk_c - 1$ donde k_c es el grado común a todos los nodos tal que el grafo queda como una estructura perfectamente conexa.

$$S_g^2 = \frac{1}{Mk_c - 1} \sum_{i=1}^{M} \sum_{j=1}^{k_c} (e_{gi} - e_{gN(i,j)})^2 \tag{2.3}$$

Nos proponemos ahora cuantificar la performance del criterio de filtrado de genes según los tres parámetros de modelado: el número de vecinos (k) considerados en el armado de la red, el nivel de ruido gaussiano que afecta la expresión de genes que obedecen la misma dinámica (**std**) y el porcenaje de genes que no presentan dinámica parametrizada por el tiempo (**porcen**).

Para cada estudio, fijamos dos parámetros mientras variámos el tercero (Tabla 2.2). Para el caso del ruido en expresión (**std**) y el porcentaje de genes sin orden temporal (**porcen**), éstos se variaron entre tres valores: bajo, medio y alto. Para el caso de los vecinos (**k**) se estudiaron 10 valores.

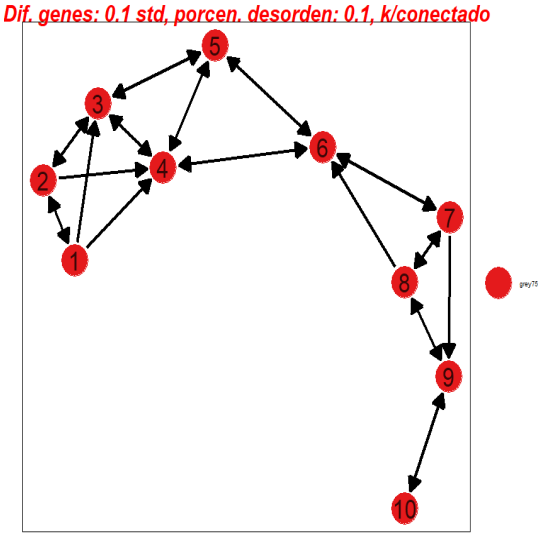
Objeto de estudio	**Parámetros fijos**	**Parámetro variable**	**Bajo**	**Medio**	**Alto**
% Genes sin orden temporal	Std = 0.5 k tq conectado	Valor de porcen	0.1	0.4	0.8
Ruido en dinámica de genes	Porcen = 0.1 k tq conectado	Valor de std	0.5	1	2
Cantidad vecinos	Std = 0.5 Porcen = 0.1	k	1-10		

TABLA 2.2: Valores de los parámetros fijos y variables para cada uno de los estudios realizados.

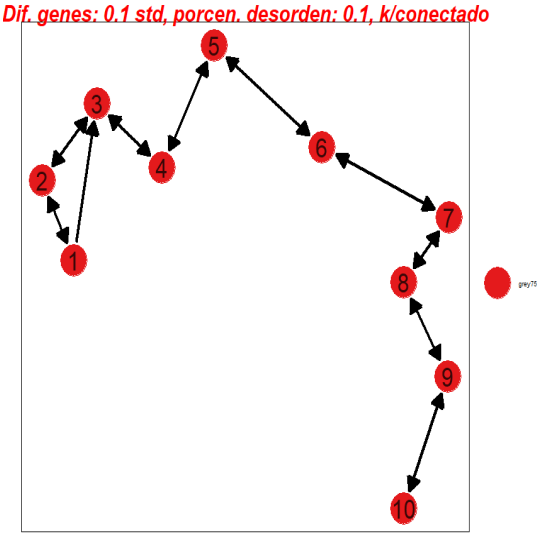
2.3. Selección de genes por noción de suavidad entre células vecinas

Comenzamos analizando los resultados que se obtienen al seleccionar genes para una matriz E caracterizada por un con una fracción baja (0.1) de genes sin dinámica y bajo ruido en la expresión (0.1). Para ello, graficamos un subgrafo en el panel A de la Fig.2.6 y la distribución de grado mutuo del grafo total en el panel C previo a aplicar el filtro. Observamos que los nodos se enlazan, en mayor medida, entre 3 nodos. Una vez aplicado el filtro, vemos en el panel D que la red se caracteriza por su mayoría de nodos con 2 enlaces mutuos y además vemos en el panel B que el enlazamiento es según el orden temporal creciente pues la etiqueta de los nodos es según el orden que modelamos con el tiempo de la simulación. Como luego de la selección

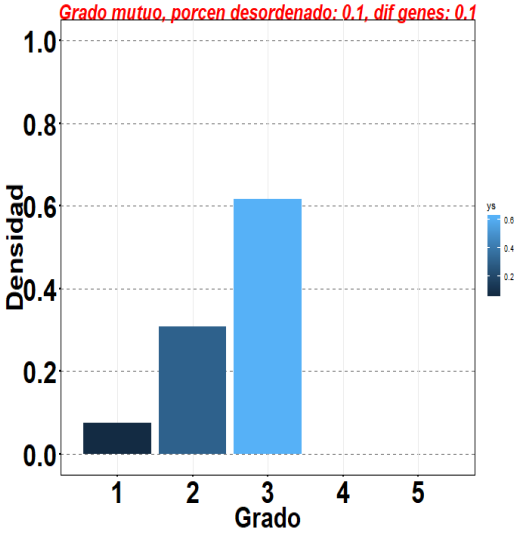
de genes encontramos que los nodos se enlazan según el orden subyacente modelado, afirmamos que el criterio de selección de genes es exitoso.



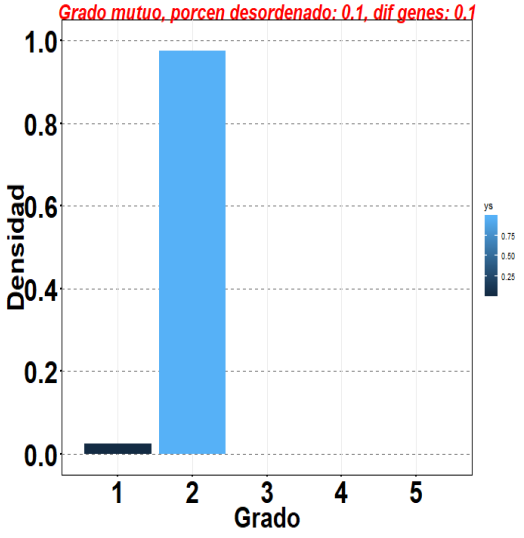
(a) Subgrafo de muestras sin filtrar.



(b) Subgrafo de muestras filtrado.



(c) Distribución de grado mutuo de la red sin filtrar.



(d) Distribución de grado mutuo de la red filtrada.

FIGURA 2.6: Estudio de la distribuciones de grado para la red sin filtrar y filtrada con una fracción baja (0.1) de genes sin dinámica y bajo ruido en la expresión (0.1).

En las siguientes subsecciones estudiamos la eficiencia del filtrado para cada uno de los parámetros mencionados. Se realizaron 100 iteraciones en cada uno de los estudios dado que los genes que se eligen para eliminar su dependencia con el tiempo son elegidos al azar y en consecuencia, en cada ejecución la matriz E resultante es distinta.

Para cuantificar el rendimiento de la técnica definimos los observables. Calculamos sus valores promedios y su respectiva desviación estándar.

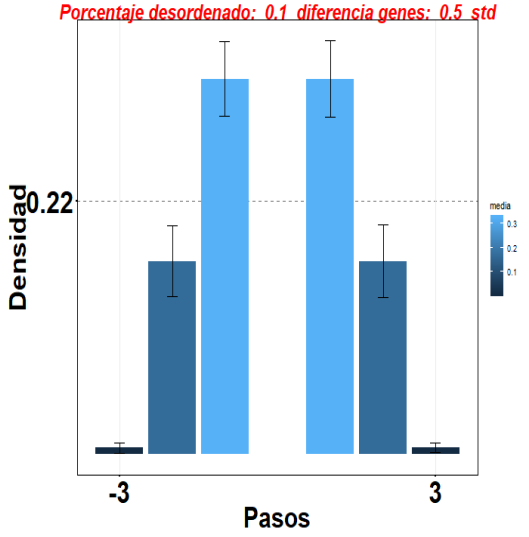
- (a) La distribución de los pasos temporales entre los nodos vecinos. Esto depende de la distribución de grado mutuo y de la compatibilidad de numeración que obtienen los vecinos con el orden que fue modelado.
- (b) El menor camino que une al nodo muestra inicial y al nodo muestra final.
- (c) La fracción sobre los genes que se les eliminó la dependencia con el tiempo que fueron filtrados por la técnica de selección.

Definimos como paso temporal a la cantidad de muestras intermedias entre dos nodos. E.g el nodo 5 se encuentra a tres pasos temporales del nodo 6. Esperamos que si la técnica es eficiente, al filtrar, los nodos vecinos estén a un paso temporal tal que sea una secuencia ordenada de muestras según el tiempo de simulación e.g el nodo 3 tenga como vecinos al nodo 2 y al nodo 4. También esperamos que no hayan caminos cortos entre la muestra inicial y la final, sino que el orden secuencial se respete en mayor medida para toda la población. Por eso estudiamos el menor camino que une a estas dos muestras. Y por último, aquellos genes carentes de dinámica no son informativos y esperamos que sean filtrados.

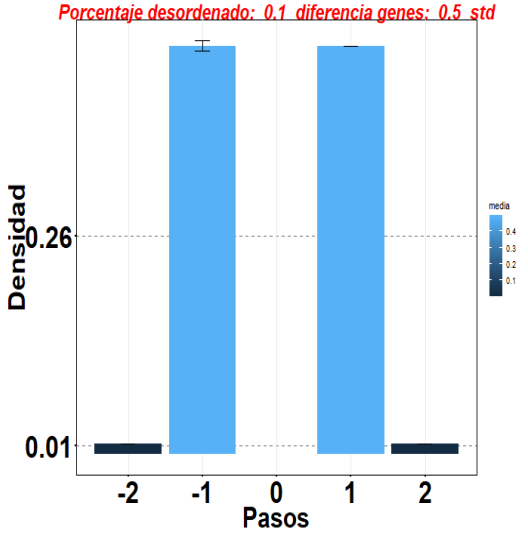
2.3.1. Fracción de genes sin dependencia explícita con el tiempo

Comenzamos estudiando el cambio en la distribución de los pasos temporales entre cada nodo y sus respectivos vecinos para una fracción de genes que se les elimina la dependencia temporal baja (0.1), media (0.4) y alta (0.8). En la Fig. 2.7, para cada uno de estos valores, se muestra a izquierda sin aplicar el filtro y a derecha, al aplicarlo. Observamos que previo al filtrado los nodos tienen mayor densidad de probabilidad asociada hasta 3 pasos, hacia el pesado (valor negativo) y hacia el futuro (valor positivo). Vemos también que a medida que el porcentaje de genes que no siguen un orden temporal aumenta, se obtienen probabilidades no nulas para valor más altos de pasos de forma simétrica hacia el pasado y futuro.

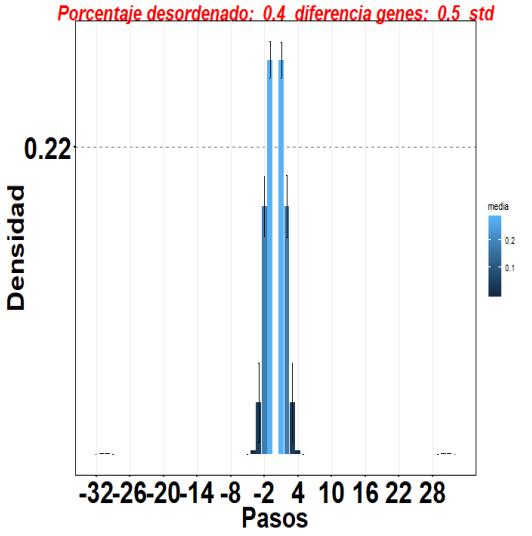
Vemos que al aplicar el filtro, la densidad en los pasos menores (valor absoluto) se incrementan significativamente. Para el caso de fracciones baja (panel B) y media (panel D), el máximo se da para un paso hacia el pasado y hacia el futuro y las densidades para ± 2 es despreciable. Por el contrario, para el caso de una fracción alta vemos en el panel F que se da densidad no nula para valores de pasos más altos. Esto se explica porque previo al filtrado (panel E) por la fracción de genes que no obedecen la evolución alta, no se define un orden sobre la variedad subyacente (la VBR) que pueda ser reconstruido.



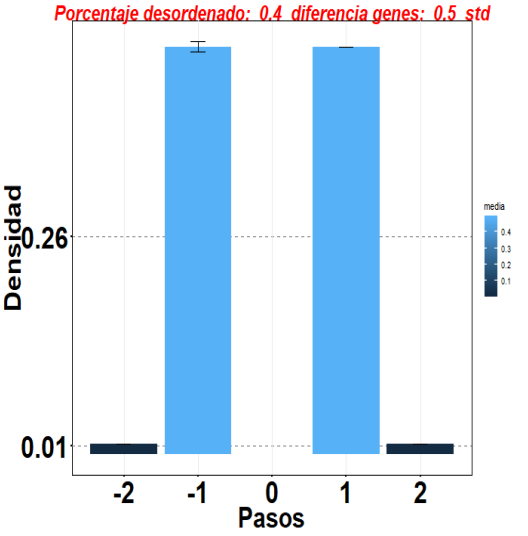
(a) Sin filtrar. **Fracción de genes sin dinámica bajo**.



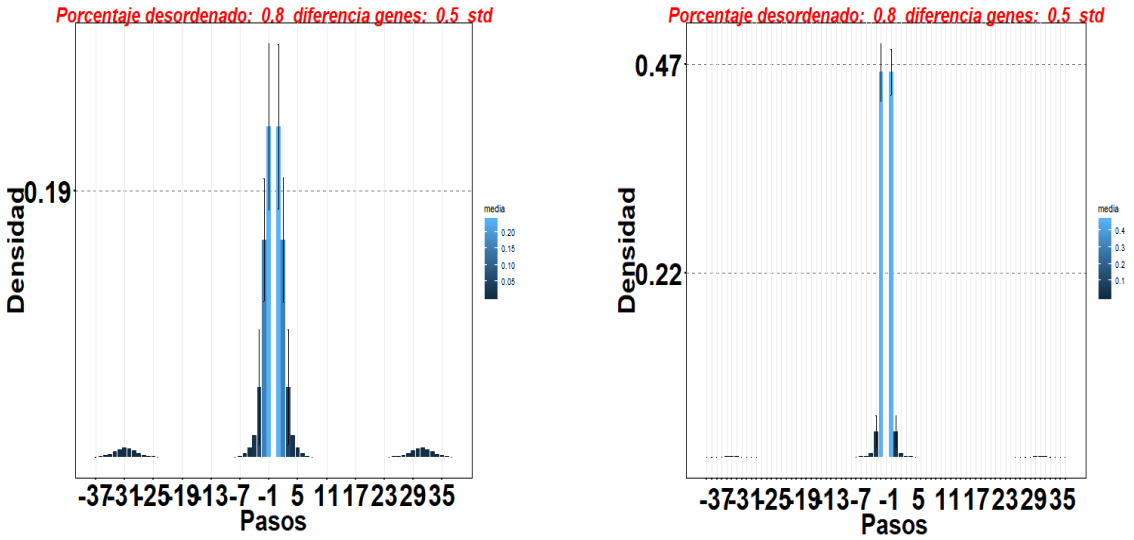
(b) Filtrado. **Fracción de genes sin dinámica bajo** .



(c) Sin filtrar. **Fracción de genes sin dinámica medio** .



(d) Filtrado. **Fracción de genes sin dinámica medio**.



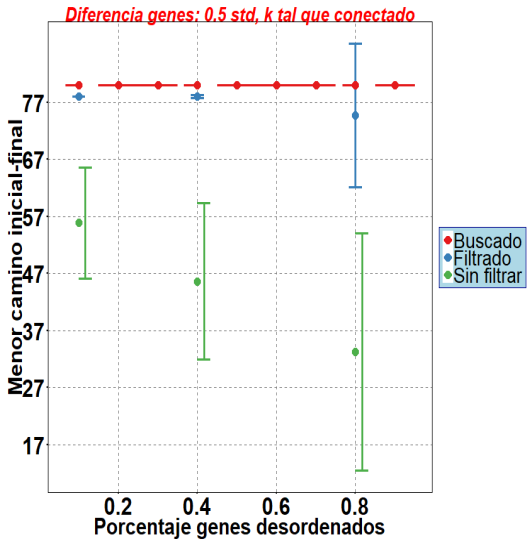
(e) Sin filtrar. **Fracción de genes sin dinámica alto** .

(f) Filtrado. **Fracción de genes sin dinámica alto** .

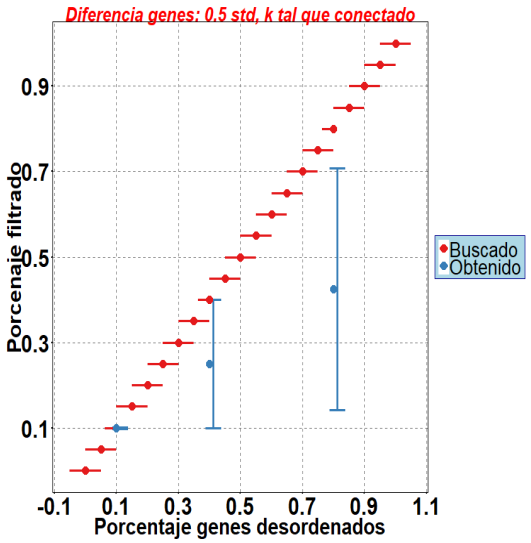
FIGURA 2.7: Distribuciones de cantidad de pasos temporales para la red de muestras sin filtrar (a izquiera) y filtrada (a derecha). Es para valores de fracción de genes a los que se les elimina la dependencia explícita con el tiempo bajo (0.1), medio (0.4) y alto (0.8). Pasos negativos se corresponden con muestras previas en el tiempo de la muestra y los positivos, posteriores.

Por otro lado, graficamos el mínimo camino que une al nodo muestra inicial y al final, en función de la fracción de genes que les eliminamos la dependencia temporal en el panel A de Fig. 2.8 para la red sin filtrar (verde) y filtrada (azul). Observamos que a medida que aumenta la fracción de genes sin dinámica, el menor camino para la red sin filtrar disminuye. Por el contrario, vemos que para las redes filtradas este camino permanece constante (por su incerteza asociada) y muy cercano a 80 ya que la red está compuesta por 80 nodos. Se afirma que se reconstruye el orden subyacente.

También calculamos la fracción de genes eliminados por la técnica de los que les removimos intencionalmente la dependencia con el tiempo. Graficamos el porcentaje de genes filtrados en función de los genes que no siguen la tendencia temporal en el panel B de Fig.2.8. Vemos que a medida que aumenta la fracción de genes sin orden temporal, la fracción sobre los mismos que es filtrada disminuye y además tienen asociada una desviación estándar creciente. Para la fracción alta, se remueven menos genes del conjunto de los que no obedecen el orden. Afirmamos que se está filtrando de menos.



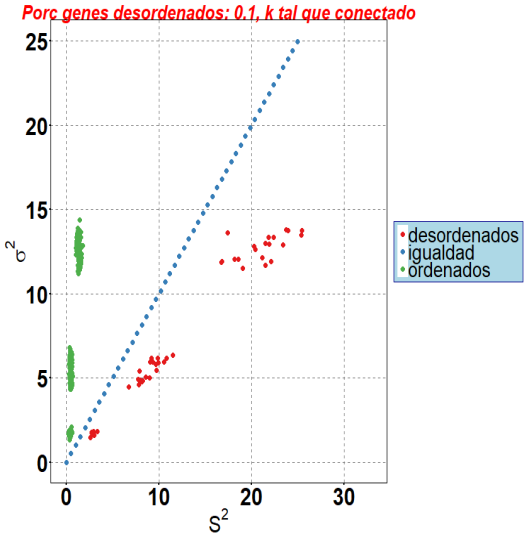
(a) Menor camino entre el nodo muestra inicial y el nodo muestra final para el grafo filtrado (azul) y sin filtrar (verde). El valor esperado (rojo) es el de la cantidad de nodos de la red: 80.



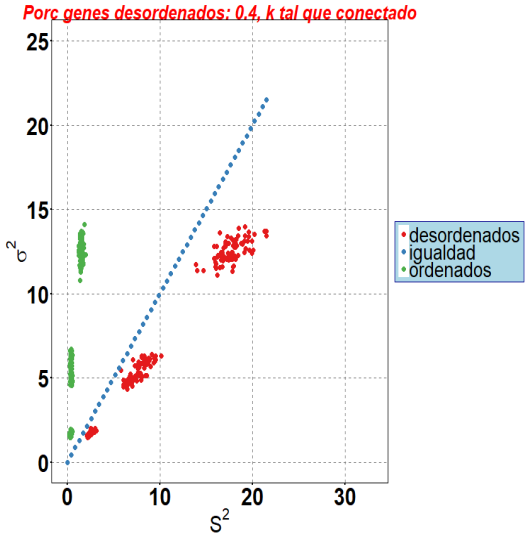
(b) Fracción de genes eliminados por el filtrado en función de la fracción de genes sin dinámica. El valor esperado es la igualdad de la fracción de genes sin dinámica y de los que son filtrados (rojo).

FIGURA 2.8: Estudio de la efectividad del método de filtrado para valor fijo de ruido gaussiano y valores de fracción de genes a los que se les elimina la dependencia explícita con el tiempo bajo (0.1), medio (0.4) y alto (0.8).

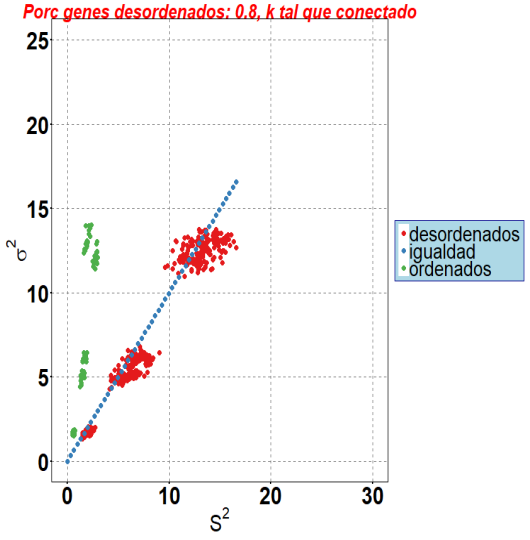
Por último, para visualizar cuáles son los genes seleccionados y cuáles son los filtrados, graficamos la varianza global (σ^2) en función de la varianza entre vecinos (S^2) para cada uno de los genes de la *matriz de expresión* en Fig.2.9 donde cada panel se corresponde con las distintas fracciones de genes sin dependencia explícita con el tiempo. Se espera que en un buen filtrado los genes sin dependencia con el tiempo estén por debajo de ésta recta y los que sí siguen el orden, por encima ya que son estos los seleccionados.



(a) **Fracción de genes sin dinámica bajo** .



(b) **Fracción de genes sin dinámica medio** .



(c) **Fracción de genes sin dinámica alto** .

FIGURA 2.9: Desviación estándar al cuadrado (σ^2) en función de la varianza entre vecinos (S^2) para valores de fracción de genes a los que se les elimina la dependencia explícita con el tiempo bajo (0.1), medio (0.4) y alto (0.8). Verde: genes que tienen una dependencia explícita con el tiempo. Rojo: genes sin orden temporal. Azul: relación uno a uno.

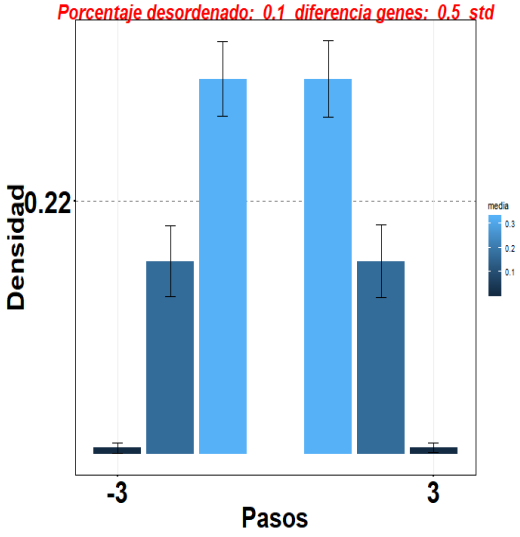
Se observa una agrupación de los valores de algunos genes, que se corresponden con los grupos dinámicos simulados, de forma que las varianzas asociadas a estos genes son similares. Vemos que a medida que aumenta la fracción de genes sin orden temporal, éstos genes sin dinámica (color rojo) toman valores cada vez más altos de σ^2 relativos a S^2 de forma que se acercan a la recta que representa la relación uno a uno (color azul). Para el caso de una fracción baja y media (panales A y B), se seleccionan los genes con dependencia explícita (color verde) y se filtran los carentes de dinámica. Para el caso de un alto porcentaje vemos en el panel C que la

mayoría de éstos genes (color rojo) supera la relación uno a uno de forma que no son eliminados por la técnica sino que son seleccionados indicando que se está sub-filtrando.

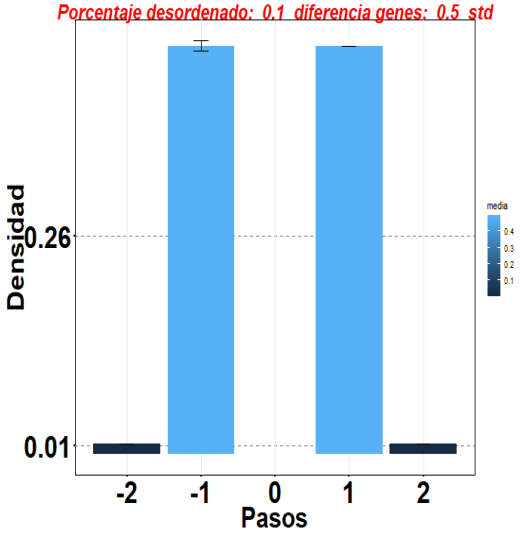
Afirmamos que para una fracción alta de genes carentes de dinámica (fracción de 0.8) la técnica no es efectiva porque no sólo no elimina a la totalidad de los genes sin orden sino también porque no reconstruye un orden de acuerdo al tiempo que simulamos. La razón de esto es que la red sobre la cual se seleccionan los genes no presenta una VBR subyacente sobre la cual las células se disponen de forma ordenada. En consecuencia, como vimos en el panel E de Fig.2.7 no hay un orden que reconstruir. Sin embargo, afirmamos que para valores bajo y medio la técnica es efectiva bajo la perspectiva de los tres observables considerados.

2.3.2. Ruido en la dinámica de los genes

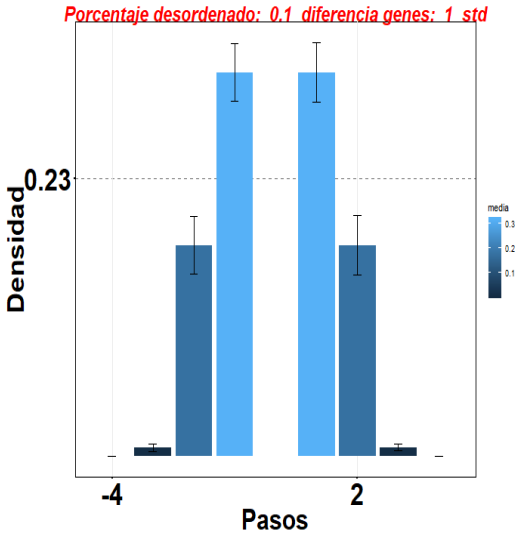
En esta subsección vamos a estudiar qué tan eficiente es la selección de genes según este criterio para distintos valores de ruido gaussiano en la expresión de genes que obedecen una dinámica. Para ello, comenzamos viendo el cambio en la distribución de los pasos temporales (para cada uno de los nodos hacia sus nodos vecinos) para los tres valores propuestos de **std** que caracteriza a la distribución normal que se le suma a la expresión de los genes. Vemos en Fig.2.10 las distribuciones para los valores bajo (0.5), medio (1) y alto (2) mostrando a izquierda sin aplicar el filtro y a derecha, al aplicarlo.



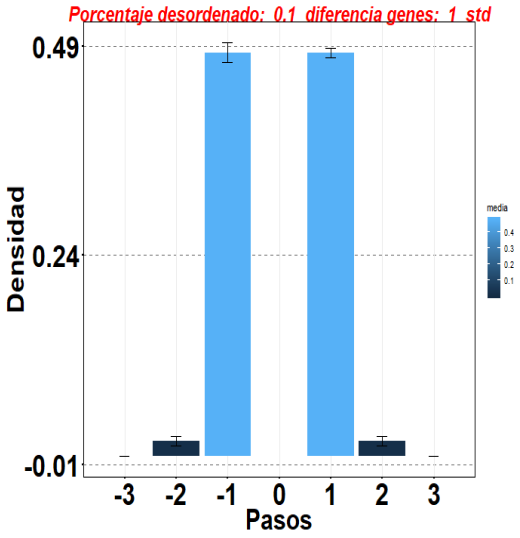
(a) Sin filtrar. **Ruido bajo**.



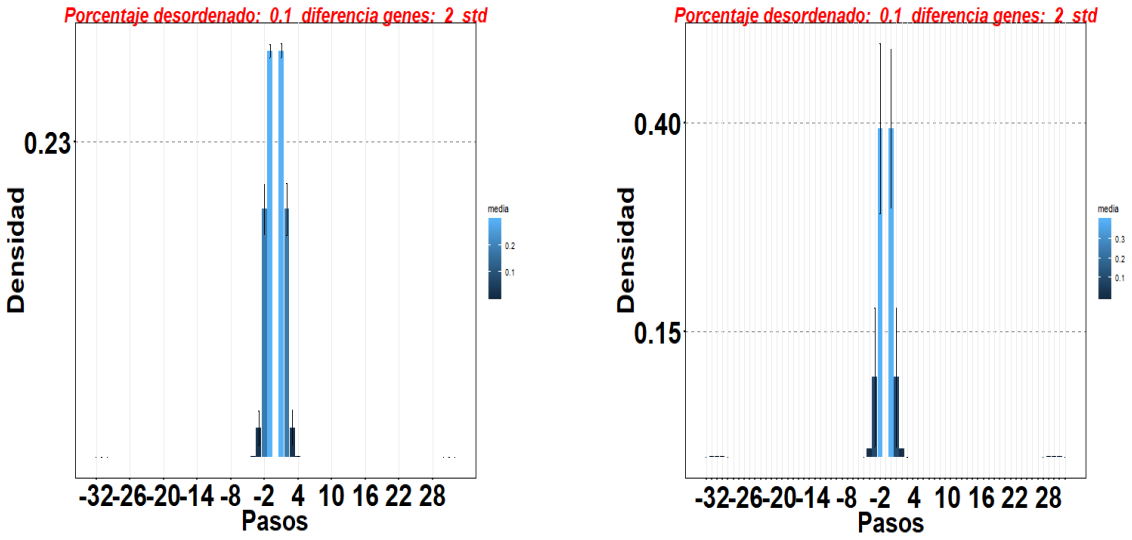
(b) Filtrado. **Ruido bajo.**



(c) Sin filtrar. **Ruido medio.**



(d) Filtrado. **Ruido medio.**



(e) Sin filtrar. **Ruido alto.**

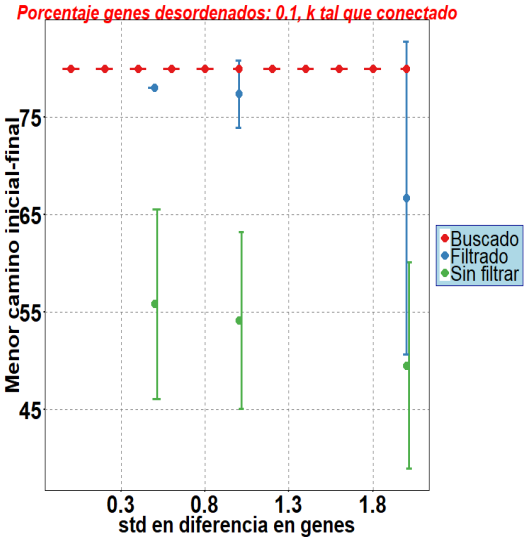
(f) Filtrado. **Ruido alto.**

FIGURA 2.10: Distribuciones de cantidad de pasos temporales para la red de muestras sin filtrar y filtrada para para valores de std bajos (0.5), medios (1) y altos (2) que representan el ruido en la expresión de genes pertenecientes al mismo grupo de dinámica. Pasos negativos son previos en el tiempo de la muestra y los positivos, posteriores.

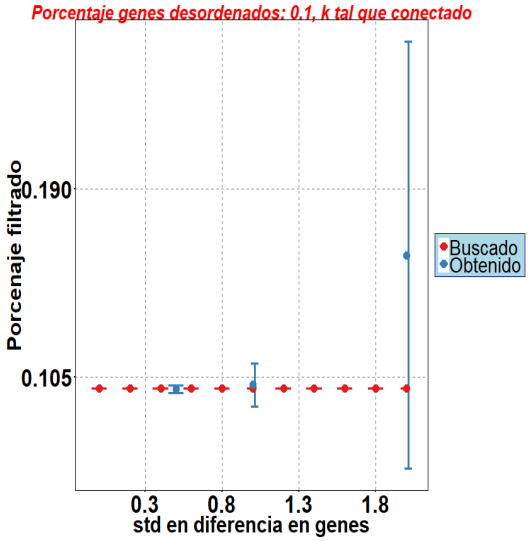
Observamos en los paneles izqueirdos de Fig.2.10 que para la red sin aplicar el filtro a medida que aumenta el ruido, la densidad de probabilidad toma valores no nulos para cada vez pasos en valor absoluto más grandes. Por el contrario, para la red filtrada vemos en los paneles derechos que para el caso de un ruido bajo y medio las densidades se concentran para ± 1 y ± 2 pasos recuperándose un orden mientras que para un ruido alto se dan densidades no nulas para valores muy altos (en valor absoluto) de pasos.

Por otro lado, graficamos en el panel A de Fig.2.11 el mínimo camino que une al nodo muestra inicial y al nodo final en función de la fracción de genes que les eliminamos la dependencia temporal. Vemos que al aumentar el ruido, representado por **std**, el camino mínimo en el grafo sin filtrar permanece constante (considerando la incerteza asociada) siendo este valor mucho mayor que el de la cantidad total de nodos. Al filtrar vemos que, para los tres valores de **std**, se encuentra el camino mínimo esperado para un buena selección también considerando la incerteza asociada.

También calculamos la fracción de genes eliminados que intencionalmente les removimos la dependencia con el tiempo (fijado en 0.1) y graficamos el porcentaje de genes filtrados en función de **std** en el panel B de Fig.2.11. Se observa que para los tres valores estudiados se filtra la fracción de genes correspondiente (considerando el error asociado).



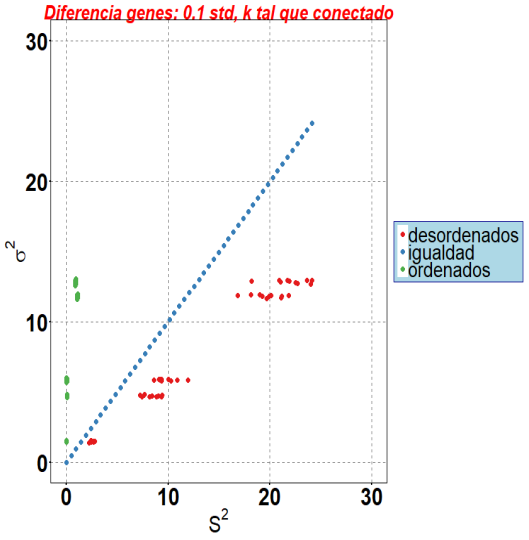
(a) Menor camino entre el nodo muestra inicial y el nodo muestra final para el grafo filtrado (azul) y sin filtrar (verde). El valor esperado (rojo) es el de la cantidad de nodos de la red: 80.



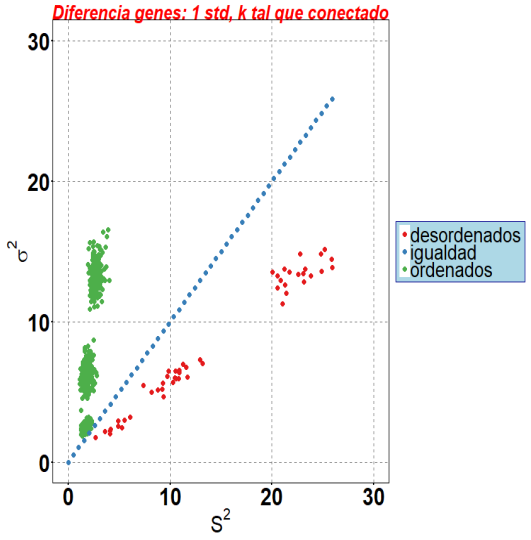
(b) Azul: porcentaje de genes filtrados en función del std que representa al ruido en la expresión dinámica de genes. Rojo: fracción de genes que no siguen el orden temporal.

FIGURA 2.11: Estudio de la efectividad del método de filtrado para valor fijo de fraccion de genes a los que se les elimina la dependencia explícita con el tiempo bajo (0.1) y para valores de std bajos (0.5), medios (1) y altos (2) que representan al ruido en la expresión de los genes.

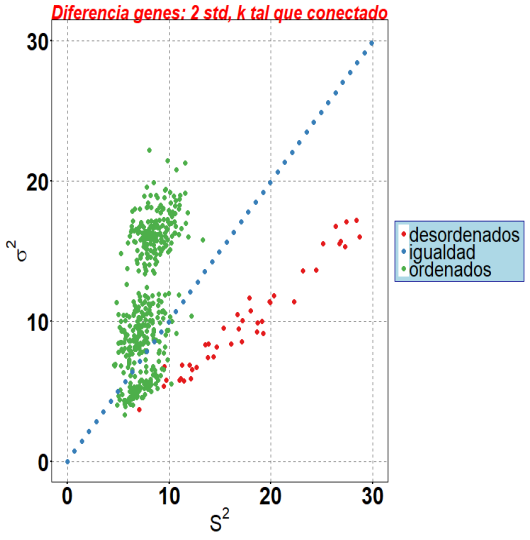
Por último, graficamos la varianza global (σ^2) en función de la varianza entre vecinos (S^2) para cada uno de los genes de la *matriz de expresión* en Fig. 2.12. Se puede ver que para valores bajo y medio (paneles A y B), se obtiene un buen filtrado ya que se seleccionan los genes con dependencia explícita con el tiempo (color verde) y se filtran los que no (color rojo). Por el contrario, para el caso de un ruido alto vemos en el panel C que genes que sí presentan el orden temporal se encuentran por debajo de la relación uno a uno representada por la recta azul. Estos genes son removidos de modo que se está sobre-filtrando.



(a) **Ruido bajo**



(b) **Ruido medio**



(c) **Ruido alto**

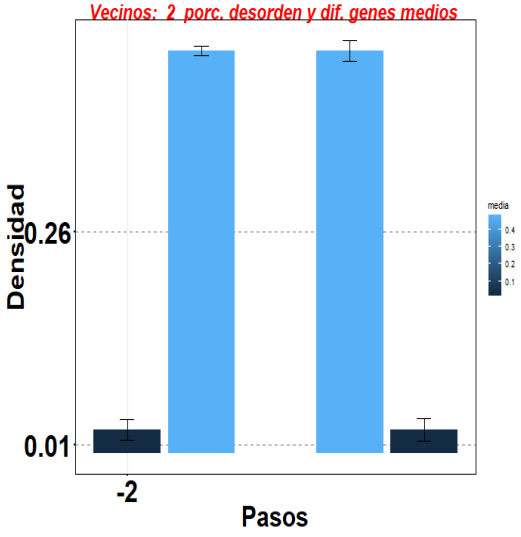
FIGURA 2.12: Varianza global (σ^2) en función de la varianza entre vecinos (S^2) para valores de std bajos (0.5), medios (1) y altos (2) que representan al ruido en la expresión de los genes. Verde: genes que tienen una dependencia explícita con el tiempo. Rojo: genes sin orden temporal. Azul: relación uno a uno.

Así concluimos que la técnica no es efectiva para un valor alto ($std = 2$) de ruido en la expresión de los genes porque aunque sus valores de camino mínimo inicial-final y de fracción de genes filtrado es el esperado (por la incerteza asociada), la distribución de pasos al filtrar toma hasta valores muy altos y se está sobre-filtrando. Esto se debe a que al ser el valor de ruido muy alto, esto rompe con la idea de que se da un orden sobre la VBR (como vimos en el panel E de Fig. 2.7). Esto es muy similar a lo que sucede con un valor alto de fracción de genes sin dependencia temporal como discutimos en la Secc.2.3.1. Por el contrario, para valores bajo y medio afirmamos que la técnica es efectiva bajo la perspectiva de los tres observables considerados.

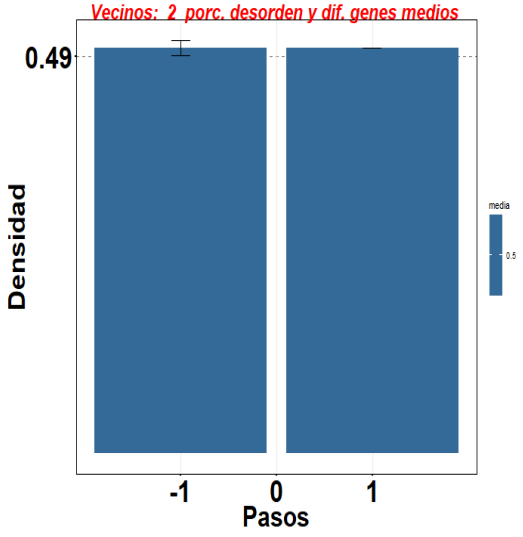
2.3.3. Cantidad de vecinos de los nodos del grafo

Por último, vamos a ver en esta subsección cómo influye la cantidad de vecinos que se utiliza para el armado del grafo que se utiliza como una aproximación a la VBR. También, como se puede ver en ec.2.3, el cálculo de S^2 es dependiente del k_c tal que se propone tomar el valor de k tal que la red queda conectada (Welch et al., 2016). Para entender cómo este parámetro influye en la reconstrucción de un orden según este criterio de selección, vamos a estudiar para distintos valores de k (incluyendo el k_c) la efectividad del filtrado según los observables que ya introducimos.

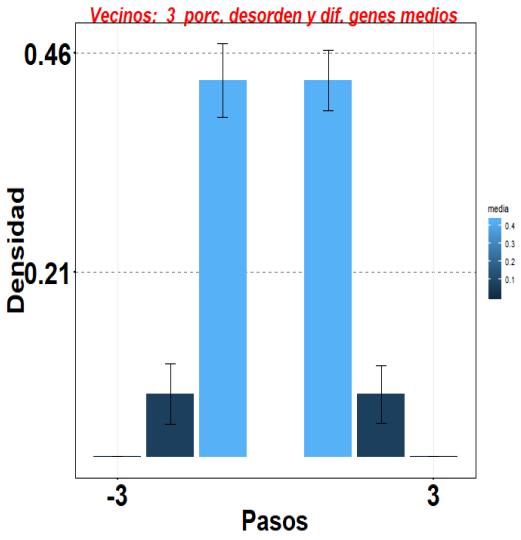
Comenzamos graficando la distribución de los pasos a los que se encuentra los vecinos de los nodos para tres valores de k (incluyendo k_c) en Fig.2.13. Observamos que a medida que aumenta k, la distribución tanto antes de aplicar el filtro como para la red filtrada, toma valores de densidad de probabilidad no nula para valores más grandes (en valor absoluto). Se obtiene un orden secuencial de las muestras perfecto para $k = 2$, muy óptimo para $k = 3$ con densidades bajas para pasos de ± 2 y no bueno para $k = 3$ con pasos de hasta ± 3 asociados a densidades no nulas.



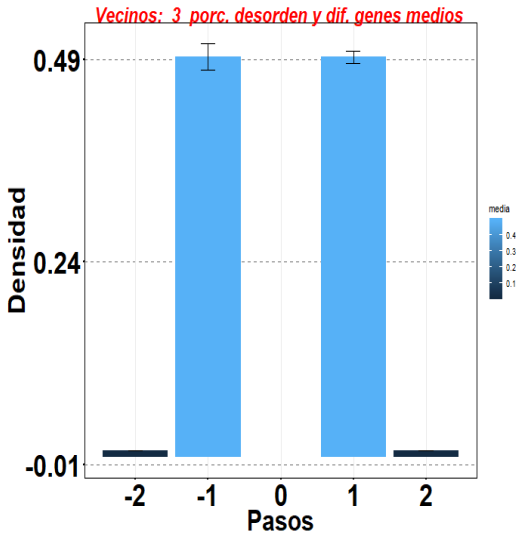
(a) Sin filtar. $k = 2$



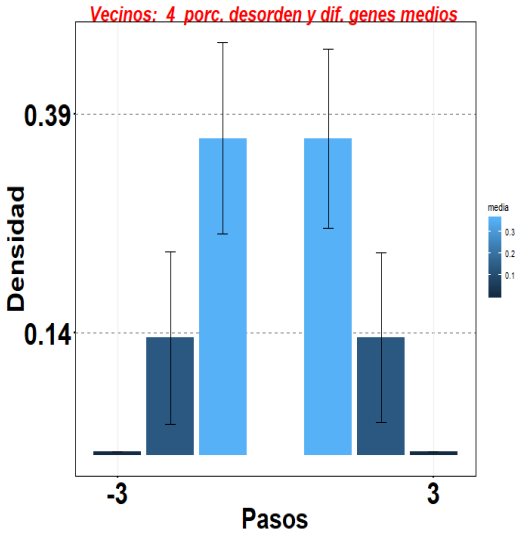
(b) Filtrado. $k = 2$



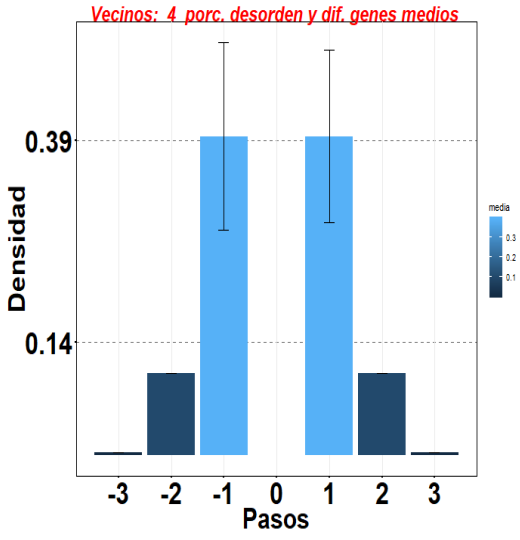
(c) Sin filtrar. $k = k_c = 3$



(d) Fitrado. $k = k_c = 3$



(e) Sin filtrar. $k = 4$



(f) Filtrado. $k = 4$

FIGURA 2.13: Distribuciones de cantidad de pasos temporales para la red de muestras sin filtrar y filtrada para distinta cantidad de vecinos por nodo $k = 2, 3, 4$. Pasos negativos son previos en el tiempo de la muestra y los positivos posteriores.

Aunque pareciera en el sentido del orden temporal que $k = 2$ es el valor más óptimo, si graficamos el porcentaje de genes eliminados por el filtro (fracción de genes sin dinámica está fijo en 0.1) en función del k (Fig. 2.14), notamos que la fracción de genes eliminados se acerca asintóticamente al valor esperado (rojo) a medida que aumenta k. Afirmamos que un $k = k_c = 3$ es una buena relación de compromiso entre éstos dos criterios. Se obtiene una distribución de pasos aceptable que es mejor que la se obtiene para $k = 4$ pero, peor que para $k = 2$. También se consigue una filtración de genes más cercana al valor óptimo que para el caso de $k = 2$ pero, más lejana que para $k = 4$.

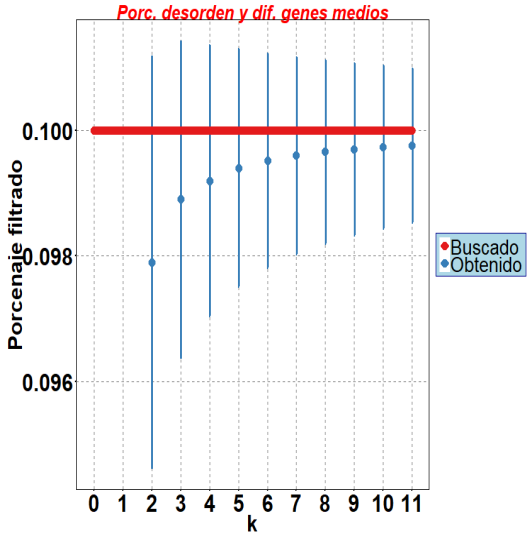


FIGURA 2.14: Porcentaje de genes filtrados en función del valor de vecinos por nodo que es común a todos (k) (azul) y valor esperado (rojo).

A modo de resumen, estudiamos la dependencia de la selección de genes con dos variables que caracterizan a la *matriz de expresión*: la fracción de genes sin orden en el tiempo y el ruido en la expresión de los genes. También estudiamos la dependencia con la cantidad de vecinos (k) de los nodos del grafo de células que viene de la proyección de la *matriz de expresión*. Encontramos valores de los parámetros tal que la selección de genes es satisfactoria e insatisfactoria. Estos se muestran en Tabla.2.3 para cada parámetro estudiado.

	% Genes sin orden temporal	**Ruido en dinámica de genes**	**Cantidad de vecinos**
Satisfactorio	bajo (0.1) y medio (0.4)	std bajo (0.5) y medio (1)	k = k_c
Insatisfactorio	alto (0.8)	std alto (2)	

TABLA 2.3: Valores para los cuales el filtrado se considera satisfactorio e insatisfactorio para los parámetros estudiados. k_c es la cantidad de vecinos tal que la red de muestras queda conectada.

En conclusión, en este capítulo modelamos con un esquema muy simple una *matriz de expresión* que podría obtenerse en un experimento scRNASeq. En nuestro escenario, modelamos la evolución de las células a través de la evolución de grupos de genes que caracterizan a la función biológica de la célula. Se trata de grupos de genes que evolucionan de forma colectiva según funciones suaves en el tiempo. En nuestro modelo incorporamos variabilidad a distintos niveles: hay una fracción de genes que no siguen una dinámica modelado por **porcen**, y también existe una variabilidad en la expresión entre genes que pertenecen al mismo grupo modelado por un ruido según **std**.

En el modelo estudiamos la selección de genes por el criterio de tener una varianza entre vecinos S^2 menor que la varianza global σ^2. Para ello, armamos un grafo de células donde los enlaces

se dan según una medida de similitud entre los vecinos, parámetro estudiado, tal que se arman vecindades. Cuantificamos la efectividad de la selección según la distribución de los pasos temporales entre los nodos, el menor camino que une a la muestra inicial y la muestra final y la fracción sobre los genes que les eliminamos el orden en el tiempo que fueron no seleccionados.

Encontramos que para una fracción de genes sin orden temporal baja y media la selección de genes es muy buena según los tres observables estudiados. Sin embargo, para una fracción alta se afirma que el filtrado no es bueno porque no se elimina la totalidad de genes sin orden temporal. Con respecto al ruido dentro del mismo grupo de genes, para un ruido bajo y medio se afirma que la selección de genes es buena mientras que para un ruido alto no lo es. La razón es que se están eliminando genes que sí presentan un orden. Y por último, con respecto a la cantidad de vecinos, vimos que $k = k_c$ (tal que el grafo queda conectado) es una buena relación de compromiso entre una buena distribución de pasos y una buena fracción de genes eliminados.

Capítulo 3

Reconstrucción de variedades de baja dimensión. Modelo sintético II

En el capítulo anterior, Cap.2, desarrollamos un modelo de *matriz de expresión* que simula la dinámica de grupos de genes según funciones parametrizadas por el tiempo. También modelamos una variabilidad en las expresiones simulando un ruido y modelamos genes que no siguen una dinámica. Sin embargo, hay ciertos genes que participan en más de un proceso biológico tal que su nivel de expresión es la suma de las expresiones que se corresponden con cada grupo funcional que representa a cada proceso. También hay genes que al participar en varios procesos, presentan niveles de expresión altos en la mayoría de las células muestreadas.

Esto nos lleva en este capítulo modelar una *matriz de expresión* de un experimento scRNASeq un poco mas realista, que tenga en consideración la participación de genes en más de un proceso. Para hacerlo, tendremos en cuenta la asignación de funcionalidad biológica a genes consignada en la ontología de *Gene Ontology*.

3.1. Idea del modelo y la biología subyacente

El modelo que presentamos considera que dentro de la población total de genes hay un grupo que sigue una dinámica parametrizada por el tiempo, que son los que llamaremos genes dinámicos cuya evolución en el tiempo es compartida por los genes del mismo grupo funcional. También como parte de este grupo, modelamos ciertos genes que participan en más de un proceso.

Por otro lado, modelamos un conjunto diferente de genes que presentan niveles de expresión alto en la mayoría de las células. A estos genes los llamamos genes persistentes. Por último, modelamos un grupo de genes que se expresan de forma estocástica a lo largo de las células muestreadas, que son los que llamamos genes ruido blanco.

La elección de cuáles genes pertenecen a cada grupo es a partir de un conocimiento de a qué procesos biológicos los mismos están asignados. Para ello, utilizamos la ontología *Gene Ontology* (GO) que es una base de datos construida a partir de un vocabulario controlado de conceptos biológicos. La misma está implementada sobre una estructura de grafo dirigido acíclico, donde cada nodo se corresponde a un concepto y los enlaces que los unen caracterizan relaciones del tipo "es un" o "es parte de". Para ilustrar mostramos un subgrafo de *Gene Ontology* (Fig.3.1) donde se puede observar como los conceptos más generales aparecen típicamente en los estratos más altos de la ontología, y los más específicos son descendientes de los mismos.

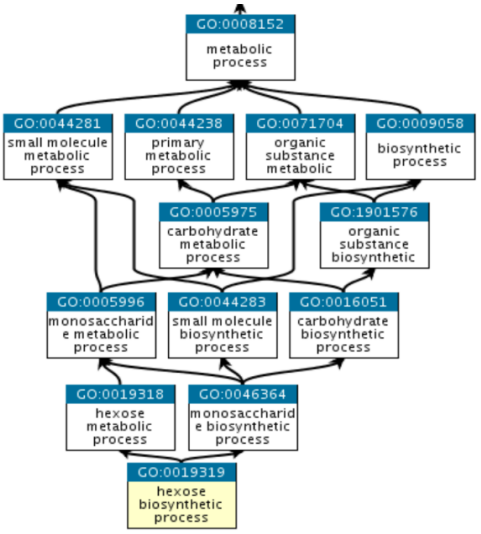


FIGURA 3.1: Subgrafo *Gene Ontology* para conceptos que se enlazan al proceso más general, proceso metabólico. Se observa la forma de grafo dirigido acíclico donde conceptos más específicos son descientes de más generales, e.g proceso metabólico de pequeñas moléculas.

Como vemos, esta red lleva adelante dos objetivos. Por un lado, uniformar los términos y nomenclaturas utilizados, y por otro, organizar los conceptos biológicos de una manera sencilla y accesible, para simplificar la búsqueda de relaciones entre ellos. *Gene Ontology*, en principio, es un vocabulario controlado y organizado de conceptos biológicos. Sin embargo resulta en una valiosa herramienta de análisis porque existen consorcios especializados en la caracterización de diferentes organismos biológicos, que continuamente se ocupan de asociar genes con nodos del grafo de GO de acuerdo a evidencia. Por ejemplo, cuando se encuentra experimentalmente que un determinado gen está relacionado con un determinado proceso biológico o función molecular, se asocia dicho a gen a los nodos respectivos (los más específicos pertinentes) en la ontología y automáticamente quedan también asociados a los nodos ancestros de los mismos.

De esta forma cada nodo del grafo, un concepto biológico, tiene anotados genes que se corresponden a ese proceso. Estos tienen asociado un código de evidencia que indica el tipo de evidencia que soporta la asociación del gen al proceso.

Por otro lado, los conceptos GO se clasifican según tres tipos (o tres sub-ontologías):

- Procesos biológicos (BP): conceptos que describen el proceso biológico en el que participa el gen (i.e. qué es lo que hace).
- Componente Celular (CC): conceptos relacionados con las estructuras celulares donde se ha encontrado que la proteina asociada se localiza poara cumplir su función (i.e. dónde lo hace). Ejemplos incluyen: núcleo, telómeros, complejos de reconocimiento de origen, etc.
- Función Molecular (MF): tareas específicas que llevan adelante las proteinas. Algunos ejemplos son: factor de transcripción, helicasa de ADN, etc.

Ahora que ya hemos presentado y explicado brevemente cómo funciona la ontología de conceptos biológicos GO vamos a explicar en la siguiente sección cómo es que a partir de esta información es que seleccionamos los genes para modelarles una dinámica y así construir un modelo de *matriz de expresión*.

3.2. Construcción del grafo de conceptos GO para armado de grupos de genes

Nuestro objetivo es armar un grafo de conceptos GO que nos permita, a partir de ciertos criterios, seleccionar grupos de genes determinados que obedezcan a ciertos comportamientos dinámicos. Para ello, en primer lugar, vamos a seleccionar los conceptos GO de interés.

Tomamos el criterio de seleccionar aquellos conceptos GO que tienen una cantidad razonable de genes anotados en humanos: decidimos que el rango sea entre 10 y 200 genes. Además pedimos que éstas anotaciones sean por código de evidencia experimental.

Armamos una red bipartita de dos tipos de nodos: *Conceptos GO* (4705), denominados como N_1, que se enlazan con nodos del tipo *gen* que son los anotados a los respectivos conceptos (9991), llamados N_2, siendo 187276 el total de enlaces establecidos entre nodos de uno y otro tipo. Una red $B(N_1, N_2, E)$ se dice bipartita si su conjunto de nodos N puede expresarse como la unión disjunta de dos conjuntos (N_1, N_2). Es decir que verifica $N_1 \cup N_2 = N$, $N_1 \cap N_2 = 0$ y además ningún arco $c_i \in E$ une nodos de un mismo conjunto (N_1 ó N_2). Los elementos de $E = e_1, e_2, \ldots, e_M$, enlaces, son pares diferentes de elementos no-ordenados que enlazan a nodos de tipo-N_1 con otro de tipo-N_2.

Para esclarecer cómo es esta red bipartita, graficamos una subred de la misma en Fig.3.2 que consiste en tres *conceptos GO* (rojo) y cinco *genes* (azul). Los nodos de tipo *conceptos GO* se enlazan con los nodos *genes* que están anotados en dichos conceptos. Nótese que algunos genes están anotados en más de un concepto.

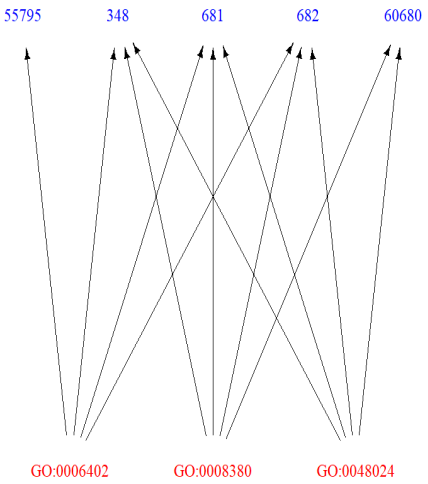


FIGURA 3.2: Subred de la red bipartita de nodos *conceptos GO* (rojo) y nodos *gen* (azul). Se trata de enlaces dirigidos de los nodos concepto GO hacia los nodos genes que están anotados en los mismos respectivamente. Nótese que hay genes que están anotados por más de un concepto.

Para el armado de los grupos queremos limitar los conceptos a aquellos que son biológicamente específicos. Para ello, seleccionamos a aquellos GO que son *hojas*, es decir que no tiene hijos directos en la estructura de árbol. Un ejemplo es el *hexose biosynthetic process* en el árbol de la Fig.3.1. En consecuencia, elegimos 344 conceptos que provienen de distintas ontologías como vemos en Tabla.3.1.

Ontología	Función molecular (MF)	Proceso biológico (BP)	Componente celular (CC)
Nodos	152	119	73

TABLA 3.1: Cantidad de nodos concepto GO considerados para el armado de lo que grupos que se corresponden con cada tipo de ontología.

Proyectamos la red bipartita, que incluye nodos de conceptos GO y genes ($B(N_1, N_2, E)$), para obtener una red compuesta unicamente por conceptos GO (G(N, E)), donde los nodos se enlazan según la cantidad de genes anotados que comparten. Para ello, calculamos la matriz de co-citas (C_{ij}) del grafo. Dos vértices del conjunto de nodos GOs están co-citados si hay otro vértice del conjunto nodos genes que cita ambos (tiene un enlace). El elemento C_{ij} expresa la cantidad de co-citas entre los conceptos GO_i y GO_j.

A partir de la matriz C_{ij} construimos la matriz de adyacencia (A_{ij}) del grafo pesado de conceptos GO (G(N, E)). Esta matriz se define según ec.3.1 donde k_i es el grado del nodo i-ésimo en la red G, es decir el total de genes asociados a dicho concepto. De esta forma, los enlaces entre los GOs son pesados según la cantidad de genes que comparten normalizados por la suma de la

cantidad de genes asociados de cada uno. Graficamos un subgrafo del grafo G(N, E) en Fig.3.3 que resulta de proyectar el subgrafo bipartito que mostramos en Fig.3.2.

$$A_{ij} = \frac{C_{ij}}{\sqrt{k_i k j}} \tag{3.1}$$

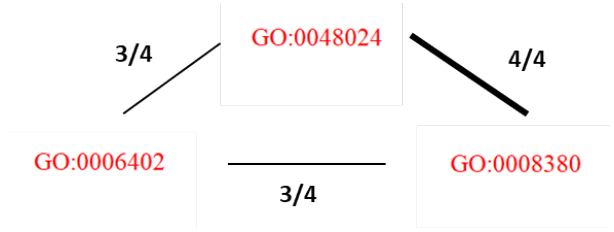


FIGURA 3.3: Subgrafo de G(N, E) de tres nodos que se obtiene de la proyección del subgrafo $B(N_1, N_2, E)$ de Fig.3.2. Se indica el peso de cada uno de los enlaces como la fracción de genes compartidos normalizados según ec.3.1 y también como el grosor de lo enlaces.

Entonces, a modo de resumen, partimos del grafo acíclico dirigido de *Gene Ontology*. A partir de este, armamos la red bipartita B(N, E) siendo las caracterísicas de éste grafo las que se ven en azul en Tabla. 3.2. Esta red la proyectamos en una red de conceptos GO, G(N, E), pesando los enlaces según la cantidad de genes compartidos a partir de la matriz de co-citas C_{ij}, sus características se ven en rojo en Tabla.3.2.

GO	**Genes**	**Enlaces bipartitos**	**GO hojas**	**Enlaces entre GOs**	**Peso enlace medio**
4705	9991	187276	344	475	0.2/1

TABLA 3.2: En azul se muestran características de la red bipartita (B(N_1, N_2, E)) y en rojo de la red proyectada sobre los nodos GOs (G(N,E)) donde nos reducimos a conceptos GOs hojas, que no tienen hijos directos.

Utilizaremos esta red de conceptos GO para definir los diferentes grupos de genes que queremos modelar. Los genes de cada grupo serán los genes mapeados a ciertos grupos de conceptos GO que son seleccionados según un criterio que representa al grupo de genes y la evolución de la expresión de los respectivos genes será común para cada grupo. Se explican los criterios a continuación y se esquematiza en Fig.3.4.

- Consideraremos que los genes persistentes son genes involucrados en varios procesos, por lo que deberían estar mapeados a conceptos GO de alta centralidad en el grafo que llamaremos PGO. El modelado de su evolución a lo largo de la células es el siguiente. Para cada célula seleccionamos un 80 % al azar de los genes anotados en los conceptos de PGO y los prendemos, es decir que les damos un valor de expresión alto.

- Los genes con dinámica siguen una evolución temporal definida parametrizada por el tiempo. Decimos que genes que intervienen en procesos biológicos similares, evolucionan según la misma función dependiente del tiempo. Modelamos grupos de genes que siguen cuatro funciones distintas. Para ello, seleccionamos cuatro grupos de conceptos DGO (A, B, C y D) cuyos genes siguen las funciones f_i respectivamente. Estos grupos son armados a partir de las comunidades del grafo pues estas agrupan nodos similares, donde la similitud viene dada por compartir genes.

- Y por último, los genes que presentan expresiones aleatorias, los RGOs, son una selección aleatoria de genes. Para modelarles un ruido del nivel de expresión para distintas células les otorgmos valores de expresión seleccionados al azar en el intervalo [0,1].

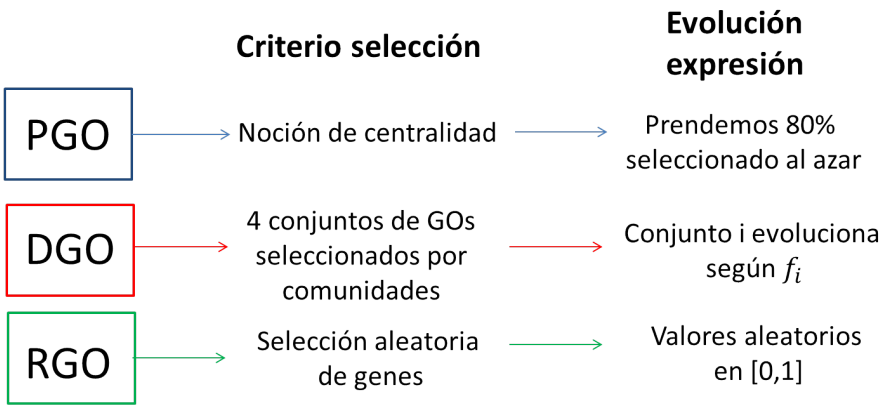


FIGURA 3.4: Esquema de los grupos de conceptos GO, el criterio de selección de cada uno y la forma en que se modela la evolución de los niveles de expresión en la *matriz de expresión* .

Comencemos viendo cómo seleccionamos los PGO. Consideramos dos medidas de centralidad: el grado pesado y el caparazón k, del inglés k core. Tomamos los doce GOs con mayor medida de centralidad según cada una de las medidas y resultaron ser lo mismos. Por otro lado, para determinar los DGOs, particionamos al grafo en comunidades por la técnica llamada edge betweeness en inglés y consideramos las cuatro comunidades que tienen una población por encima o igual que cuatro nodos y seleccionamos subconjuntos de nodos (conceptos GO) que son asociados a los DGO_A, DGO_B, DGO_C y DGO_D.

Estudiamos cómo es la comunicación entre las comunidades resultantes de la partición del grafo calculando la modularidad y la fracción de enlaces entre nodos de comunidades distintas. Los resultados que se obtienen para estas dos medidas nos permiten afirma que se trata de un grafo con mayor comunicación intra-comunidad que la que se espera por azar y tiene pocos enlaces inter-comunidades. En consecuencia, es muy baja la probabilidad de que dos nodos pertenecientes a distintas comunidades compartan genes anotados. Es poco probable que un gen perteneciente a un DGO_i esté contenido también en otro DGO_j. Sin embargo, este es el fenómeno que buscamos modelar pues si un gen pertenece a dos grupos DGO distintos, este

es expresado por los dos grupos funcionales asociados a estos grupos $f_1(t)$ y $f_2(t)$. Así su nivel de expresión será $(f_1 + f_2)(t)$. Para modelar esto, a los cuatro grupos ya construidos de nodos GO les agregamos los nodos pertenecientes a cuatro comunidades que son elegidas al azar de forma de que pueden ser seleccionada una comunidad por más de un conjunto. Así tenemos una probabilidad más alta de que se repitan algunos conceptos GO en distintos conjuntos y por tanto, los genes.

Así construimos a los DGO como dos conjuntos de 16 conceptos GO y otros dos conjuntos con 10 conceptos respectivamente. Identificamos a los genes mapeados a los conceptos de cada conjunto siendo en total 319, 293, 229 y 116 genes que siguen las $f_i(t)$ evoluciones temporales respectivamente. De esta forma modelamos 856 genes en total, de los cuales 66 se repiten en por lo menos dos conjuntos que serán expresados como la suma de los grupos funcionales a los cuales está asociado.

Por último, seleccionamos genes al azar para construir un grupo de genes que llamaremos de ruido blanco, RGO, cuyos niveles de expresión fluctúan de forma aleatoria durante el tiempo. Proyectamos al grafo de los conceptos GO en un grafo de genes y seleccionamos sobre ellos 96 genes al azar.

3.3. Modelado de la evolución temporal

Ya seleccionados los genes para el modelo, resta simular su dinámica para el armado de la *matriz de expresión* (E_{ij}). La cantidad de genes es fija, $N = 1110$, mientras que la cantidad de células M es un parámetro regulable.

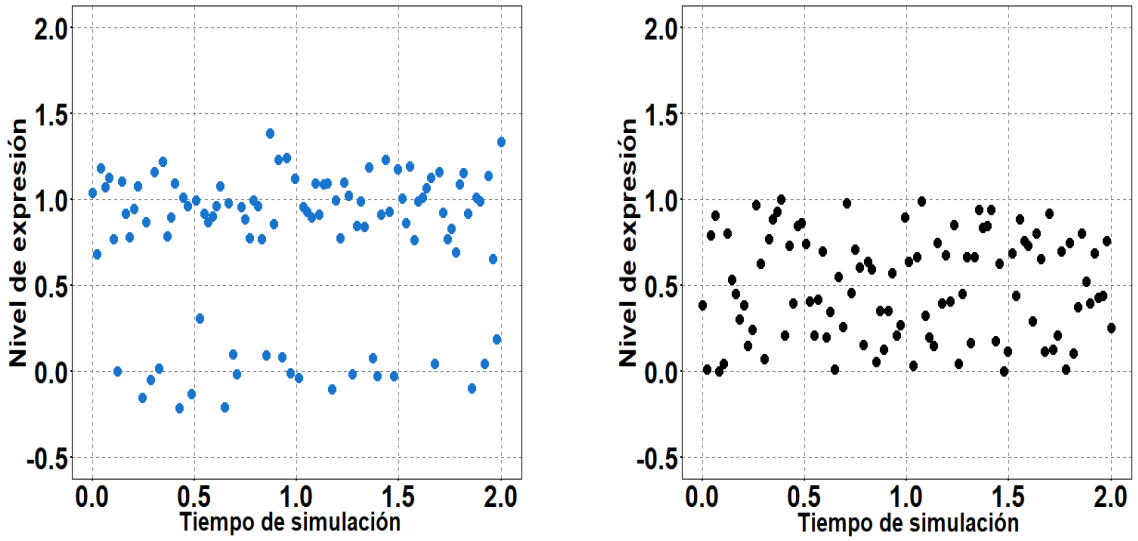
En primer lugar, en la dinámica reconocemos dos estados para un gen: prendido, su nivel de expresión es alto en comparación con las demás expresiones y apagado, cuya expresión es lo opuesto, es baja. De esta forma, es posible identificar a genes que al estar prendidos están involucrados activamente en el proceso biológico que asumimos que caracteriza a la célula en el tiempo en que fue muestreado dicho nivel de expresión del gen. Por el contrario, diremos que un gen que está apagado no participa del proceso.

La forma de simular este compotamiento es modelando a genes prendidos como aquellos cuya expresión tomará un valor dentro de una distribución normal centrada en 1 mientras que un gen apagado, lo tomará de una distribución centrada en el 0. El valor de la desviación estándar vendrá dado por la restricción que impondremos de poder distinguir un comportamiento del otro. Vimos que para $sd = 0.15$ se cumple esta condición de forma que se encontró una cota tal que para expresiones (E_{ij}) por debajo de esta, diremos que el gen i está apago en la célula j y por encima, prendido:

$$E_{i,j} = \begin{cases} > 0.5 & \text{gen prendido} \\ < 0.5 & \text{gen apagado} \end{cases}$$

Una vez ya definidos estos dos estados, modelamos estados intermedios que vienen por la dinámica en el tiempo. En particular, nosotros modelamos un grupo de genes que llamamos persistentes cuya expresión es de tipo constante, estándo prendidos, al que le simularemos ruido técnico seleccionando aleatoriamente, para cada célula, un 20 % de genes persistentes que serán apagados.

Por otro lado, para construir la evolución de los genes que funcionan como ruido blanco asociamos al nivel de expresión un valor seleccionado aleatoriamente en el rango [0,1]. De esta forma, en cada una de las muestras de célula única, cada uno de estos genes expresa un nivel de expresión aleatorio acotado. Supongamos como ejemplo un modelo con 100 células. Entonces, el vector tiempo de simulación, t, está consittuido por 100 puntos equiespaciados en [0,2]. Graficamos la evolución de un gen persistente y de uno ruido blanco para entender la diferencia (Fig. 3.5).



(a) Genes persistentes

(b) Genes ruido blanco

FIGURA 3.5: Evolución modelada en el tiempo de simulación en una matriz de 100 células para dos tipos distintos de genes.

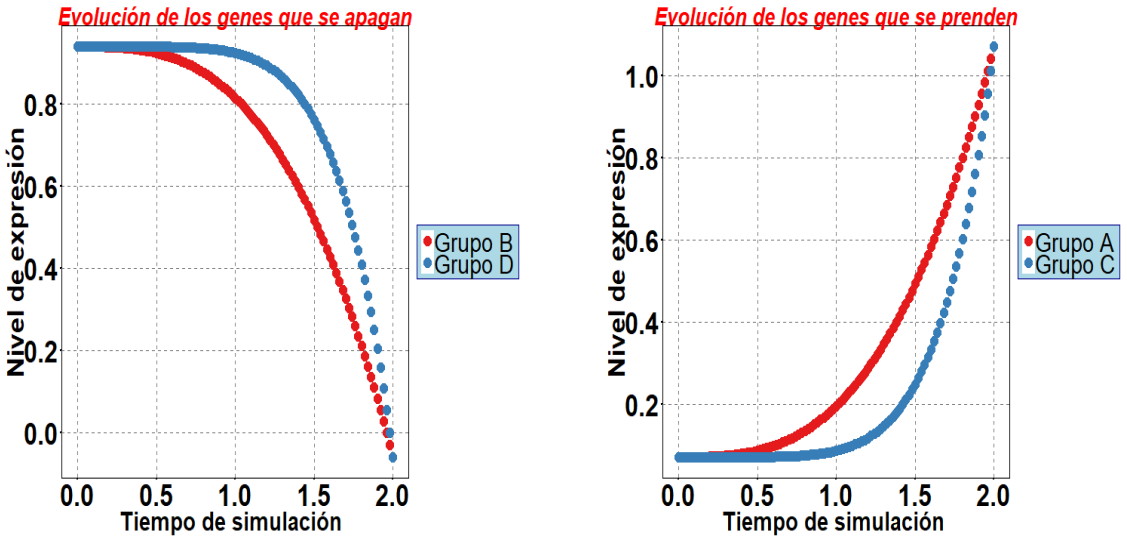
Observamos que para el caso de los persistentes, en el panel A, el nivel de expresión es acotado en un rango que viene dado por el ancho de la campana de la distribución y se dan expresiones de apagado (en entorno del 0) que se deben a que en cada célula se apaga un 20 % de los genes persistentes seleccionados al azar. Nótese que como la distribución de valores que toma un gen apagado es una normal de $\mu = 0$ y $std = 0.15$, se dan expresiones nulas. Sin embargo, los niveles de expresión de genes se definen para valores positivos. Por ello, colapsaremos a todos los valores < 0 a valores nulos. Por otro lado, vemos en el panel B que los genes ruido blanco muestran la aleatoriedad en sus niveles de expresión en el rango [0,1].

Con respecto a los genes dinámicos, asociamos los conjuntos DGO_A y DGO_C a funciones que prenden a los genes, funciones crecientes sueves en el tiempo, f_1 y f_2 (ec.3.2) mientras que los conjuntos DGO_B y DGO_D son asociados a funciones que apagan a los genes, decrecientes, f_3 y f_4 (ec.3.2). A su vez, imponemos que la diferencia entre las respectivas funciones sea la velocidad con que crece o decrece respectivamente. Utilizamos funciones polinómicas cuyo exponente es una medida de la rapidez con que crece/decrece el nivel de expresión. Elegimos usar t^3 y t^6, donde la última crece/decrece al doble de velocidad.

Las constantes multiplicativas (A_i) y de adición (c_i) para las funciones en ec.3.2 se encuentran determinadas para que las funciones $f_1(t)$ y $f_2(t)$ tengan como estado inicial el de apagado y como final, el de prendido y las funciones $f_3(t)$ y $f_4(t)$, al revés. Así c_i para $i = 1, 2$ es un valor dentro de la distibución de gen apagado y para $i = 3, 4$ es de la distribución de gen prendido. Por otro lado, el vector tiempo de simulación son N valores equiespaciados entre 0 y 2 donde se deja como parámetro a fijar N para poder regular la cantidad de células muestreadas en el modelo.

$$\begin{aligned} f_1(t) &= A_1 t^3 + c_1, \quad A_1 = \tfrac{1}{8} \\ f_2(t) &= A_2 t^6 + c_2, \quad A_2 = \tfrac{1}{64} \\ f_3(t) &= A_3 t^3 + c_3, \quad A_3 = -A_1 \\ f_4(t) &= A_4 t^6 + c_4, \quad A_4 = -A_2 \end{aligned} \tag{3.2}$$

Graficamos la evolución en el tiempo de simulación de los genes dinámicos cuyas expresiones obeceden la funciones en Fig. 3.6. Observamos en el panel A la evolución asociada a los genes de los grupos DGO_B y DGO_D que comienzan sus trayectorias en un estado de prendido y terminan en uno apagado. Vemos en el panel B que los genes de los grupos DGO_A y DGO_C muestran un comportamiento inverso, comienzan en un estado apagado y terminan en un estado final de prendido.



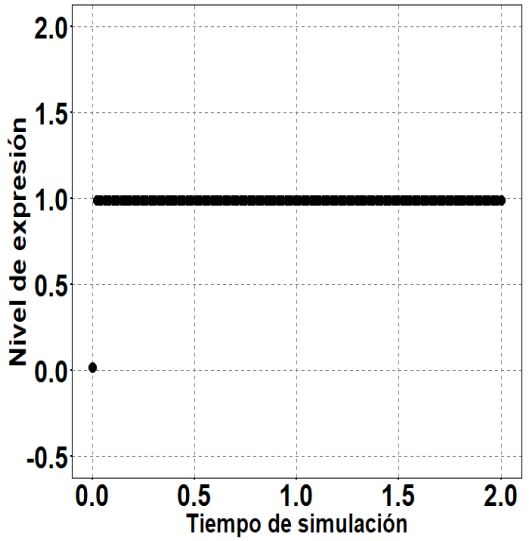
(a) Evolución creciente

(b) Evolución decreciente

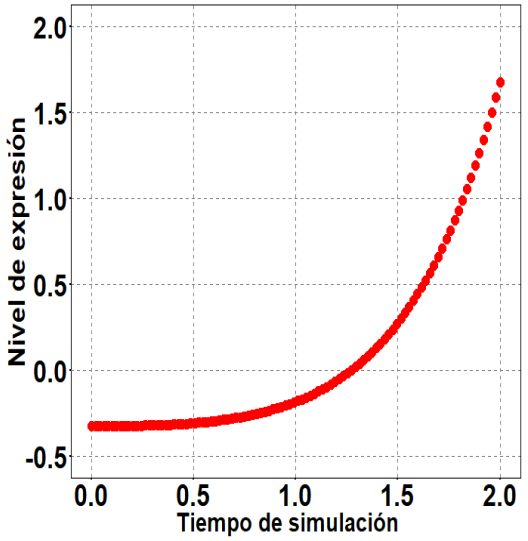
FIGURA 3.6: Evolución de los genes según cada grupo funcional dinámico que siguen respectivas funciones parametrizadas por el tiempo de simulación, según t^6 (rojo) y según t^3 (azul). Se muestra a costado la asociación del color a cada grupo.

Como ya se comentó, modelamos 66 genes que son expresados por más de un grupo dinámico de forma que se encuentran anotados en más de un conjunto DGO. En consecuencia, su nivel de expresión se modela como $\sum_i f_i(t)$ donde i representa a cada uno de los DGO en que el gen se encuentra anotado. Graficamos las evoluciones para algunos de ellos en Fig.3.7 donde se indica en cada pie de figura en qué grupos dinámicos se expresa. Se observa en el panel A que la evolución de un gen que es expresado por DGO_A y por DGO_B que muestra un comportamiento similar al de un gen persistente al haber sido expresado por estos dos grupos.

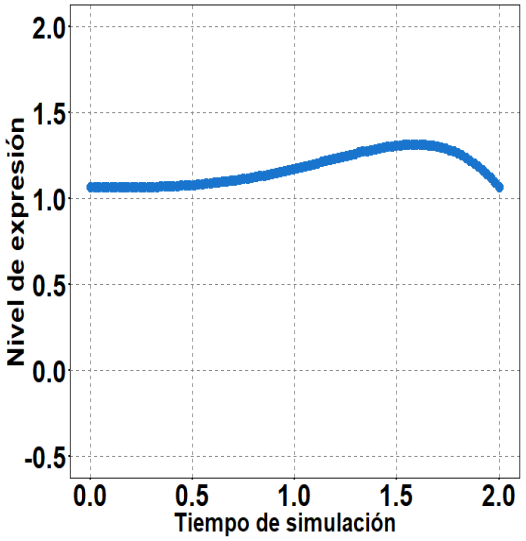
También vemos en el panel B la evolución de un gen que llega a tomar, hacia el final del intervalo del tiempo de simulación, valores por encima de 1.5. Por otro lado, se puede ver que se toman valores por debajo de 0, estos van a ser colapsados a valores nulos tal como explicamos antes. Por último, observamos en el panel C que se da un comportamiento similar al de un persistente pero, con un poco más de variabilidad en un rango superior a 1 ($[1, 1.5]$).



(a) Gen expresado por DGO_A y DGO_B, $(f(t) = f_1 + f_3)(t)$.



(b) Gen expresado por DGO_A y DGO_C $f(t) = (f_1 + f_2)(t)$.



(c) Gen expresado por DGO_A y DGO_D, $f(t) = (f_1 + f_4)(t)$.

FIGURA 3.7: Evoluciones en el tiempo de simulación de niveles de expresión de genes que son expresados por dos grupos DGO distintos tal que sus niveles de expresión se modelan como la suma de la expresión de cada grupo.

Por lo visto de que algunos genes que son expresados por más de un grupo dinámico alcanzan valores por encima del 1.5, este valor resulta una buena cota tal que aquellos genes que tomen valores por encima de dicho valor están siendo sobre-expresados. Es así como se tienen definidos rangos de expresiones común a todos los genes de la matriz tal que podemos identificar si están siendo no expresados, expresados o sobre-expresados según ec. 3.3.

$$E_{i,j} = \begin{cases} < 0.5 & \text{gen no expresado} \\ > 0.5, < 1.5 & \text{gen expresado} \\ > 1.5 & \text{gen sobre-expresado} \end{cases} \tag{3.3}$$

A modo de resumen, construimos un modelo de *matriz de expresión* scRNASeq que simula distintas dependencias de las expresiones de los genes con el tiempo de simulación agrupándolos en cuatro grandes grupos cuya fracción de genes se muestra en Tabla.3.3.

Estos son los genes persistentes (PGO) que, con una cierta variación, están prendidos en la totalidad de células. El conjunto de genes con dinámica (DGO) que se subdivide en dos grupos. Uno que llamamos dinámicos únicos, subdividido a su vez en cuatro grupos cuya evolución está determinada por una función parametrizada por el tiempo (f_i, $i = 1, 2, 3, 4$) respectivamente. El segundo grupo que conforma a los DGO es el de los dinámicos múltiples compuesto por genes que son expresados por más de uno de lo cuatro grupos de los dinámicos únicos. Sus niveles de expresión son el resultado de sumar los niveles de expresión correspondiente a cada grupo donde se encuentra anotado.

Y por último, los genes ruido blanco que toman valores aleatorios en el intervalo $[0, 1]$ y modelan un típico ruido blanco experimental. De esta forma construimos una *matriz de expresión* de 1110 genes y N células, que es un parámetro libre.

Persistentes	**Dinámicos únicos**	**Dinámicos múltiples**	**Ruido blanco**
0.14	0.71	0.06	0.09

Tabla 3.3: Fracción de genes que se corresponden a cada una de las categorías que modelamos.

3.4. Selección de genes por noción de suavidad entre células vecinas

Una vez armado por completo el modelo, en esta sección vamos a estudiar la selección de genes según el criterio de la suavidad entre células vecinas representado por la varianza entre vecinos S^2 que fue estudiado en Secc.2.2. Para ello, construimos el modelo explicado para 100 células. El vector t que representa al tiempo de simulación será una sucesión de 100 valores equiespaciados en el rango [0,2].

Tal como hicimos en Secc.2.2, a partir de la *matriz de expresión*, calculamos las distancias euclideanas entre los perfiles (muestras de célula única), y armamos la matriz de adyacencia del grafo donde los nodos representan a las células. La cantidad de vecinos k es uniforme a todos

los nodos y es el k_c la menor cantidad de vecinos tal que el grafo queda conectado. La decisión se debe a que encontramos que es el valor más óptimo (ver Secc. 2.3.3).

En las siguientes dos subsecciones cuantificamos qué tan buena fue la selección de genes según la reconstrucción de los pasos en el grafo de muestras tal como hicimos en Secc. 2.3.3 y también según la fracción de los genes que fueron seleccionados para cada grupo modelado (los persistentes, los dinámicos y los de ruido blanco). Diremos, en este caso, que una buena selección de genes es aquella que selecciona sólo a aquellos que presentan dinámica y elimina a los genes persistentes y los de ruido blanco.

3.4.1. Fracción de genes eliminados

Al realizar la selección se eliminó una fracción de 0.23 genes. Para ver cuáles genes fueron removimos, graficamos los valores de ambas varianzas en función de cada gen (Fig.3.8) . Aquellos genes con un índice contenido en $[1, 158]$ son persistentes, en $[159, 1014]$ son dinámicos y en $[1015, 1110]$ son de ruido blanco. Se observa cómo se distribuye la fracción de eliminación en los distintos grupos de genes. Para los genes persistentes (1-158) y los de ruido blanco (1015-1110) el valor de la varianza entre vecinos (S^2) es significativamente mayor que la global (σ^2), de forma que la mayoría de los genes, de ambos grupos, son eliminados. Una fracción de 0.94 de los persistentes y un 0.99 de los de ruido blanco.

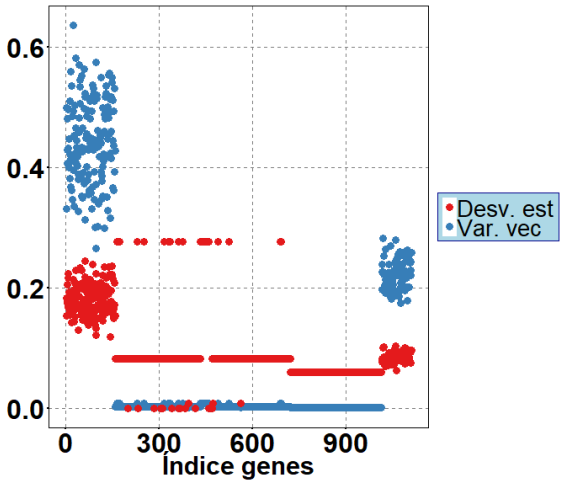


FIGURA 3.8: Desviación estándar de cada gen con respecto a la totalidad de genes de la *matriz de expresión* (rojo) y la varianza con respecto a los vecinos de cada gen (azul) en función de los genes. Se pueden visualizar 3 regiones delimitadas con a grupos de genes. Grupo de genes persistentes (1-158), dinámicos (1-1014) y los de ruido blanco (1015-1110).

También observamos que los genes que siguen una evolución temporal determinada, aquellos con índice contenido en [159,1014], tienen asociada una varianza entre vecinos (S^2) que es significativamente menor que la global (σ^2). Es éste el comportamiento que esperamos pues según el modelo propuesto, la dinámica biológica del proceso estudiado se encuentra expresada

en éste grupo de genes. Sin embargo, notamos que algunos genes son eliminados (fracción de 0.01 del grupo).

Para más detalle respecto a este grupo de genes, graficamos la diferencia entre éstas magnitudes estadísticas ($S^2 - \sigma^2$) (Fig. 3.9). Nótese que se dan cuatro regiones. La primera se observa como una línea continua en el rango de [159, 718] de genes que se corresponden a los conjuntos DGO_A y DGO_B que evolucionan como t^3 mientras que la otra línea continua en el rango de [719, 1014] de genes se corresponde a DGO_C y DGO_D que van como t^6 (ver Fig.3.6). También reconocemos un región para valores negativos y en valor absoluto los más grandes que se corresponden a genes que podemos considerar como que son ampliamente seleccionados y otra, para valores cercanos al cero, que serían genes que no presentan una variabilidad entre vecinos significativamente menor que la global.

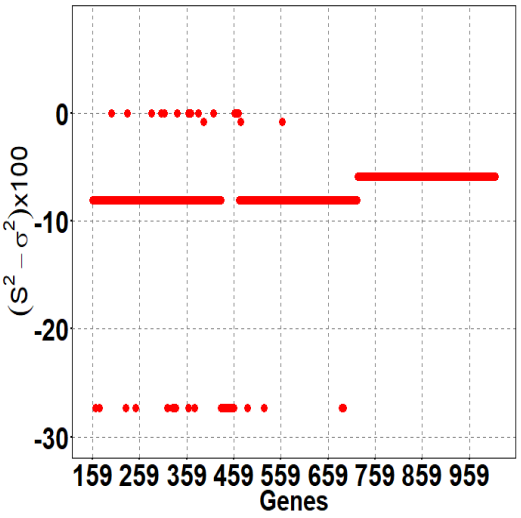


FIGURA 3.9: Diferencia entre varianza entre vecinos (S^2) y la global (σ^2) (x100). El criterio empleado elimina a un gen si $S^2 - \sigma^2 > 0$.

Aquellos que toman valores cercanos al cero, en su mayoría (fracción de 0.83), se corresponden a genes anotados tanto en DGO_A y como en DGO_B, tal que sus expresiones son definidas por $(f_1 + f_3)(t)$. Toman valores de expresión en un entorno de un valor constante dado por las condiciones iniciales (ver en Fig.3.10 la forma funcional de la sumatoria). La fracción 0.17 restante se compone, en misma proporción, de genes anotados en DGO_A y DGO_D y, en DGO_B y DGO_C respectivamente. Son expresiones que toman valores acotados en un rango (como se ve en el panel C de Fig. 3.7). Para el caso de la suma de expresiones que se corresponden a DGO_B y DGO_C, el rango es acotado para valores bajos de expresión y son estos eliminados, mientras que los que se corresponden a DGO_A y DGO_D el rango se compone de expresiones altas, por tanto son seleccionados.

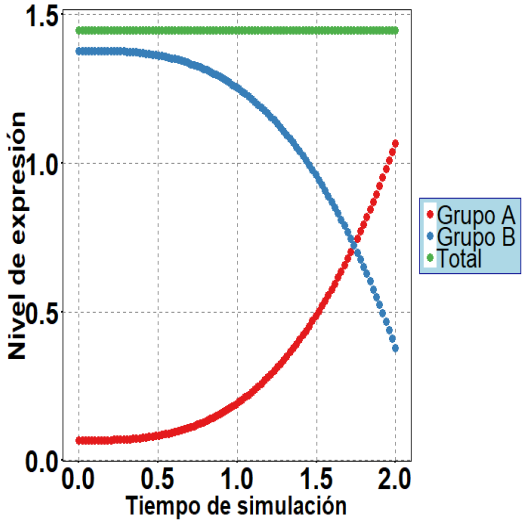


FIGURA 3.10: Nivel de expresión en función del tiempo de simulación que se corresponde a un gen pertenience al grupo DGO_A (rojo), uno del DGO_B (azul) y uno que es doblemente expresado por ambos (verde). Se observa que se modela la expresión del doblemente expresado como la suma de ambos niveles.

También se distingue una componente que toma los valores de $S^2 - \sigma^2$ más grandes en valor absoluto y negativos que se corresponden, en su mayoría (fracción de 0.9), a genes doblemente expresado por DGO_A y DGO_C, con una evolución modelada por $(f_1 + f_2)(t)$. Los niveles de expresión crecen como $t^3 + t^6$ y llegan a tomar (en el tiempo final) los valores más altos, mayores a 1.5 (ver panel B de Fig.3.7). Esto indicaría, por cómo definimos a los estados de los genes en ec.3.3, que están siendo sobre-expresados.

En consecuencia, la puntuación otorgada por este criterio permite reconocer un grupo de genes que son ampliamente expresados por dos funciones crecientes suaves. Al trabajar con un data-set experimental, esto nos va a permitir detectar aquellos genes que al participan activamente en más de un proceso biológico.

A modo de resumen, se muestra en Tabla 3.4 la fracción de genes que fueron eliminados por el criterio para cada grupo. Para el caso de los dinámicos múltiples se elimina una fracción de 0.18 que se corresponde a los casos en que la expresión permanece constante o acotada, un comportamiento similar al de los genes persistentes. Son seleccionados aquellos que son sobre-expresados tal que alcanzan, hacia el final del tiempo de simulación, valores por encima de 1.5. Por otro lado, los genes que siguen una dinámica dada por una única función parametrizada por el tiempo de simulación son seleccionados por completo y aquellos que son persistentes y de ruido blanco son eliminados en su mayoría pues la fracción de 0.01 de genes aceptados resulta insignificante.

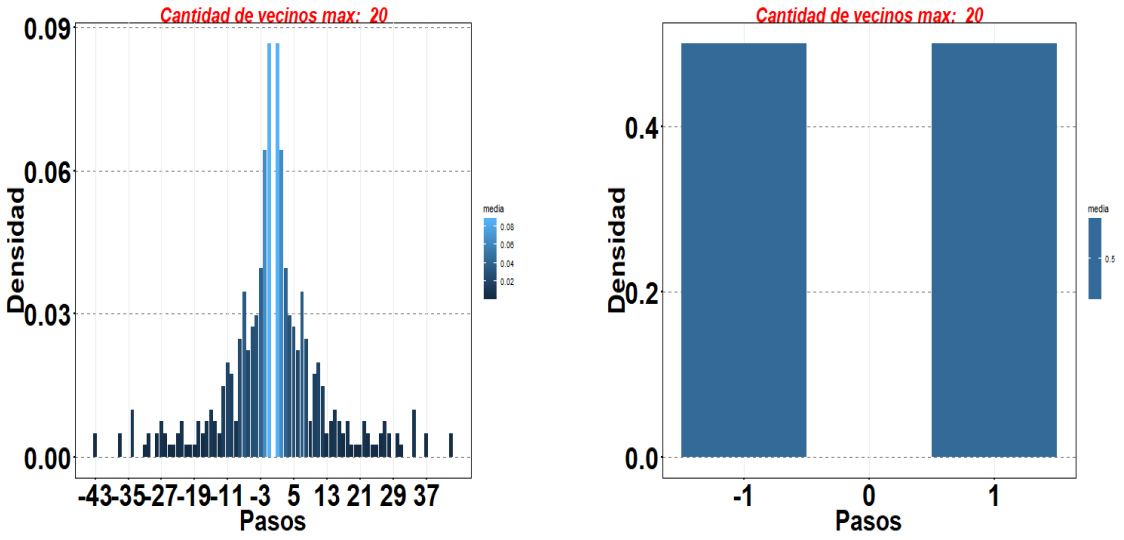
	Persistentes	**Dinámicos únicos**	**Dinámicos múltiples**	**Ruido blanco**
Fracción eliminados	0.99	0	0.18	0.99

TABLA 3.4: Fracción de genes eliminados según cada grupo modelado. La fracción es relativa a la cantidad de genes que pertenece a cada una de sus grupos respectivamente. Los dinámicos múltiples se refiere a genes que son expresados por dos o más grupos dinámicos tal que su expresión es la suma de la expresión de cada grupo.

3.4.2. Distribución de los pasos temporales

Otra forma de cuantificar la performance del filtro es comparando la distribución de pasos entre una célula y sus vecinas como se hizo en Secc. 2.3. Para ello, graficamos la distribución de los pasos del grafo previo al filtrado y después del mismo en Fig.3.11. Observamos en el panel A que antes de aplicar el filtro, el grafo queda conectado para $k = 3$ y la distribución tiene asociada una densidad no nula para pasos muy grande en valor absoluto y los valores de densidad más significativos se dan en el rango [0,5].

Al filtrar, para $k = 2$ el grafo queda conectado, es decir que todo nodo muestra del grafo tiene grado hacia afuera de 2. Observamos en el panel B que la distribución de pasos se limita a un paso, a -1 y 1, indicando que todo nodo muestra tiene como vecina a su respectiva muestra asociada a un tiempo de simulación previo y posterior respectivamente. Esto nos lleva a afirmar que el filtrado de genes permite visualizar el orden dado por el tiempo de simulación cuya evolución gobierna la dinámica.



(a) Distribución del grafo previo a ser filtrado.

(b) Distribución del grafo filtrado.

FIGURA 3.11: Histogramas de la distribución de pasos de los primeros vecinos de los nodos muestras. Se encontró que para $k = 3$ el grafo muestras previo a ser filtrado queda conectado y para el grafo filtrado éste queda conectado para $k = 2$.

A partir de los resultados discutidos de las fracciones de genes filtrados para cada grupo asociado a una dinámica particular en tiempo de simulación y también a partir de cómo cambia la distribución de pasos que se da sobre los nodos enlazados en el grafo al considerar sólo los genes de interés según este criterio, podemos afirmar que dicho criterio resulta satisfactorio para este modelo de datos scRNASeq.

A modo de resumen, en este capítulo construimos un modelo de *matriz de expresión* más robusto en el sentido de que considera que hay genes que son expresados por más de un grupo funcional. La selección de los genes la hicimos a partir del conocimiento de a qué proceso biológico está asociado el gen utilizando la ontología GO. Armamos un grafo bipartito de conceptos GO seleccionados y de los genes anotados a los mismo que luego proyectamos en un grafo unipartito de conceptos GO que nos permitió a partir del mismo modelar grupos de genes que siguen funciones suaves en el tiempo y también genes cuyas expresiones siguen más que una función parametrizada por el tiempo. Por otro lado, modelamos también genes que se expresan en la mayoría de las células, los persistentes y genes no siguen una evolución parametrizable por el tiempo, los ruido blanco.

Realizamos la selección de genes por la noción de suavidad en las vecindades cuantificada por la varianza entre vecinos y cuantificamos su efictividad según la fracción de genes eliminados para cada grupo y por la distribución de pasos temporales entre nodos enlazados en el grafo. Encontramos que la selección es muy buena porque los genes persistentes y de ruido blanco son eliminados en un su mayoría (fracción de 0.99) mientras que los genes que expresan una única dinámica son en su totalidad seleccionados.

Con respecto a aquellos genes que presentan una dinámica múltiple (por más de un grupo funcional), son eliminados solamente aquellos que, por cómo se da la forma de la expresión, muestran valores constantes. Aquellos que llegan a tomar valores altos de expresión son los que tienen asociadas una puntuación más alta en valor negativo, de forma que es detectado que son genes que son sobre-expresados y con una posible implicancia biológica relevante.

Capítulo 4

Expresión de célula única en el giro dentado del hipocampo de ratones

En los capítulo anteriores hemos estudiado la técnica de selección de genes para un modelo de *matriz de expresión* simple y otro más robusto en el sentido biológico de la dinámica y encontramos que la técnica de Slicer que selecciona genes según la noción de variabilidad suave en vecindades de células resulta satisfactoria. Proyectando a las muestras de transcriptoma en un grafo, con sólo aquellos genes seleccionados, se logra la reconstrucción de una VBR compatible con los datos simulados. En este capítulo nos proponemos aplicar esta técnica de selección de genes a un data-set experimental de scRNASeq de células del giro dentado del hipocampo de ratones que fue medido y estudiado por el Dr. Linnarson y colaboradores (Hochgerner et al., 2018).

4.1. La tecnología de célula única scRNASeq

Como mencionamos en la introducción, gracias a la tecnología scRNASeq es posible observar simultáneamente los niveles de expresión de miles de genes que tiene lugar dentro de decenas/miles de células, de a una por vez (Eberwine et al., 2014) (Nawy, 2013). Así, los datos scRNASeq nos permiten monitorear estados y cambios transcripcionales con un detalle sin precedentes.

Una manera de estudiar estos experimentos consiste en analizar el ensamble de mediciones en un espacio de estados celulares. Existe mucha información en la manera en que los puntos asociados a estados celulares se distribuyen en este espacio. Zonas densamente pobladas pueden ser asociadas a estados celulares estables y/o metaestables relacionados con, por ejemplo, tipos celulares definidos, así como criterios de proximidad y conectividad en este espacio pueden utilizarse para estimar un ordenamiento *pseudo-temporal* que de cuenta de trayectorias dinámicas entre los mismos.

El método de célula única (scRNASeq) no es uno sólo, sino que es una colección de protocolos compatibles con varias aplicaciones. Por ejemplo, una aplicación popular es la de identificar poblaciones de células raras (fracción < 0.01) para lo cual hay que examinar grandes cantidades de células. Por ejemplo, secuenciando 20921 células del hipotálamo de ratones se han podido identificar una sub-población neuronal de menos de 50 células (fracción < 0.002) (Campbell et al., 2017).

El protocolo de un experimento scRNASeq consiste en tres grandes pasos: i) aislamiento de una célula única, ii) preparación de la librería de ADNc, y iii) la secuenciación.

- i) El aislamiento de las células requiere de la disociación de la muestra seguido por una clasificación en partes separadas por captura de células individuales en gotas (droplets) o en cámaras de microfluídica. Posteriormente, se realiza la captura del ARNm (ARN mensajero) celular, guardando informacíon sobre la provenencia celular de cada molécula.
- ii) La preparación de la librería involucra transcripción inversa para convertir el ARNm en ADNc (ADN complementario). Se hace una amplificación (por PCR o por transcripción in vitro) del ADNc en su totalidad o del extremo $3'$ o $5'$, de acuerdo al protocolo utilizado. La tasa de amplificación varía desde 25000 lecturas por célula (Macosko et al., 2015) hasta un promedio de 5 millones (Kolodziejczyk et al., 2015). Sin embargo, se ha demostrado que para determinar la identidad de tipo de una célula es suficiente entre 25000 y 50000 lecturas por célula (Jaitin et al., 2014).
- iii) Finalmente se realiza la secuenciación de las moléculas de ADNc de la librería, utilizando plataformas de secuenciación (Van Dijk et al., 2014).

Estos protocolos son típicamente utilizados en conjunto con identificadores únicos moleculares (UMIs, del inglés unique molecular identifiers) y/o con RNA exógenos para resolver el alto ruido técnico. Los UMIs son secuencias de entre 4 y 12 nucleótidos que funcionan como códigos de barras únicos de los 5' o 3' finales de cada ARNm individual en cada transcripto. Son colocados previo a la transcripción inversa. De esta forma, es posible marcar con un identificador único a las múltiples copias que surgen del proceso de amplificación, de manera de poder estimar el número de moléculas originalmente presentes en la muestra y reducir el bias producto de la amplificación (Dal Molin and Di Camillo, 2018). Tomemos como ejemplo el caso esquematizado en Fig.4.1 donde se observa un transcripto X que se expresa en 4 copias y uno Y que se expresa en 6.

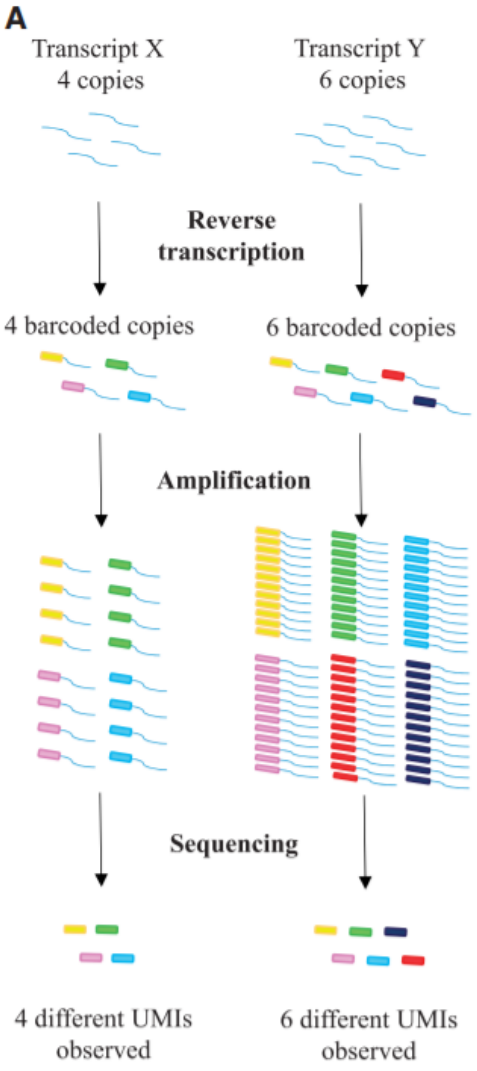


FIGURA 4.1: Esquema del proceso de asignación de UMIs a los transcritomas X e Y. Se observa la conservación de las copias de cada uno. Figura extraída (Dal Molin and Di Camillo, 2018).

Observamos en Fig.4.1 que estos transcriptos van a tener asignados 4 y 6 UMIs diferentes previo a la amplificación. Durante la amplificación estos son multiplicados, pero en la sequenciación se reconoce la cantidad de UMis y por tanto, la cantidad de copias. Así la abundancia de las copias de los respectivos transcriptos es conservada. Sin embargo, los experimentos scRNASeq se caracterizan por una limitación técnica inherente que es la ineficiente captura del RNAm, estos eventos son denominados *drop-outs* en inglés (Grün et al., 2014).

Uno de los grades desafíos del análisis de datos que brindan estos experimentos es la poca fracción de información debido a los *dropouts*. La ineficiente captura induce la aparición de expresiones nulas en la *matriz de expresión* haciendo que esta sea muy esparsa. La fracción de elementos nulos depende del protocolo utilizado y típicamente van desde el 40 % hasta el 95 %. Más adelante en la tesis veremos los efectos que esto trae y cuáles son nuestras propuestas para mitigarlos.

4.2. Introducción a los datos experimentales y análisis de calidad

Para esta parte del trabajo, consideramos los datos experimentales generados en el contexto del trabajo (Hochgerner et al., 2018) liderado por Hannah Hochgerner y Amit Zeisel en el Linnarson Lab. Los colaboradores se propusieron estudiar la diversidad de los tipos celulares en el giro dentado del hipocampo en ratones. Se trata de una región del cerebro en la cual la neurogénesis persiste en la adultez. Sin embargo, hasta el momento la relación entre el desarrollo y la neurogénesis en la giro dentado no ha sido examinada en detalle. Por eso en este trabajo estudian la dinámica molecular y la diversidad de células en ratones perinatales, juveniles y adultos.

Tomamos los datos de la base de datos GEO (Gene Expression Omnibus) del NCBI (National Center for Biotechnology Information) de código GSE95312 que consisten en una *matriz de expresión* de 14545 genes y 5454 células correspondientes a experimentos scRNASeq. Las muestras fueron tomadas en dos días juntando muestras de ratones machos y hembras en un único experimento. El giro dentado de los machos fue diseccionado y disociado en los días post-natales 24 (P24) y 35 (P35) y el de las hembras fue en P12 y P16.

En la *matriz de expresión* E_{ij} figura el identificador para cada gen y célula. Los genes se identifican con un ID y las células con otro. Al mismo tiempo, se consigna en qué día post-natal la célula fue muestreada de forma que cada ID tiene asociado el día de la muestra. En la tabla 4.1 se muestra la cantidad de células que se midieron para cada día post-natal. Adicionalmente, recopilamos la información sobre el tipo celular siendo estos 22 (Hochgerner et al., 2018).

Día post-natal	12	16	24	35
Cantidad células	1129	1440	1063	1822

Tabla 4.1: Distribución de la cantidad de células para cada día post-natal en que se realizaron mediciones.

Cada elemento de la *matriz de expresión* E_{ij} da cuenta de la cantidad de moléculas únicas (i.e. UMI's diferentes) asociados a la expresión del gen i-ésimo en la célula j-ésima. Se trata de una matriz esparsa, la mayoría de sus elementos son cero (un 90 %) debido a los *dropouts* que fueron presentados en Secc.4.1.

Realizamos un análisis de calidad de las células y genes estudiados (representados en la *matriz de expresión*) para seleccionar a aquellas células y genes que sean representativos y relevantes para continuar nuestro analisis. En particular, consideraremos a una célula de interés si satisface los siguientes criterios (Hochgerner et al., 2018).

i La fracción moléculas detectadas por gen sea mayor a 1.2.

ii Al menos la expresión de 600 genes haya sido relevada en la célula.

iii El número de moléculas relevadas total $\in [800, 20.000]$.

(i) Cada elemento de la *matriz de expresión* E_{ij} es el número de moléculas que se midió para el gen i-ésimo en la célula j-ésima. De esta forma, para la célula j-ésima el total de moléculas es $\sum_{i=1}^{14545} E_{ij}$ que normalizamos por el total de genes que consideramos que son expresados. Se dice que un gen es expresado si la cantidad de moléculas para ese gen, dada la célula, es mayor o igual que 1.

Graficamos el promedio de moléculas por gen para cada una de las células identificadas por un índice y por el tipo según el color (Fig. 4.2). En línea roja se indica el valor cota (Hochgerner et al., 2018) tal que sólo consideramos a las células que tienen valores por encima de esta. Se observa que el intervalo de valores es acotado y más o menos constante para todos los tipos de células aunque algunas del tipo endothelial toman valores mucho más altos. Se afirma que la población completa de células cumplen ésta condición.

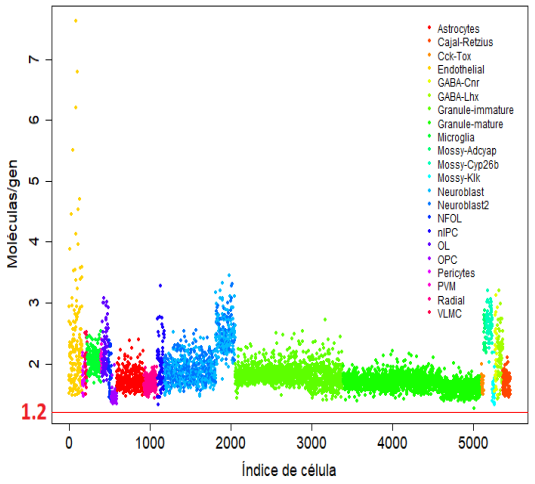
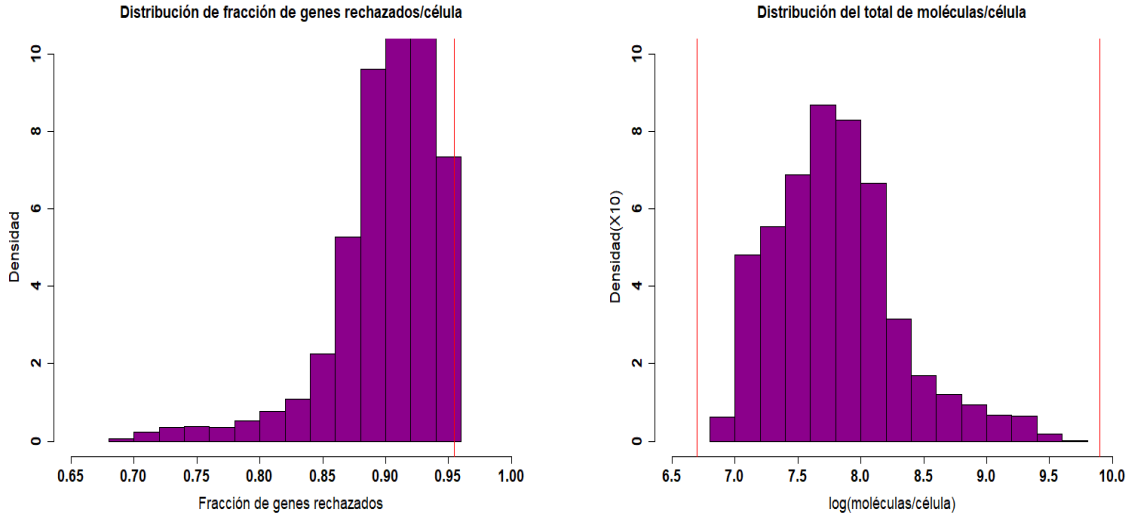


FIGURA 4.2: Fracción de moléculas por gen para cada una de las células en función del índice de cada célula. Se considera sólo las cuentas (moléculas) por gen mayores o iguales que 1. Se identifica cada tipo de célula con los colores y se marca en rojo el valor cota de 1.2 propuesto (Hochgerner et al., 2018).

(ii) Como ya presentamos, este tipo de experimentos se caracterizan por los *dropouts* que son una ineficiente captura del RNAm. La consecuencia de esta limitación técnica es que la mayoría de los genes no son expresados en las células y son genes eliminados del análisis. Buscamos trabajar sobre células que son informativas en el sentido de que expresen una fracción de genes que sea razonable. El criterio adoptado (Hochgerner et al., 2018) es considerar células que expresan más de 600 genes (fracción de 0.04) que es análogo a considerar como cota máxima de fracción de

genes eliminados (*dropouts*) de 0.96. Graficamos la distribución de la fracción de genes por célula que no son expresados, denominados genes rechazados, en el panel A de Fig.4.3. Observamos que en promedio las células no expresan en un 0.90 de la fracción de genes. En la línea roja se indica la cota tal que las siete células que tienen asociada una fracción por encima de esta son eliminadas.

(iii) El criterio consiste en no considerar a aquellas células con muy bajo o muy alto nivel de expresión global. Para determinar cuáles son estas, calculamos la cantidad de moléculas para cada célula y realizamos un histograma (panel B de Fig. 4.3) para ver su distribución. El rango determinado por el criterio se expresa por las líneas rojas en la figura. Se ve que todas las células muestreadas cumplen dicha condición.



(a) Distribución de la fracción de genes rechazados por célula (no expresados). Se consideran células válidas a las que se encuentran por debajo de la cota.

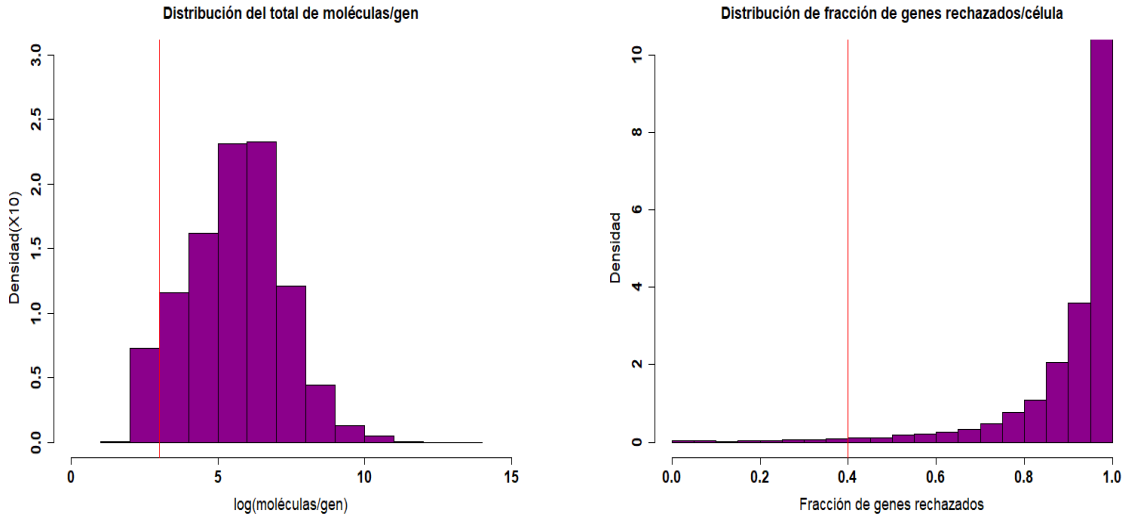
(b) Distribución del total de moléculas para cada célula muestreada. El eje x está en escala logarítmica. Las células válidas están delimitadas por ambas cotas.

FIGURA 4.3: Distribuciones estudiadas para la aplicación de criterio. Las líneas rojas, en ambos histogramas, indican los valores cotas propuestos (Hochgerner et al., 2018).

Con respecto a los genes, removemos a aquellos que son o bien poco o muy expresados. Definimos a un gen poco expresado como aquel que es detectado en menos de 20 moléculas en todas las células y un gen muy expresado es aquel que se expresa en más de una fracción del 0.6 células, que es lo mismo que tenga una fracción de expresiones nulas en el total de células menor o igual que 0.4.

Para identificarlos, graficamos la distribución del promedio de moléculas por célula para cada gen (panel A de Fig. 4.4) en escala logarítimica. En rojo se marca el valor cota mínima del criterio y se observa que dicho valor se encuentra lo suficientemente alejado del valor medio. Por otro lado, en el panel B graficamos la distribución para la fracción de genes rechazados (*dropouts*) y en rojo se indica el valor cota inferior tal que aquellas células con una fracción menor son

eliminadas. También podemos ver que dicho valor se encuentra lo suficientemene alejado del valor medio. De esta forma, eliminamos 1291 genes de forma que nos hemos reducido a una matriz de 13244 genes y 5447 células.



(a) Distribución de la cantidad de moléculas en que un gen es expresado en todas las células.

(b) Distribución de la fracción de genes rechazados por célula por ser expresados en dicha célula por menos de una molécula.

FIGURA 4.4: Distribuciones estudiadas para la aplicación de criterio de filtración de genes que son sobre-expresados o expresados insignificativamente. Las líneas rojas, en ambos histogramas, indican los valores cotas propuestos (Hochgerner et al., 2018) tal que se considera válido por encima de este.

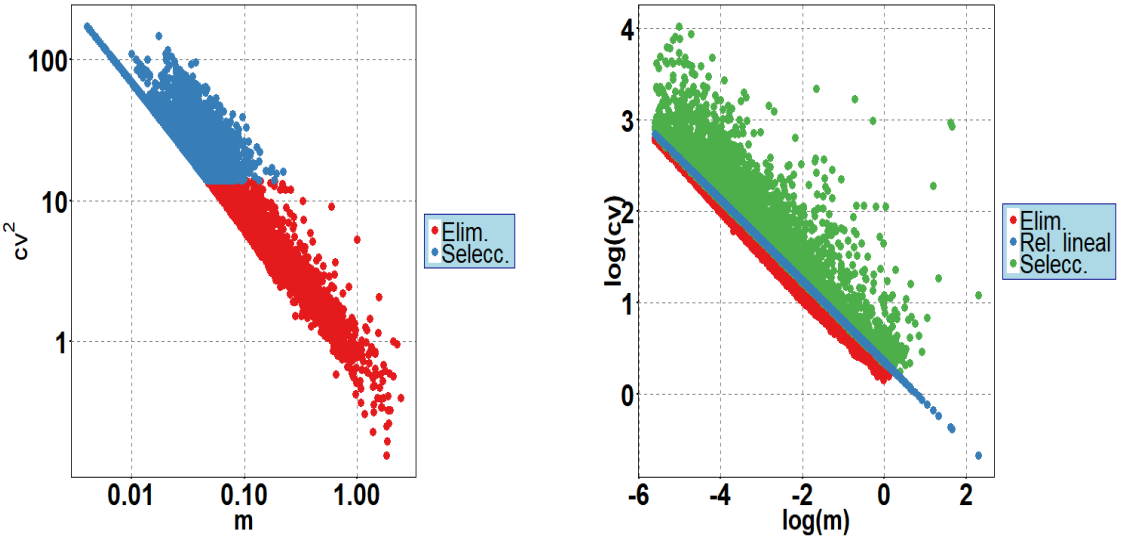
Por último, buscamos estudiar genes que sean informativos de los procesos biológicos que caracterizan a los tipos de células medidas en el experimento. Genes que no presentan una variabilidad alta entre la población de células no están relacionados directamente a algunos procesos en particular. Por el contrario, aquellos genes que sí presentan una variabilidad son característicos de algunos procesos siendo estos los genes que nos interesa estudiar. Por este motivo seleccionamos los 5000 genes que presentan la mayor variabilidad.

En particular estudiamos dos medidas de variabilidad basadas en los conceptos de coeficiente de variación (cv) y de valor medio (m) que se aplican al objeto de estudio en ec.4.1 para el gen i-ésimo donde N es la cantidad de células ($N = 5447$) y M es la cantidad de genes ($M = 13244$). Una forma de poder detectar a los genes con más variación es en un gráfico de cv^2 en función de m en escala logarítmica como se muestra en el panel A de Fig.4.5.

$$\begin{aligned} m &= \frac{\sum_{j=1}^{N} E_{ij}}{M} \\ cv &= \frac{\sigma(E_{ij})}{m} \end{aligned} \tag{4.1}$$

Se observa en la figura que seleccionamos a los 5000 genes (color azul) cuyos coeficientes de variación cuadrado (cv^2) son los más altos mientras que el resto es eliminado (color rojo). También se ve que este criterio de selección presenta un bias hacia genes de expresión baja. Buscamos seleccionar a aquellos genes con más variabilidad y que a su vez tengan valores medios no restringidos en un cierto rango, sino que se de una distribución extensa. Para corregir esto proponemos el siguiente criterio.

Graficamos $log(cv)$ en función de $log(m)$ en el panel B donde se realiza una regresión, un ajuste de un modelo lineal (Chambers et al., 1992) (color azul), y se calculan los residuos de éste ajuste. Aquellos con mayor residuo son los que más se alejan en distancia euclideana del ajuste lineal tal como se muestran a los 5000 con mayor residuo en color verde. Se puede ver que estos no sólo presentan una mayor variabilidad sino que también toman valores medios (m) en la totalidad del intervalo que presenta la población completa de genes. En consecuencia, son estos los 5000 genes con los que seguiremos trabajando de ahora en adelante.



(a) cv^2 en función de m del nivel de expresión para cada uno de los genes (escala logarítmica). Se seleccionan (azul) aquellos 5000 genes de mayor variabilidad y se elimina al resto (rojo).

(b) $log(cv)$ en función de $log(m)$. Regresión lineal (azul) tal que seleccionan los 5000 genes con mayor residuo (verde) mientras que el resto es descartado (rojo).

FIGURA 4.5: Criterios para la selección de genes de gran variabilidad basados en las nociones estadísticas de coeficiente de variación (cv) y valor medio (m).

Como ya presentamos, estos experimentos se caracterizan por una fracción baja de captura de expresión en los transcriptomas (*dropouts*) que dificultan el análisis. Vemos cómo cambia la distribución de la fracción de *dropouts* por célula al limitar el análisis a los 5000 genes más variables (Fig. 4.6). Se observa que para la población original de genes (rojo) el máximo se da para 0.95 mientras que para los 5000 más variables (de acuerdo al criterio adoptado), se reduce la densidad asociada al valor máximo y aumenta para valores de fracciones menores (color celeste).

Es importante notar que una expresión nula no implica biyectivamente que se trata de una limitación técnica (un *dropouts*) sino que puede ser que el gen no se expresa biológicamente en esa célula. Veremos más adelante cuáles son las problemáticas que trae una presencia mayoritaria de elementos nulos y cuáles son las propuestas para lidiar con éste problema.

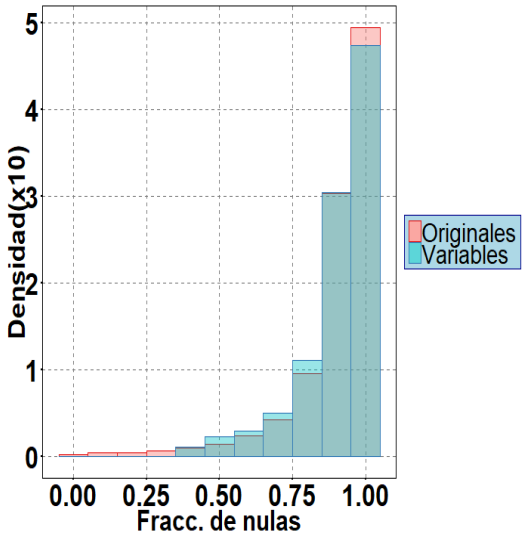


FIGURA 4.6: Distribución de la fracción de expresiones nulas en la población de genes para el caso de la población original (rojo) y de la resultante al limitarnos a los 5000 genes más variables según el criterio empleado para ello (celeste).

Posterior al análisis de calidad, la *matriz de expresión* se reduce a unas dimensiones de 5000 genes y 5447 células. Se eliminan diez marcadores que están relacionados con el sexo (se realizó el experimento con ratones hembras y machos) y con el estrés (Hochgerner et al., 2018) reduciéndonos a 4996 genes.

4.3. Criterios de selección de genes por noción de variabilidad

Como ya explicamos en Secc.2.1, la reducción de dimensionalidad para aproximar la VBR y la posterior inferencia de una trayectoria de evolución celular sobre esta variedad (ver esquema en Fig.2.2) depende fuertemente del proceso de selección de genes que denominamos de interés. Buscamos seleccionar genes que sean informativos sobre ciertos procesos biológicos como los genes cuyas evoluciones son esquematizadas en los paneles A y B de fig.2.3.

Estos presentarán variabilidad alta a lo largo de los experimentos pero, al mismo tiempo, están involucrados en determinados procesos biológicos que ocurren en el tiempo de desarrollo del tejido estudiado. Por esta razón la variabilidad que presentan estos genes debe ser compatible con la topología de la trayectoria celular inferida sobre la VBR.

En los Cap.2 y Cap.3 estudiamos la técnica de Slicer (Welch et al., 2016) para seleccionar estos genes informativos reconociéndolos como aquellos que presentan una varianza sobre las

vecindades de las células promedio menor que la global. Trabajamos sobre dos modelos de datos de experimentos scRNASeq y concluimos que la técnica de selección resulta ser efectiva en el sentido que permite construir una aproximación de la VBR sobre la cual inferimos una trayectoria de evolución dinámica. En esta sección analizaremos la efectividad de la técnica en el mismo sentido que en los capítulos anteriores sobre los datos experimentales ya estudiados previamente en el capítulo.

Para construir las vecindades de células, y disponer de una aproximación a la VBR, construimos un grafo donde cada nodo representa una célula. Tal como explicamos en Secc.2.2, a partir de la *matriz de expresión* construimos un grafo de k-primeros vecinos y armamos la matriz de adyacencia a partir de una matriz de distancia euclideana entre los nodos (muestras de célula única). En este caso en vez de que las células tengan un grado k_c tal que el grafo queda conectado, tomamos $k = 20$ (Hochgerner et al., 2018).

Utilizando las medidas de variabilidad global y local presentada en ec.2.3, nos concentramos en la cantidad $P = S^2 - \sigma^2$, una puntuación asociada a cada gen, para identificar a aquellos que presentan una reducción significativa en la estimación de la variabilidad local respecto a la global tal que $P < 0$. Graficamos la distribución de estas puntuaciones (Fig.4.7).

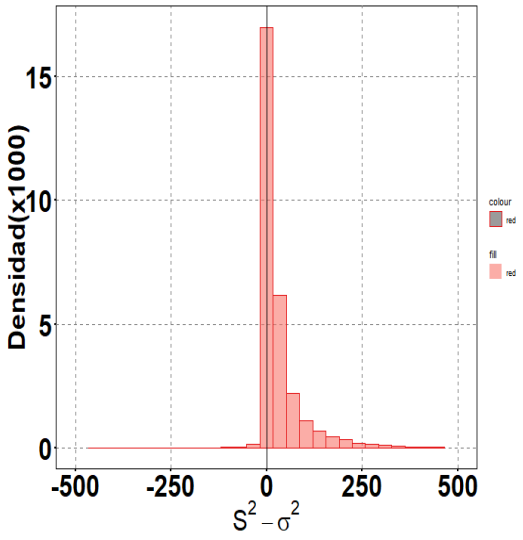


FIGURA 4.7: Puntuación $S^2 - \sigma^2$ en función de cada gen designado por un índice numérico que no expresa un orden particular. La línea horizontal roja señala el valor 0 donde $S^2 = \sigma^2$ de forma que el filtro selecciona a aquellos genes con un valor por debajo del cero.

Como se ve a partir de la figura, sólo una pequeña fracción de genes (4 %) presenta una variabilidad local menor a la global, sugiriendo que sólo 201 del total de genes profilados presentan campos suaves sobre el grafo celular. A continuación nos proponemos desarrollar otro criterio que relaje más la noción de suavidad para poder seleccionar una mayor fracción de genes y así no perder tanta cantidad de información.

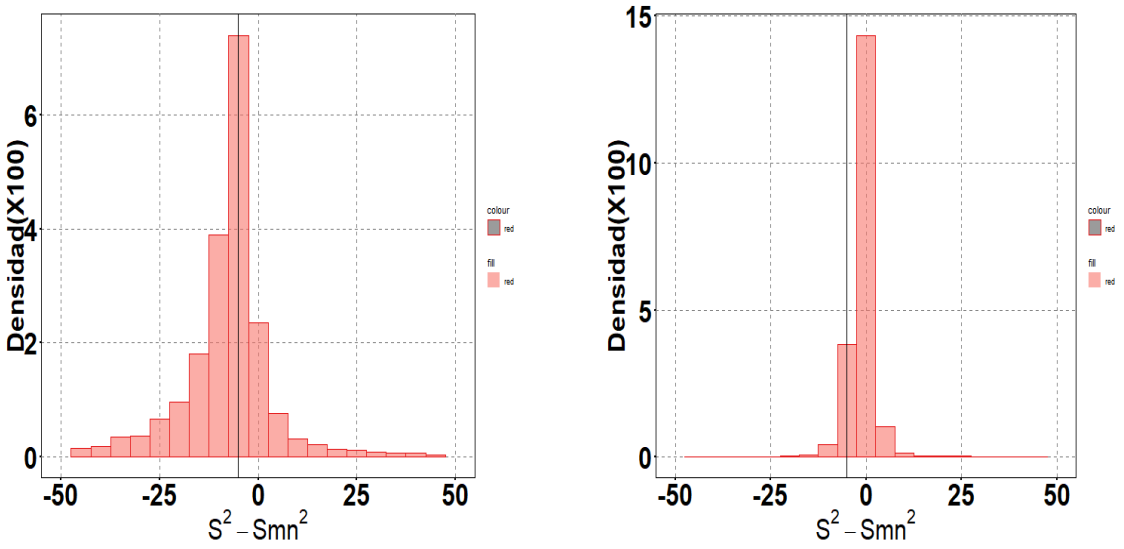
En vez de comparar a la varianza local con la global, desarrollamos un nuevo criterio que consiste en comparar esta medida S^2 con un modelo nulo. Este modelo implica calcular la varianza que llamamos Smn^2 que se espera encontrar cuando las vecindades de cada célula se definen de manera aleatoria. Smn^2 la comparamos con la que presenta el grafo, es decir la que se corresponde con el enlazamiento de nodos que presenta la red original. Si $S^2 < Smn^2$ decimos que se da una variabilidad local menor que la que se espera por azar.

Como hacemos una selección de células aleatorias para el cálculo Smn^2, realizamos 100 iteraciones con el objetivo de armar una distribución caracterizada por un valor medio de Smn^2 y una desviación estándar σmn. De esta manera, compararemos el S^2 topológico contra el valor medio de Smn^2 en unidades de σmn.

Como ya hemos corroborado (ver Fig.4.6) el experimento que estamos analizando presenta altísimos niveles de *dropouts*. Al estudiar la variabilidad global de genes, en su mayoría, se compara una expresión *dropouts* con la media y para el caso de la local, con la de sus vecinos que también tienen una probabilidad alta de ser *dropouts*. Esto oscurece la variabilidad real de los genes.

Una forma de disminuir estos efectos es considerar variabilidades locales S_0^2 o globales (σ_0^2) que sólo tengan en cuenta elementos no nulos de expresión. Así definimos criterios basados sobre las puntuaciones $S_0^2 - Smn_0^2$ y $S_0^2 - \sigma_0^2$ para implementar nuevos criterios de seleccion de genes.

Por otro lado, el criterio de Slicer para aceptar un gen pide que esta tenga una puntuación $P < 0$. Para el caso del modelo nulo nosotros consideramos una cota c tal que si $S^2 - Smn^2 < c$, el gen es seleccionado. Para determinar esta cota c, en Fig. 4.8 graficamos la distribución de la puntuación $S^2 - Smn^2$ en unidades de σmn para las varianzas definidas considerando *dropouts* en el panel A y sin hacerlo en el panel B. Tomamos el criterio de que las cotas c (para cada caso) se correspondan con el valor para el cual se centra el bin de la máxima densidad siendo este 5 para el caso en que se tienen en cuenta los elementos nulos y para cuando no, es 0 (en unidades de σ).



(a) Con elementos nulos

(b) Sin elementos nulos

FIGURA 4.8: Distribución de las puntuaciones $S^2 - S^2_{mn}$ medidas en unidades de σmn. La línea negra indica el valor de -5σ (valor usado en la literatura para reconocer un comportamiento lo suficientemente particular).

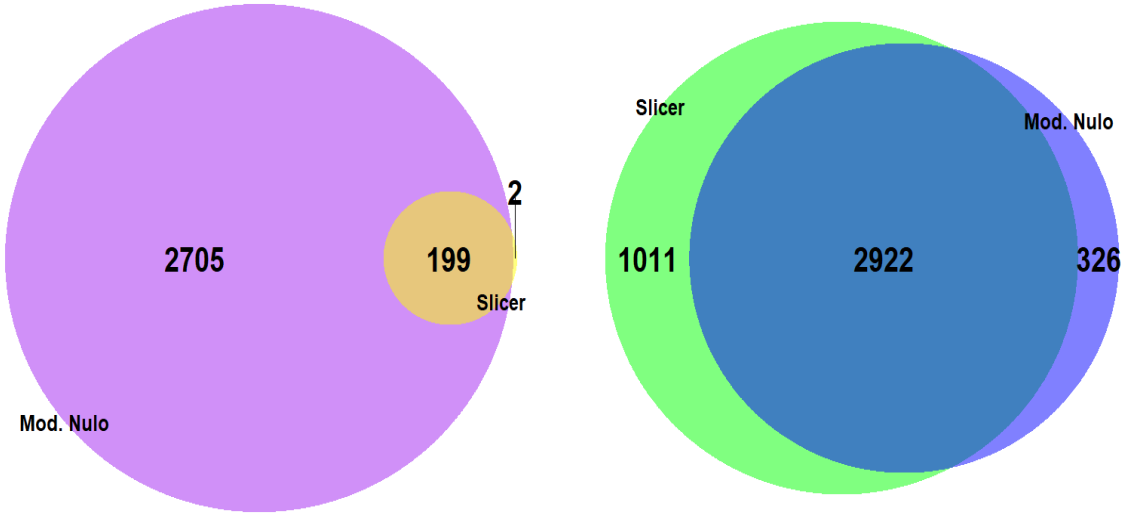
De esta manera, hemos definido cuatro criterios de selección de genes: comparando la variabiliad local con una global (criterio Slicer), y otro en el que la comparamos con una variabilidad obtenida en un modelo nulo (criterio del modelo nulo). En ambos casos se da la posibilidad o no de considerar los valores nulos de expresión. Esto último nos permite analizar si es posible mitigar los efectos de los eventos *dropouts* definiendo estos criterios.

Para analizar la concordancia de los diferentes criterios, se muestran diagramas de Venn (Fig.4.9 y 4.10). En la primera se muestran las respectivas intersecciones entre los conjuntos de genes seleccionados según el criterio de Slicer y del modelo nulo. En el panel A vemos que al considerar elementos nulos, los genes seleccionados por Slicer son muy pocos en comparación a los que selecciona el modelo nulo y están en su mayoría incluidos en éste. Esto indica que la selección por Slicer es un caso particular de la que se realiza por el modelo nulo. Por consiguiente, una variabilidad en vecidades menor que la global es un criterio de *suavidad* más fuerte que la que se corresponde con una variabilidad en vecindades menor que la de por azar.

En en panel B observamos que sin considerar los *dropouts*, la tendencia que vimos cambia porque se da una intersección y por ambos criterios quedan genes que no son compartidos por el otro criterio. En consecuencia no se puede afirmar que una selección sea un caso particular de la otra.

Por otro lado, comparamos para cada criterio, el caso en que se consideran expresiones nulas y para el que no (Fig.4.10). En el panel A se observa la diferencia en la cantidad de genes aceptados por Slicer en un caso y en el otro. En particular es notoria cómo se incrementa la cantidad al no considerar las expresiones nulas.

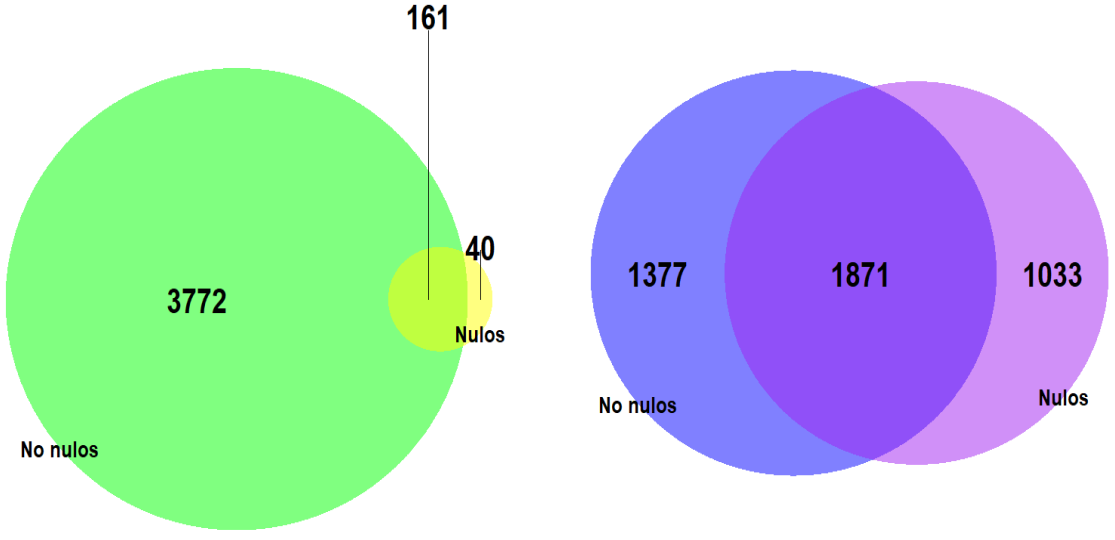
En el panel B vemos que para el modelo nulo se da un subconjunto en la intersección mientras que se dan otros dos particulares siendo el de los no nulos del 42 % y de los nulos 36 %. Esto indica que, al no considerar elementos nulos, la selección de genes por el criterio del modelo nulo se ve menos impactada que por Slicer.



(a) Considerando los elementos nulos.

(b) Sin considerar los elementos nulos.

FIGURA 4.9: Conjuntos de genes seleccionados por cada criterio y su respectiva intersección. A izquierda los criterios cuando se consideran los elementos nulos y a su derecha cuando no.



(a) Criterio de Slicer.

(b) Criterio del modelo nulo.

FIGURA 4.10: Conjuntos de genes seleccionados por el mismo criterio pero, comparando el caso en que se consideran los elementos nulos con el que no y su respectiva intersección. A izquierda el criterio de Slicer y a derecha el criterio el modelo nulo.

Para entender la razón de la diferencia en cómo se ve impactado cada criterio al no considerar *dropouts*, en Fig.4.11 realizamos graficos de caja de las distribuciones para el cociente entre el observable sin considerar los elementos nulos (S_0^2 o σ_0^2) asociado a los genes. Estos son designados genéricamente al no considerarlos como V_0^2 y al considerarlos, V^2.

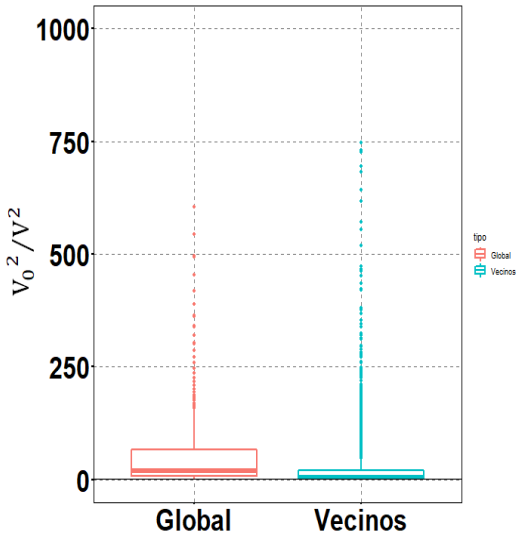


FIGURA 4.11: Gráfico de caja de las distribuciones sobre la población total de genes del cociente del observable estadístico sin considerar elementos nulos V_0^2 y el mismo observable considerándolos V^2. Se hace para la varianza global (σ^2) en rojo y para la de vecinos (S^2) en verde. Se indica con la línea negra horizontal el valor unidad tal que para valores por debajo es mayor la varianza al considerar ceros que al no hacerlo y para valores por encima, al revés.

Observamos que en ambos casos las medianas se dan por encima del 1 (indicado por la línea negra) siendo para la de vecinos (S^2) 5.9 y para la global (σ^2), 19.5. Esto indica que ambas medidas se ven incrementadas al no considerar los eventos *dropouts*. En particular, para σ^2 (color celeste) el tercer cuartil se encuentra muy por encima que para S^2 (color salmón). Por ello, concluimos que la varianza global se ve fuertemente aumentada al no considerar los elementos nulos en comparación con la de vecinos.

Por otro lado, como ya fue explicado, el criterio de Slicer compara dos varianzas distintas (la global y la de vecinos) que, como concluimos, al no considerar *dropouts* se ven aumentadas en distinta proporción. Esto puede llevar a sacar conclusiones erróneas porque puede ocurrir que para un gen $S^2 < \sigma^2$ no por una variabilidad local suave sino por el hecho de no haber considerado *dropouts*. Por el contrario, el criterio del modelo nulo al comparar S^2 con una varianza del mismo tipo obtenida por un modelo nulo no presenta este problema y es por tanto, más confiable.

Otro punto a tener en cuenta es que la existencia de altos valores de *dropouts* atenta contra cualquier metodología de reconstrucción de trayectorias y bases de atracción de los genes sobre la variedad de dimensión reducida, la VBR. Por eso consideramos adicionalmente en el proceso de filtrado a aquellos genes que son poco expresados. Estos los idenficamos como aquellos que

presenten un fracción de *dropouts*, a lo largo de las células muestreadas, mayor que la media observada siendo estos una fracción de 0.67. También tendremos en consideración la fracción sobre aquellos genes poco expresados que son seleccionados pues buscamos seleccionar la menor cantidad de estos.

Para determinar cuál de los criterios estudiados hasta el momento es más óptimo en relación a esta idea, se observan en Tabla.4.2 la fracción de genes aceptados, la fracción sobre estos que son poco expresados (tienen mayor *dropouts* que la media) y la fracción sobre la población total de genes poco expresados que son seleccionados. Se muestra para el criterio de Slicer y del modelo nulo indicando la fracción de la forma "considerando *dropouts*/sin hacerlo".

Vemos que al no considerar expresiones nulas la fracción de los genes que selecciona Slicer que son poco expresados disminuye al igual que la selección por el criterio del Modelo nulo. Sin embargo, por este último criterio se selecciona una fracción menor en ambos casos que es a su vez igual y menor que la media observada en la población total.

Por otro lado, la fracción sobre los genes poco expresados que son seleccionados aumenta al no considerar los *dropouts* para ambos criterios. La diferencia es que por Slicer se acepta una cantidad muy por encima que la mitad mientras que por el Modelo nulo la fracción es levemente mayor que 0.5.

En consecuencia, afirmamos que ambos criterios del Modelo nulo resultan ser más óptimos con respecto a la filtración de genes que son poco expresados con respecto a la fracción *dropouts* media.

Criterio	**Slicer**	**Modelo nulo**
Fracc. aceptados	0.04/0.79	0.58/0.65
Fracc. de aceptados poco expresados	0.82/0.73	0.67/0.61
Fracc. de poco expresados aceptados	0.05/0.85	0.39/0.59

TABLA 4.2: Relación entre genes que son aceptados y los que son poco expresados (por debajo de la media) para ambos criterios de comparación de medidas de varianza. Se señala en cada caso el valor considerando elementos nulos/sin elementos nulos.

Una vez que ya hemos estudiado las dependencias de estos criterios con los *dropouts*, es de interés estudiar algunas propiedades de estas medidas de varianza para casos particulares. De especial interés será hacerlo considerando/buscando genes marcadores (concepto explicado en Secc.4.3). Estos genes permiten identificar etapas, o mejor dicho, subconjuntos diferenciales de muestras, asociadas, por ejemplo, a diferentes estadíos celulares.

4.4. Marcadores: comparación de criterios de selección

Un grupo especial de genes que resultan de interés identificar y caracterizar en este tipo de estudios incluye a genes que actuan como marcadores de determinados estadíos celulares. Para estudiar cómo las diferentes consideraciones sobre variabilidad y *dropouts* afectan este tipo de análisis estudiamos las variabilidades en las expresiones de dos genes marcadores que son relevantes para el proceso de neurogénesis.

Estos genes se encuentran involucrados en procesos biológicos distintos y según la bibliografía uno sucede previo a otro en el tiempo de desarrollo del ratón en que se han realizado las muestras. Estos son el gen Ascl1, marcador del estadío celular llamado radial glia like, y Igfbpl1, marcador de neuroblastos 2 (Hochgerner et al., 2018). También tienen como diferencia la fracción de elementos nulos siendo Ascl1 muy poco expresado (98 % de elementos nulos) mientras que el gen Igfbpl1 bastante expresado (78 % de nulos) en comparación con la fracción media observada.

Con el fin de visualizar cómo sería la evolución dinámica en la expresión de estos genes se busca darle un orden temporal a las células en un eje *pseudo-temporal* a partir de una célula que representa el estadío inicial de la pseudo-dinámica que estamos estudiando. Seleccionamos, en consecuencia, una célula tipo nIPC de la muestra obtenida en el primer punto temporal del experimento. Esto se justifica ya que dicho tipo de células aparecen en los estadíos iniciales del camino de desarrollo neuronal (Hochgerner et al., 2018).

A partir de dicha muestra inicial las distancias geodésicas sobre el grafo nos proveen de una medida que asociamos al *pseudo-tiempo* sobre la trayectoria reconstruida. Una forma de calcular distancias es por el camino más corto que une dos nodos, es decir, la cantidad mínima de nodos intermedios. Para el caso de un grafo pesado se define la misma noción de distancia pero, teniendo en cuenta el peso de los enlaces que unen a los nodos del camino más corto por el algoritmo de Dijkstra (West et al., 1996).

Nos proponemos comparar ambas medidas de distancias con el fin de elegir cuál emplear y reconstruir un orden *pseudo-temporal* dado por el orden celular. Para ello, procedemos a pesar los enlaces según una medida de similitud que es la medida de Jaccard (Jaccard, 1901). Esta, para dos nodos (i, j), se calcula cuantificando la proporción de vecinos comunes:

$$J = \frac{\sum_{k=1}^{M} A_{ik} A_{kj}}{\sum_{k=1}^{M} A_{ik} A_{ik} + \sum_{k=1}^{M} A_{jk} A_{jk}} \tag{4.2}$$

donde A_{ij} es el elemento de la matriz de adyacencia.

A partir de la muestra inicial que designamos calculamos ambas nociones de distancia sobre el grafo. En un caso sin pesar los enlaces y en otro pesando por Jaccard. Graficamos sus distribuciones en Fig.4.12.

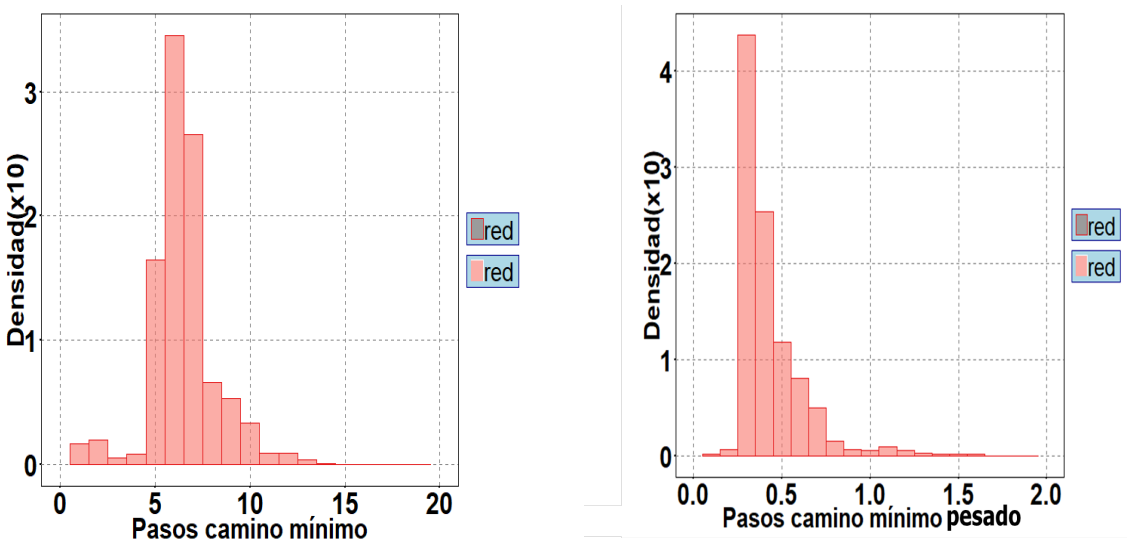
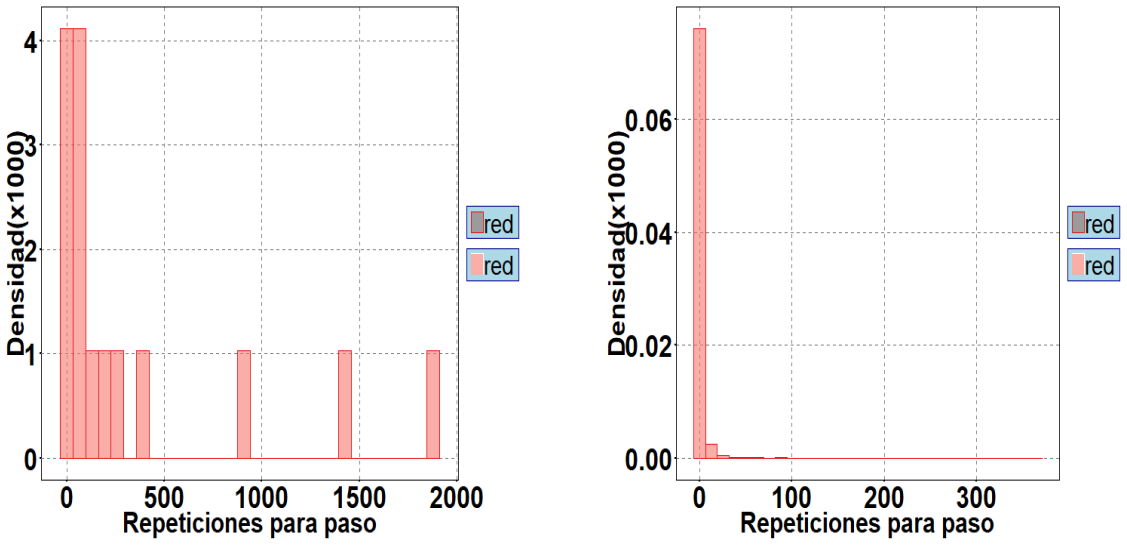


FIGURA 4.12: Distribuciones de medidas de distancia para la población de células para con una célula inicial. A izquierda la medida de cantidad de pasos (nodos) mínima y a derecha la misma medida pero para los enlaces del grafo pesados por Jaccard.

Observamos que mientras que para el grafo no pesado (panel A) se da una densidad significativa para pasos más chicos que el que se corresponde con el máximo, para el grafo pesado (panel B) esto no ocurre. Sin embargo, esto no parecería ser una diferencia suficiente para determinar cuál criterio utilizar.

Por otro lado, a una misma cantidad de pasos del camino mínimo de la muestra inicial se encuentran varios nodos dificultando el armado un orden dinámico de las células. Por eso, buscamos emplear una medida que presente una menor cantidad de nodos a la misma distancia del nodo inicial, es decir que se de una menor cantidad de repeticiones por medida de distancia.

En Fig.4.13 graficamos la distribución de repeticiones que se dan para los pasos del camino mínimo para el caso del grafo sin pesar (panel A) y pesado (panel B). Observamos que para la medida sobre el grafo pesado se da una concentración de la densidad de probabilidad para valores bajos de repeticiones en comparación con los pasos sobre el grafo sin pesar. Por esta razón decidimos optar por el criterio del camino mínimo pesado y los subconjuntos de células que muestran el mismo valor de distancia son ordenadas según la semana post-natal en que fue realizada la medición.

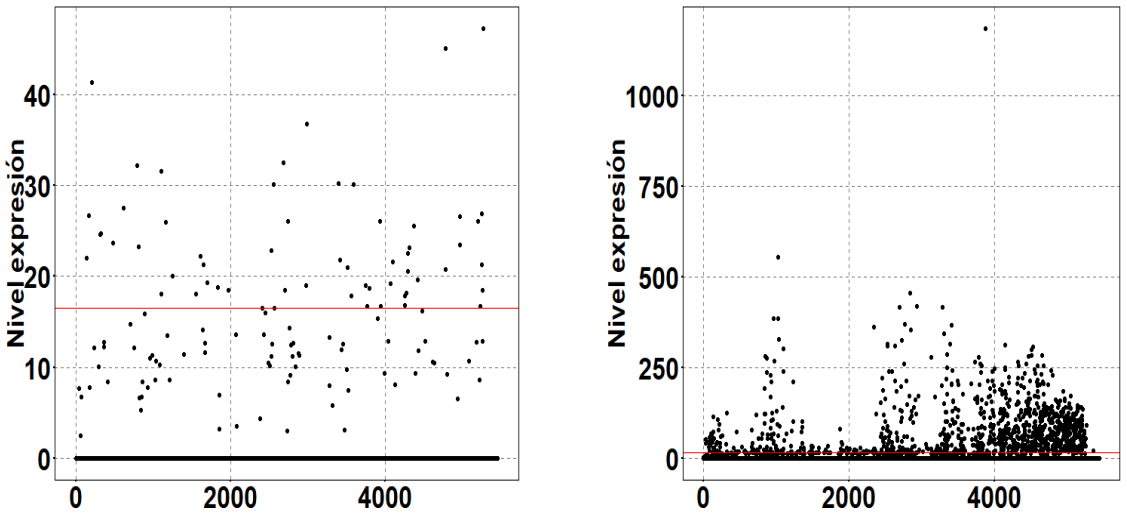


(a) Camino mínimo

(b) Camino mínimo pesado

FIGURA 4.13: Distribuciones de las cantidad de repeticiones para un mismo valor de distancia según la respectiva medida.

Una vez que hemos encontrado un pseudo-orden en el grafo de células a partir de una muestra inicial, graficamos la trayectoria de la expresión de estos genes marcadores (Fig. 4.14) en función del *pseudo-tiempo* reconstruido. Se pueden observar dos comportamientos muy diferentes, en el panel A vemos que los niveles de expresión relevados del gen Ascl1 son mucho más chicos que para el gen Igfpl1 que se ilustra en el panel B. Finalmente a lo largo de las trayectorias se observa para el gen Ascl1 un comportamiento errático mientras que el otro gen presenta una cierta correlación *pseudo-temporal* asociada a una dinámica en su expresión que se localiza en ciertas regiones más o menos bien definidas.



(a) Gen Ascl1 marcador de radial glia like.

(b) Gen Igfbpl1 marcador de neuroblastos 2.

FIGURA 4.14: Evolución de los niveles de expresión para cada gen marcador según el orden en pseudo tiempo encontrado. La línea roja horizontal representa el valor medio. Se muestra en ambos casos la cantidad de valores que toman cero.

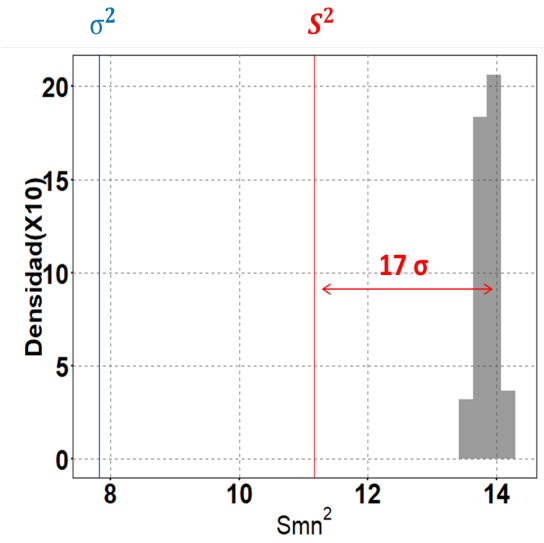
Afirmamos que se trata de dos genes que presentan un comportamiento en sus expresiones muy diferente. Esto nos permite estudiar cómo funciona nuestro análisis de selección de genes para casos muy distintos. Para ello, a continuación, analizamos cómo es la variabilidad de la expresión de estos dos genes con respecto a la noción que presentan los cuatro criterios con los que vinimos trabajando.

Graficamos la distribución de la varianza local asociada al modelo nulo para cada uno de los genes (Fig.4.15). Por un lado dándo la posibilidad de considerar elementos nulos (parte superior) y por otro, sin hacerlo (parte inferior). Se señala en azul el valor de σ^2 y en rojo, el S^2 asociada a la topología. También en Tabla. 4.3 se muestran los puntuaciones según cada criterio.

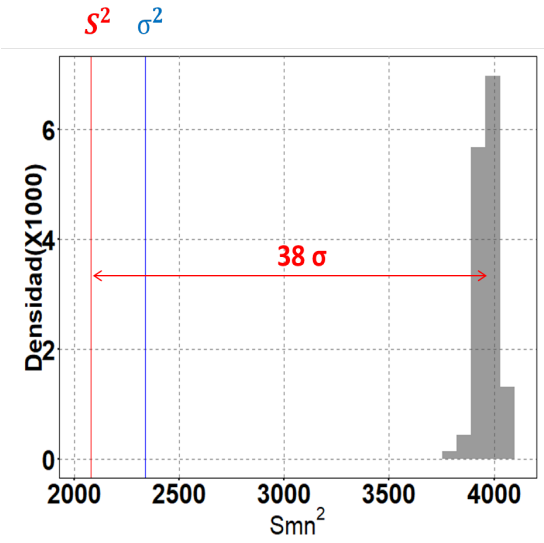
Observamos en el panel A que para el gen Ascl1 aunque $S^2 > \sigma^2$, es decir que no presenta una variación local suave del nivel de expresión en comparación con la global, S^2 es significativamente menor al compararla con la varianza del modelo nulo. Esto indica que la variabilidad local es suave sólo en comparación a lo que se esperaría por azar.

Al no considerar los elementos nulos, vemos en el panel C que el valor de σ^2 supera significativamente al modelo nulo y al S^2 topológico. Indica que el gen no sólo tiene una variabilidad local más suave de lo que se espera por azar sino que también en comparación con la población global de células.

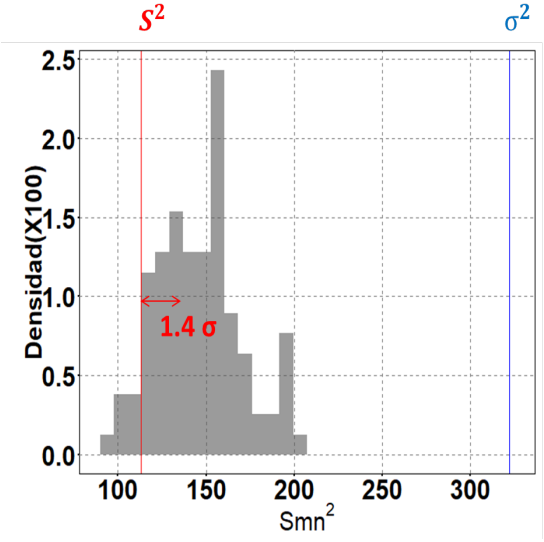
Esta es una cualidad que se hace visible al no considerar los elementos nulos, que para el caso de este gen Ascl1 representa a la mayoría de sus expresiones (98 %). Se plantea el interrogante de si este fenómeno también ocurrirá para el caso del gen que tiene mucho menos nivel de *dropouts* (78 %).



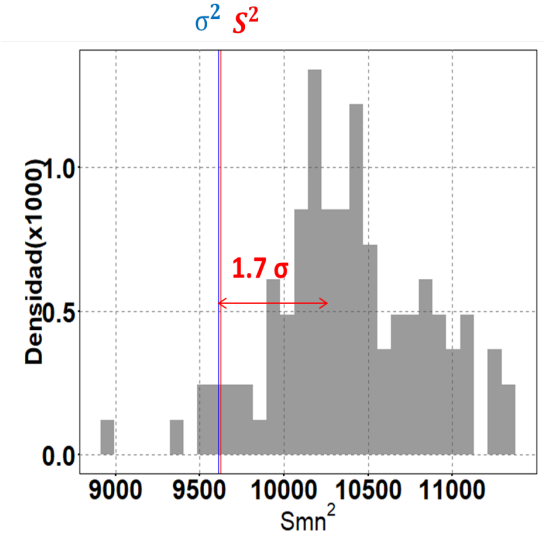
(a) Gen Ascl1 marcador de radial glia like.



(b) Gen Igfbpl1 marcador de neuroblastos 2.



(c) Gen Ascl1 marcador de radial glia like.



(d) Gen Igfbpl1 marcador de neuroblastos 2.

FIGURA 4.15: Distribuciones del modelo nulo para cada gen (a derecha e izquierda) considerando elementos nulos (superior) y sin hacerlo (inferior). Se marca el líneas verticales distintas medidas. En azul σ^2 y en rojo S^2. Las distribuciones se corresponden con 100 iteraciones de asignación de vecinos de forma aleatoria preservando la distribución de grado del grafo.

En el panel B vemos que al dar la posibilidad de dropout, el gen que tiene una menor fracción de estos (Igfbpl1) la variabilidad local cumple $S^2 - \sigma^2 < 0$ cuyo valor se observa en Tabla.4.3 y además es lo suficientemente menor que la distribución del modelo nulo de forma que es un gen que experimenta una transición de su nivel de expresión suave entre vecinos en comparación con lo que se espera por azar y en comparación también a la población total de células.

Al no considerar los elementos nulos (en el panel D), $S^2 - \sigma^2 > 0$ con el valor que se indica en Tabla.4.3 por lo que no presenta una variabilidad suave en comparación con la población global. Sin embargo, continúa mostrando un comportamiento de transición más suave entre vecinos de lo que se espera por azar.

Nombre gen	**Ascl1**	**Igfbpl1**
Proceso	Radial glia like	Neuroblastos 2
% Elementos nulos	98	78
$S^2 - \sigma^2$	3.33/-208.8	-258.2/13.9
$(S^2 - Smn^2)(\sigma)$	-17.13/-1.43	-38.2/-1.7

TABLA 4.3: Valores de puntuación según cada criterio para los genes marcadores estudiados. Se indica para cada uno el valor considerando valores nulos/sin considerarlos.

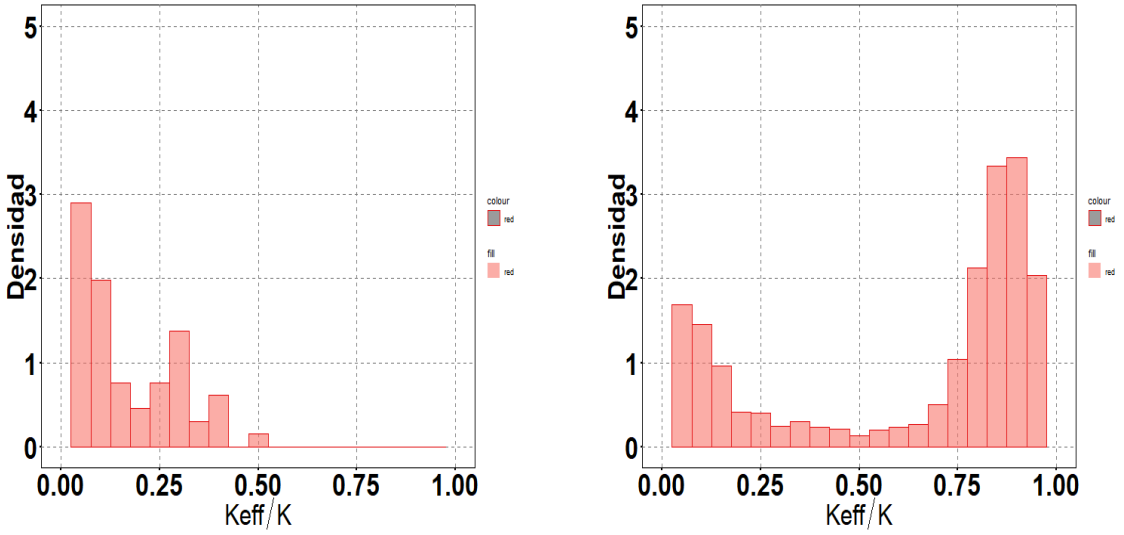
Podemos concluir que el criterio de Slicer no resulta confiable al tener en consideración la problemática de los elementos nulos. Para el caso de un gen que tiene una fracción alta de estos (por encima que la media), Ascl11, el no considerar *dropouts* para el cálculo de las varianzas permite visibilizar una transición local más suave que la global y para el caso de un gen con un bajo nivel de *dropouts*, Igfbpl1, se deja de poder visibilizar este tipo de transición.

Por el contrario, aunque cambia la puntuación asociada por el criterio del modelo nulo, no cambia la tendencia que este criterio permite observar para el caso de ambas fracciones de dropout estudiadas. Se pude ver una variabilidad en las vecindades menor que la se espera por enlazar aleatoriamente los nodos del grafo que aproxima a la VBR.

Por otro lado, al no considerar expresiones nulas para el cálculo de las medidas de varianza, es de interés preguntarse en qué proporción se ve reducido el espacio de muestreo contra la que se compara la expresión media del gen ($< g >$) en el caso de σ^2 o la expresión del gen en las células vecinas para S^2. Queremos comparar también cómo es esta reducción del espacio para el caso de un gen con nivel alto de *dropouts* y para otro de nivel bajo.

Para ello, considerando sólo aquellas células (nodos del grafo que aproxima la VBR) en las cuales los genes se expresan calculamos el cociente del grado efectivo (Keff) que considera sólo como vecinas a células donde el gen también se expresa y el grado K. Graficamos las distribuciones

de las densidades asociadas a este cociente (Fig. 4.16) donde vemos en el panel A que el espacio de muestreo del gen expresado en un 2 % (Ascl1) es ampliamente reducido concentrando la densidad por debajo del 0.50. Por el contrario, el gen expresado en un 22 % (Igfbpl1) presenta una densidad de probabilidad distribuida en todo el rango siendo esta más alta para valores altos de Keff/K.



(a) Gen Ascl1 (b) Gen Igfbpl1

FIGURA 4.16: Distribuciones de la fracción de grado efectivo al no considerar células donde el gen no se expresa.

Es interesante notar éste fenómeno en la cantidad de términos sobre lo que se promedia para calcular una varianza local promedio sobre todas las células y su relación con el porcentaje de células en que es el gen expresado (Tabla. 4.4). Observamos que la cantidad de términos total en las sumatoria para el gen que es muy poco expresado (Ascl1) se ve reducida a una fracción de 0.17 mientras que para el gen que es más expresado (Igfbpl1) se reduce a una de 14.45. Esto refuerza la idea de que estamos calculando un observable estadístico, S^2, muestreando sobre una población muy reducida y esto se ve ampliamente exacercabo para el caso de mucho *dropouts*.

Nombre gen	**Ascl1**	**Igfbpl1**
% células expresado	2	22
% términos	0.17	14.45

TABLA 4.4: Porcentaje de células de la población total en la que cada gen es respectivamente expresado y el porcentaje de términos totales de la cuenta de S^2 al no considerar elementos nulos respectivo al que tendría al considerar la totalidad de la población para el cálculo.

Estudiamos distintos criterios de selección de poblaciones de genes que muestran transiciones suaves de su expresión a lo largo de las muestras de células. Por el criterio de Slicer sólo se

selecciona una fracción de 0.04 de genes. Estos nos llevó a definir los criterios de Slicer y del modelo nulo sin considerar los eventos dropout para cuantificar las variabilidades. Observamos que σ^2 se ve fuertemente aumentada al no considerar los *dropouts* en comparación con S^2. En particular vimos un ejemplo de un gen con poco *dropouts* y otro con mucho. De este modo afirmamos que el criterio de Slicer sin *dropouts* no resulta confiable.

Con respecto a la alta fracción de *dropouts*, decidimos seleccionar genes que se expresan en la mayoría de las células. El criterio del modelo nulo (con y sin *dropouts*) es el más óptimo en este sentido. Sin embargo, vimos que el espacio de muestreo, K_{eff}/K, se ve muy reducido para el gen en que predominan los *dropouts*. En consecuencia afirmamos que el criterio que parecería ser el más confiable es el del modelo nulo considerando *dropouts*.

Hasta ahora en el capítulo propusimos distintas metodologías de selección de genes que sirven para reducir la alta dimensionalidad que caracterizan a los experimentos scRNASeq. Nos concentramos en un primer filtrado en los 5000 genes más variables y con ellos construimos un grafo KNN de k-primeros vecinos para aproximar la VBR. Luego, aplicamos diferentes criterios basados en el análisis de variabilidad local sobre la VBR para identificar genes de interés sobre los que focalizaremos futuros análisis.

En la siguiente subsección buscamos entender algunas de las propiedades que presentan los genes que identificamos por medio de los criterios. Por eso, resulta interesante pensar qué relación podríamos plantear con otros métodos de reducción de dimensionalidad. En particular, un método ampliamente utilizado que en particular es con el que trabajaron en el proyecto de Linnarson (Hochgerner et al., 2018) (ver Cap.4) es el análsis de componentes principales (PCA, del inglés principal component analysis) (K., 1901). Presentaremos la técnica y por medio de esta buscaremos cuantificar diferencias entre los genes que fueron seleccionados por los distintos criterios que estudiamos a lo largo del capítulo.

4.5. Proyección en PCA y su relación con la selección de genes

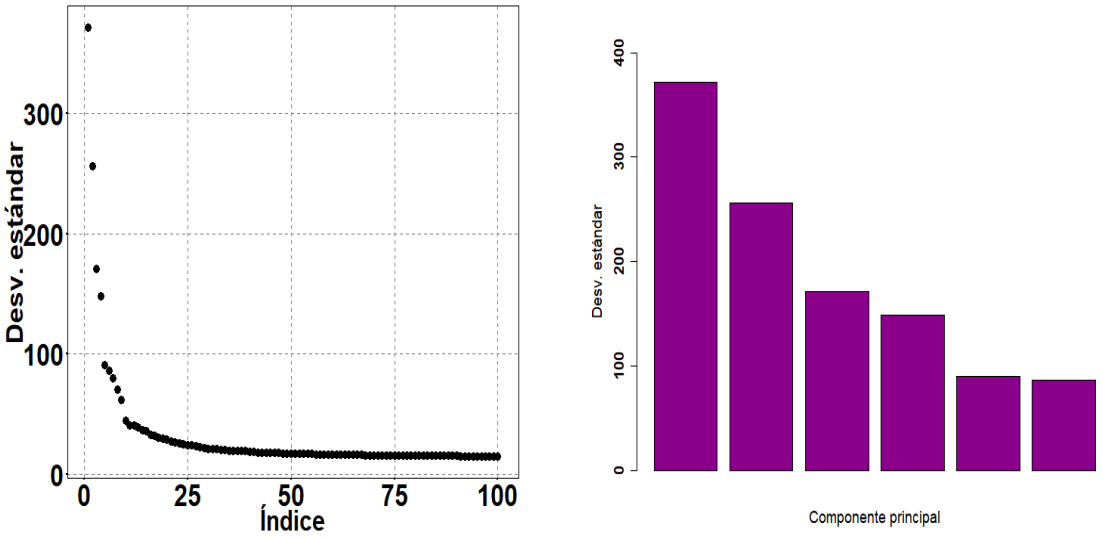
PCA simplifica la complejidad de la alta dimensionalidad conservando tendencias y patrones proyectando los datos en menos dimensiones que actúan como una síntesis de las características que posee la misma. Es un método de aprendizaje no supervisado pues encuentra patrones sin un conocimiento previo.

Se proyectan los datos en el espacio de las componentes principales (PCs, del inglés principal components) siendo la primera de ellas (PC_1) elegida para minimizar la distancia total entre los datos y su proyección en PCs y también maximizar la varianza (σ^2) de los puntos proyectados. Las siguientes PCs son elegidas de la misma forma explicando sucesivamente una σ^2 menor. Se tiene el requerimiento adicional de estar descorrelacionadas con las previas PCs tal que estas son geométricamente ortogonales. Esta condición significa, para el caso de la *matriz de expresión*

de un experimento scRNASeq que el número máximo de PCs posibles es o bien el número de muestras (M) o de genes (N).

En particular, nuestro objetivo es reducir la cantidad de variables de las observaciones, la cantidad de genes o dimensiones. Entonces, la matriz a la que le practicamos PCA es la transpuesta de la *matriz de expresión*, que tiene como columnas los genes y como filas, las observaciones de célula única. Proyectamos sobre los genes de forma que obtenemos 100 dimensiones o componentes principales siguiendo el trabajo (Hochgerner et al., 2018) de 5447 elementos.

Para entender cómo son las componentes principales de la *matriz de expresión*, graficamos la desviación estándar que estas explican en Fig.4.17. En el panel A graficamos para las 100 PCs mientras que en el panel B para las 6 primeras en forma de gráfico de barras. Observamos que el valor más alto se da para la primera PC tal como indica la teoría y éste valor cae rápidamente tal que a partir de una cierta cantidad de PCs, el σ que explican permanece más o menos constante. Esto indica que la elección de las primeras 100 PCs (Hochgerner et al., 2018) resulta arbitraria.



(a) 100 primeras PCs

(b) 6 primeras PCs

FIGURA 4.17: Desviación estándar (σ) asociada a cada una de las componentes principales.

Un criterio para determinar la cantidad de PCs es elegir la menor cantidad de PCs que explican la mayor σ acumulativa. Para ello, graficamos esta medida sin normalizar (color rojo) en función del índice de la PC en Fig.4.18 y realizamos un ajuste lineal (color azul). Observamos que a partir de las doce primeras PCs, el σ acumulado tiene un comportamiento lineal. Esto indica que la adición de PCs no aumenta el σ que estas pueden explicar. En consecuencia, seleccionamos a estas PCs.

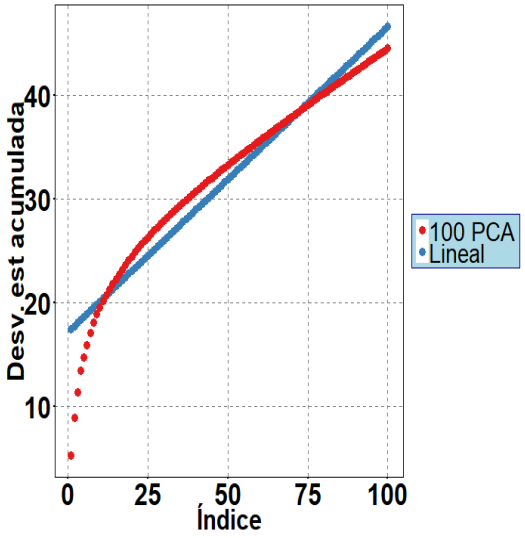


FIGURA 4.18: Gráfico de desviación estándar acumulada sin normalizar en función de la cantidad de PCs sobre la que se calculó la medida (rojo). Están ordenadas según el orden que da el algoritmo explicado previamente. En azul se grafica un ajuste lineal.

Por otro lado, el algoritmo de selección de genes de Slicer (Welch et al., 2016) le otorga una puntuación a cada gen según $P_j = S_j^2 - \sigma_j^2$ y la técnica introducida por nosotros que compara la varianza entre vecinos de la topología con la de un modelo nulo también le otorga una puntuación a cada gen $W_j = S_j^2 - Smn_j^2$ (en unidades de σ). Mientras que desde ésta perspectiva se otorgan puntuaciones a todos los genes y se toma algún valor como cota para seleccionar genes, proyectando en PCA se selecciona una cantidad de PCs. Para poder comparar ambas técnicas lo que hacemos es otorgar una puntuación a cada gen a partir de la proyección en PCAs para así comparar las puntuaciones.

La proyección en componentes principales (PCA) de las dimensiones de las variables genes se formaliza en ec.4.3 donde PCA_i es la componente principal i-ésima, g_j es el vector gen j-ésimo (la expresión de éste en las muestras), a_{ij} es un coeficiente positivo normalizado y se suma hasta M, la cantidad de genes. Se puede ver que cada PCA es una combinación lineal de los autovectores que constituyen la base que describe al subespacio de las expresiones de los genes.

$$PCA_i = \sum_{j=1}^{M} a_{ij} g_j \tag{4.3}$$

Se trata de una suma pesada por los a_{ij} que cuantifican qué tanto participa el gen j-ésimo en la proyección en la PC i-ésima. Esto nos permite definir una puntuación P_j en ec.4.4 que representa la participación del gen en la totalidad de las K PCs.

$$P_j = \sum_{i=1}^{K} |a_{ij}| \tag{4.4}$$

Graficamos la distribución de la puntuación sobre la población de genes (Fig. 4.19). Se observa que es un distribución con su sección inter-cuartil comprendida en valores bajos. El valor medio es de 0.04 mientras que el máximo es de 2.31 de forma que las puntuaciones altas son los que están como outliers de la distribución. Esto indica que la proyección en PCAs le otoga mayor participación a un grupo específico y reducido de genes.

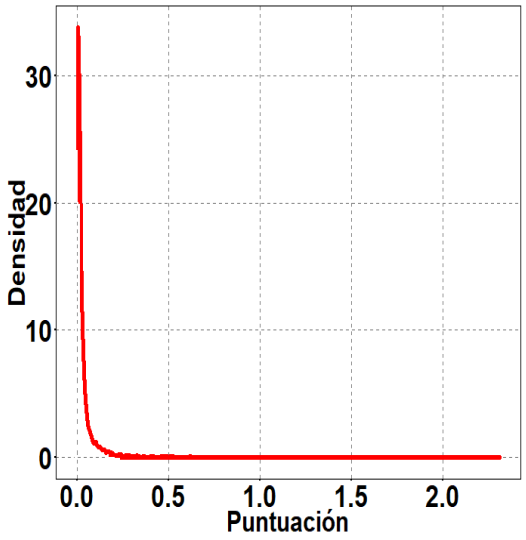


FIGURA 4.19: Distribución de la puntuación P que se obtiene de la proyección del espacio de genes en las 12 primeras PCAs para cada gen de la población de 5000 genes más variables. Un gen con una puntuación alta implica una mayor participación del mismo en la proyección de dichas componentes. Se trata de una distribución caracterizada por sus outliers.

Ahora que interpretamos a la proyección en PCA como la otorgación de un puntuación P a los genes, podemos comparar esta técnica de reducción de dimensionalidad con la de los criterios de selección de genes según una variabilidad local suave. Para ello, comparamos la distribución de la puntuación P sobre las poblaciones de los 200 genes que fueron seleccionados con la mejor puntuación por cada criterio y también sobre la población total de genes en gráficos de caja (Fig. 4.20). Se representa a cada distribución con un color y especificando el criterio donde D (por *dropouts*) referencia que se consideran expresiones nulas mientras que ND es que no lo hace.

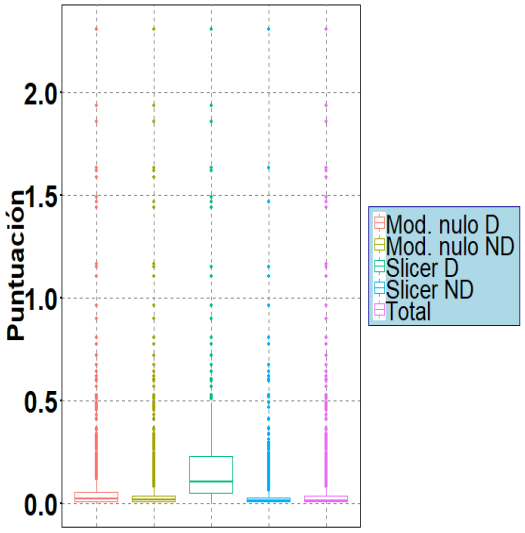


FIGURA 4.20: Distribuciones de la puntuación P para los 200 genes aceptados que obtuvieron una mejor puntuación según cada criterio (mayores en módulo y del mismo signo que la cota). Son comparadas con la distribución de la población total. D referencia que se considera elementos nulos en las expresiones mientras que ND es que no lo hacen.

Se puede ver que el criterio que tiene asociados mayor valores de la puntuación en el espacio inter-cuartil es el criterio de Slicer considerando los *dropouts* tal que por completo dicho espacio se encuentra por encima del tercer cuartil correspondiente a los demás criterios y también respecto a la población total de genes.

El siguiente criterio que es más compatible con PCA es el criterio del modelo nulo considerando *dropouts* ya que tiene un mayor P en el espacio inter-cuartil que el que presenta la población total. Se condice con nuestro análisis previo de que es un criterio exitoso en varios sentidos. Por último, los criterios tanto de Slicer como del modelo nulo sin considerar las expresiones nulas no muestran distribuciones significativamente distintas a la de la población total.

Entonces, en este capítulo caracterizamos e hicimos un análisis de calidad en un data-set experimental de células del giro dentado del hipocampo de ratones. Aplicamos el criterio de selección de genes por la varianza entre vecinos y observamos que selecciona sólo 201 genes que es una fracción de 0.04. Esto nos llevó a desarrollar un criterio que comparar la varianza entre vecinos con un modelo nulo con el objetivo de relajar la condición de suavidad y seleccionar una fracción mayor.

También observamos que una fracción de 0.90 de la *matriz de expresión* son elementos nulos de forma que esto dificulta una apropiada selección de genes. Por ello, construimos los mismos criterios (Slicer y modelo nulo) pero, sin considerar expresiones nulas para ver si así mitigabamos los efectos.

Estudiamos las limitaciones de cada uno de los criterio y vimos en particular cómo funcionan para dos genes marcadores. También comparamos la técnica de reducción de dimensionalidad por selección de genes con la reducción por componentes principales (Hochgerner et al., 2018).

Considerando varios aspectos, concluimos que el criterio que pareciera ser el óptimo es el del modelo nulo considerando los *dropouts*. Sin embargo, hay varios genes que este criterio selecciona que no son seleccionados al aplicar el mismo criterio pero, sin considerar los *dropouts* y al revés también. Esto nos hace pensar de que tal vez, al aplicar este criterio, estemos seleccionado genes que no deberíamos y no seleccionando otros que sí presentan una variabilidad suave en las vecindades.

La razón de esto son los *dropouts* que no nos permiten sacar conclusiones robustas pues difuminan las variabilidades reales en las expresiones. En consecuencia, concluimos que debemos imputar los valores nulos de la matriz para poder realizar una reducción apropiada. En el siguiente capítulo estudiamos una técnica de imputación de valores de expresión que fue desarrollada para ser utilizada en este tipo de experimentos.

Capítulo 5

Técnica de imputación de dropouts

Hasta ahora, a lo largo de la tesis, estudiamos el criterio de selección de genes según la variabilidad en vecindades de células representadas en el grafo que resulta como una aproximación para la VBR. Esta variabilidad es cuantificada por la varianza entre vecinos y comparada con la variazan global. Lo hicimos para el caso de un modelo simple de evolución de genes y para un modelo más robusto que modela genes expresados. En ambos casos encontramos resultados exitosos.

Sin embargo, al trabajar con una matriz de resultados experimentales de células del hipocampo de ratones del experimento en Linnarson Lab (Hochgerner et al., 2018) pudimos ver las complicaciones que traen los *dropouts* que se originan en la ineficiente captura del RNAm que se ve representada en una alta tasa de expresiones nulas.

El problema de los *dropouts* es inherente a esta tecnología, y se presenta típicamente en tasas que rondan entre el 40 % y 90 % de las expresiones. Eventos que involucren niveles de expresión muy bajos serán posiblemente reportados como ceros en las matrices de expresión scRNAseq. En particular los datos experimentales que analizamos en el Cap.4 presentan un tasa de un 90 %.

Altos valores de *dropouts* atentan contra cualquier metodología de reconstrucción de trayectorias sobre la VBR. Esto ha llevado a que en los últimos años se desarrollaran diversos algoritmos de imputación que son específicos para este tipo de experimentos. En general estos presentan el mismo principio de funcionamiento: infieren expresiones que fueron reportadas como nulas a partir de la información restante recopilada durante el experimento.

En este capítulo estudiamos un método de imputación en particular que se llama Scimpute (Li and Jessica Li, 2018). Lo aplicamos sobre el modelo robusto de la *matriz de expresión* que construimos y explicamos en Cap.3 sobre el cual simulamos estos eventos *dropout* con el objetivo de comprender su funcionamiento y entender cuáles son sus limitaciones para su uso.

5.1. Algoritmos de imputación

La forma en que se realiza la inferencia es distinta en cada método. El método MAGIC fue el primero introducido que explícitamente imputa los perfiles de expresión de células únicas a nivel genómico (Dijk et al., 2017). Imputa valores ausentes compartiendo información de células similares basándose en la idea de un proceso difusivo implementado sobre una red de similaridad celular. Se construye una matriz de transición de Markov normalizando la matriz de similitud de células únicas. Para la imputación de una célula los pesos de las demás células los determina la matriz de transición. También existe SAVER (Huang et al., 2017) que toma prestada información de genes usando un enfoque bayesiano para estimar niveles de expresión. Ambos métodos alteran todos los niveles de expresión de los genes incluyendo a aquellos que no son afectados por los *dropouts*. Esto introduce un nuevo bias en los datos que posiblemente elimine las variaciones de interés biológico.

A diferencia de estos dos métodos, Simpute (Li and Jessica Li, 2018) parte de la base de que es inapropiado tratar a todas las cuentas nulas como valores perdidos por una limitación técnica ya que algunos de ellos puede reflejar un valor verdadero en sentido biológico de no expresión del gen en la célula. Entonces, plantean un método que se propone diferenciar cuáles son los valores afectados por los *dropouts* y cuales corresponderían a *ceros biológicos*, para solamente imputar sobre los primeros. Para esto Scimpute realiza los siguientes dos grandes pasos.

1 Modela la probabilidad de cada gen de sufrir *dropouts* en una dada comunidad de células vecinas basándose en un modelo mixto que es ajustado a partir de los datos.

2 Imputa a los que tienen mayor probabilidad de *dropouts* en la célula considerando información sobre el mismo gen en células similares que son seleccionadas basándose en información de genes que presentan poca probabilidad en ser afectados por los *dropouts*.

El primer paso del algoritmo es normalizar la *matriz de expresión* (X^c) por el tamaño de la librería de cada muestra de célula única de forma que todas las muestras tengan un millón de lecturas (X^N). Luego se realiza una transformación de log_{10} según ec.5.1 donde $i = 1, 2, ..., N$, $j = 1, 2, ..., M$, N total de genes y M total de células. La razón es esta transformación es prevenir que observaciones de valores altos de expresión sean extremadamente influyentes y que los valores transformados sean continuos. Se le agrega una pseudo-cuenta de 1.01 para no tener valores infinitos en parámetros de estimación más adelante.

$$X_{ij} = log_{10}(X_{ij}^N + 1{,}01) \tag{5.1}$$

Como la imputación se hace tomando información sobre el mismo gen en células similares, uno de los pasos más importantes es la determinación de grupos, o vecindades, de células en algún sentido parecidas.

Scimpute procede de la siguiente manera:

- 1. Se realiza PCA sobre X^N y se obtiene Z cuyas filas son PCs. Así reducimos el impacto de las grandes proporciones de valores de *dropouts*. Las PCs se seleccionan de manera que por lo menos un 40 % de la varianza de los datos pueda ser explicada, esto quiere decir que la varianza acumulada de las PCs sea igual a 0.4.
- 2. Basándonos en los datos transformados por PCA (la matriz Z) calculamos matriz de distancias entre células que llamamos D. Para cada célula j la distancia a su vecino más cercano es l_j. Entonces, para el set $L = l_1, ..., l_j$ definimos el primer cuartil Q_1 y el tercero Q_3 y definimos el conjunto de los outliers, O, como aquellas células que no tienen vecinos cercanos:
$$O = \{j : l_j > Q_3 + 1{,}5(Q_3 - Q_1)\} \tag{5.2}$$
- 3. Las células que no son outliers son particionadas en K comunidades por una técnica espectral (Ng et al., 2001) explicado en Ápendice A y se designa como $g_j = k$ si la célula j es asignada a la comunidad k ($k = 1, ..., K$) de forma que la célula j tiene como vecinos al conjunto $N_j = \{i : g_j = g_i, j \neq i\}$. En particular, este conjunto para los outliers se define como vacio y para estas no se imputa ningún valor ni tampoco se utiliza su información para imputar las expresiones de genes en otras células.

Una vez definidos los vecinos de cada célula, se procede a inferir en qué células los genes son afectados por los *dropouts*. El punto importante está en que en vez de tomar a todos los valores nulos como eventos *dropouts* (como MAGIC y SAVER), construimos un modelo estadístico para sistemáticamente determinar cuando un valor nulo proviene de un evento *dropouts* y cuando no. El modelo se basa en afirmar que los genes tienen un patrón de expresión bimodal. Dicho patrón se lo puede describir como un modelo mixto de dos componentes: una que es una distribución gamma que representa a los eventos *dropouts* y otra componente que es una distribución normal que representa a la expresión de los genes (Welch et al., 2016).

La idea detrás de este modelo es que si sucede que un gen tiene una expresión alta y relativamente bien definida (variación baja) en la mayoría de las células pertenecientes a una comunidad, la expresión nula de dicho gen para una célula de la misma comunidad será indicativa de un evento *dropout*, una limitación técnica. Por el contrario, si el gen tiene niveles de expresión bajos con una gran variabilidad es más probable que una expresión cero sea evidencia de un *cero biológico*.

Para tener en cuenta que la distribución de la expresión para cada gen puede diferir en distintos tipos de células asociados a procesos biológicos particulares, construimos modelos mixtos diferentes para sub-poblaciones de células que son definidas por la partición en comunidades. Para cada gen, dentro cada comunidad k de células, se estiman los 5 parámetros: $\lambda_i^{(k)}, \alpha_i^{(k)}, \epsilon_i^{(k)}, \mu_i^{(k)}$ y $\sigma_i^{(k)}$ que definen la densidad de probabilidad en ec. 5.3 utilizando un algoritmo de maximización de la expectación (EM).

En consecuencia, para cada gen i-ésimo, su expresión en la población de células de la comunidad k es modelada por una variable aleatoria llamada $x_i^{(k)}$ con la función de densidad de probabilidad según ec.5.3 donde $\lambda_i^{(k)}$ pondera el peso de la distribucón Gamma asociada a *dropouts*, en la comundiad k-ésima de células, $\alpha_i^{(k)}$ y $\epsilon_i^{(k)}$ son el parámetro forma y el parámetro de proporción (la inversa del parámetro de escala), que caracterizan a la distribución gammma y $\mu_i^{(k)}$ y $\sigma_i^{(k)}$ son la media y la desviación estándar de la distribución normal utilizada para modelar niveles de expresión correctamente medidos.

$$f_{x_i^{(k)}}(x) = \lambda_i^{(k)} Gamma(x; \alpha_i^{(k)}, \epsilon_i^{(k)}) + (1 - \lambda_i^{(k)}) Normal(x; \mu_i^{(k)}, \sigma_i^{(k)}) \tag{5.3}$$

La distribución Gamma se define según ec.5.4 donde Γ es la función gamma. Se parametriza por α y ϵ tal que $E(Gamma(x)) = \frac{\alpha}{\epsilon}$ y $Var(Gamma(x)) = \frac{\alpha}{\epsilon^2}$.

$$Gamma(x; \alpha_i^{(k)}, \epsilon_i^{(k)}) = \frac{\epsilon^{\alpha}}{\Gamma(\epsilon)} x^{\alpha-1} e^{-\epsilon x} \tag{5.4}$$

A partir de la densidad de probabilidad asociada a la expresión de cada gen i-ésimo, la probabilidad de *dropout* del gen en la célula j-ésima (perteneciente a la comunidad k) puede ser estimada como d_{ij} según ec.5.5 donde $\hat{\lambda}_i^{(k)}$ es el valor medio de la fracción de *dropouts* del gen i-ésimo en las células de la comunidad k. Obsérvese que el gen i-ésimo tiene distintas probabilidades de *dropouts* en las células j-esimas dentro de la comunidad.

$$d_{ij} = \frac{\hat{\lambda}_i^{(k)} Gamma(X_{ij}; \alpha_i^{(k)}, \epsilon_i^{(k)})}{\hat{\lambda}_i^{(k)} Gamma(X_{ij}; \alpha_i^{(k)}, \epsilon_i^{(k)}) + (1 - \hat{\lambda}_i^{(k)}) Normal(X_{ij}; \mu_i^{(k)}, \sigma_i^{(k)})} \tag{5.5}$$

El valor de d_{ij} da cuenta de la fracción de probabilidad (*a-posteriori*) explicada únicamente por la distribución gamma y por lo tanto brinda información sobre si se trata de un evento *dropout* que debe ser imputado. En consecuencia, para cada célula, seleccionamos un conjunto de genes denominado A_j que requieren de la imputación basándonos en la probabilidad de estos: $A_j = \{i : d_{ij} > t\}$, donde t es un valor cota sobre las probabilidades de *dropouts*.

Por otro lado, también construimos el conjunto de genes $B_j = \{i : d_{ij} < t\}$ que, como tienen asociadas probabilidades de dropout d_{ij} por debajo de la cota, de acuerdo a nuestro modelo, son

expresiones con baja probabilidad de ser afectadas por *dropouts* y por consiguiente, no requieren de imputación. De esta forma, este es un conjunto de genes cuyas expresiones en la célula j-ésima son confiables. Por ello, se utilizarán para imputar la expresión del conjunto A_j.

Asumiendo que la similitud entre células se extiende a todos los genes estudiados ($A_j \cup B_j$) podemos inferir la expresión de los genes del conjunto A_j a partir de la expresión que tienen estos en las células caracterizadas como similares a partir de los perfiles de expresión de genes B_j.

Cuantificaremos la similitud entre células pertenecientes a una misma comunidad pues aquellas células de comunidades distintas tienen similutud nula. El grado de similitud lo expresan los coeficientes $\beta^{(j)}$ tal que para la célula j-ésima, estos coeficientes son los que minimizan la distancia (en norma dos) de $X_{B_j,j}$ (la expresión de los genes del conjunto B_j en la célula j) a una combinación lineal pesada por los coeficientes $\beta^{(j)}$ de la expresión de estos genes en las demás células de la comunidad representa por $X_{B_j,N_j}\beta^{(j)}$.

Esta minimización hace por una regresión por cuadrados mínimos no negativa (NNLs, del inglés non-negative least squares) según ec.5.6 donde B_j es el conjunto de genes confiables, N_j representa los índices de las células vecinas de la célula j-ésima (pertenecen a la misma comunidad) de manera que $X_{B_j,j}$ es el transcriptoma de la célula j-ésima en el espacio de confianza y X_{B_j,N_j} es la sub-matriz de los transcriptomas de las células vecinas en el espacio de confianza.

$$\hat{\beta}^{(j)} = argmin_{\beta^{(j)}} ||X_{B_j,j} - X_{B_j,N_j}\beta^{(j)}||_2^2, \beta^{(j)} \geq 0 \tag{5.6}$$

Así se obtiene $\hat{\beta}^{(j)}$ que es el vector de coeficientes de tamaño $|N_j|$ que es un estimador que puede ser esparso, cuyos elementos representan la similitud que tienen las células vecinos (N_j) según la partición en comunidades. Es este el estimador que utilizamos para imputar los genes del conjunto A_j en la célula j-ésima:

$$\hat{X}_{ij} = \begin{Bmatrix} X_{ij} & i \in B_j \\ X_{i,Nj}\hat{\beta}^{(j)} & i \in A_j \end{Bmatrix} \tag{5.7}$$

En conclusión, los pasos del algoritmo son los siguientes:

- 1. Construimos la matriz X, proyectamos en PCA para obtener Z a partir de la cual construimos matriz de distancias D. Reconocemos células outliers sobre las cuales no vamos a trabajar.
- 2. El sub-grafo (las células restantes) es particionado para reconocer las K comunidades.
- 3. Para la comunidad k-ésima calculamos la función densidad para cada gen. Con estos parámetros que fueron estimados por un algoritmo de maximización de la expectación

(EM) y el cálculo del valor medio de fracción de *dropouts* de cada gen en la comunidad k-ésima $\hat{\lambda}_i^{(k)}$ calculamos la probabilidad de *dropouts* d_{ij}.

- 4. Para una dada célula j-ésima, dado un valor cota t construimos los conjuntos de genes confiables B_j y el de aquellos que vamos a imputar A_j. A partir de los confiables el algoritmo aprende el grado de similitud entre la célula j-ésima y las células de su comunidad N_j. Esto es cuantificado por el estimador $\hat{\beta}^{(j)}$ que es calculado por cuadrados mínimos no negativos.

- 5. Para la dada célula j-ésima imputamos el valor para los genes del conjunto A_j a partir de la expresión de cada gen en las células similares que aprendimos gracias a la información que contienen los genes confiables B_j.

  Repetimos 4 y 5 para las M células.

5.2. Estudio del funcionamiento del algoritmo Scimpute

Utilizamos el modelo de *matriz de expresión* que simula la evolución dinámica de grupos de genes en experimentos scRNASeq descripta en Cap. 3. Para simular además el reporte de expresiones nulas debido a una limitación técnica experimental modelamos la distribución de probabilidad de eventos dropout. Construimos una distribución de densidad de probabilidad de Bernoulli (una binomial de tamaño uno) con una probabilidad de fracaso igual a la probabilidad de que haya *dropouts*. Esperamos que si el valor medio de la expresión del gen en las células muestredas es baja, entonces la probabilidad de dropout sea alta. En consecuencia, en particular, modelamos esta probabilidad como $p = e^{-0,1<g>}$ donde $< g >$ es la expresión media del gen en las células de la matriz (Welch et al., 2016). Así se logra una tasa de dropout de 0.95 que es similar a la observada en los datos experimentales de Linnarson (ver Cap.4).

Con el objetivo de entender cómo es la operatoria del algoritmo comenzamos considerando una matriz de 1250 células para la expresión de los 1110 genes. Luego de normalizar a la matriz y de calcular el logaritmo como en ec.5.1, el siguiente paso del algoritmo es seleccionar aquellos genes que son variables. Se entiende como genes variables a aquellos que en un plano donde se grafica al coeficiente de variación (CV) en función del valor medio (m) de los genes a lo largo de las células estudiadas (Fig. 5.1), son aquellos genes tales que $m > 1$ y $CV > Q_1$ (primer cuartil). La idea detrás de esto es restringirnos a aquellos genes cuya variabilidad con respecto a la media se encuentra por encima del 25 porciento más bajo y que además son expresados en promedio significativamente en el sentido de tener una expresión media superior a 1.

Observamos en Fig. 5.1 que las zonas de mayor densidad son por un lado, para un m alto y CV bajo y por otro, m y CV bajo. Las cotas propuestas (indicadas por las líneas rojas) permiten no considerar a aquellas que se corresponden con valores bajos de m y de CV y también a aquellas que aunque presentan un alto m, no tienen un nivel de variabilidad mínimo.

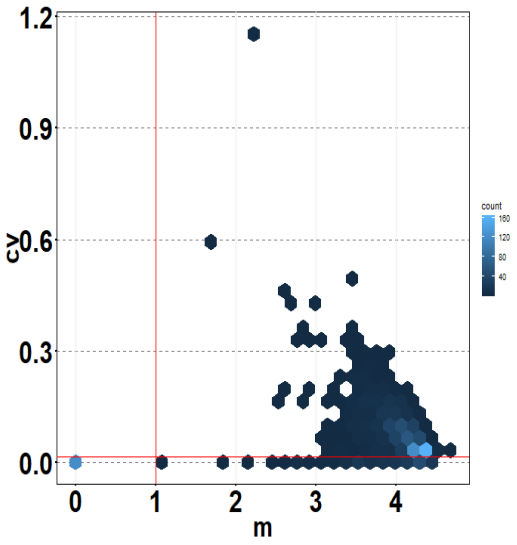


FIGURA 5.1: Coeficiente de variación (CV) en función del valor medio (m) de los 1110 genes lo largo de las células muestreadas en la *matriz de expresión* . El color representa la densidad de puntos. La línea horizontal roja representa al CV cota tal que se toman como variables aquellos con valores superiores y la línea vertical indica el valor de m tal que se selecciona por encima de éste valor.

De esta forma nos reducimos a una matriz de 832 genes que es proyectada por PCA. Se eligen la cantidad de componentes principales que tengan asociada una desviación estándar acumulada normalizada por debajo del 0.4 siendo estas 27.

La matriz D tiene información sobre la similitud entre células pues podemos calcular, para cada célula, la distancia al resto de forma que la célula a la más se asemeja es aquella que se encuentra a la menor distancia. Para entender cómo se da esta relación de similitud en toda la población de células, graficamos la distribución de las menores distancias que unen a cada célula con el resto en Fig.5.2.

Se reconocen en la figura a las células extremales a esta distribución que no tiene vecinos según ec.5.2 que son aquellas que tienen a su vecina por encima de la línea vertical negra. Se observa que la densidad para esta distancia es muy pequeña y son estas células eliminadas de la imputación porque se cree que pueden ser el resultado de errores experimentales o también representar variaciones biológicas particulares.

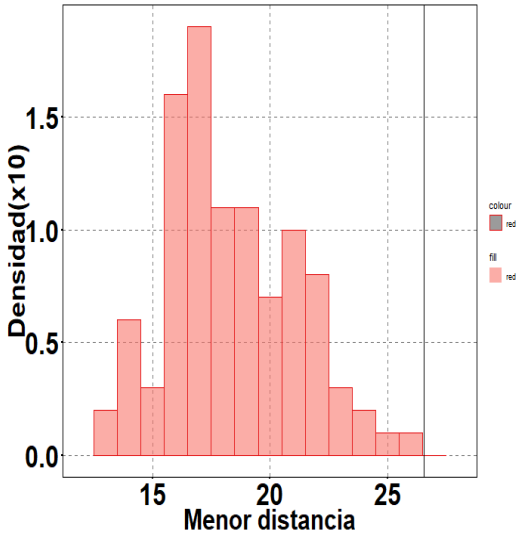


FIGURA 5.2: Distribución de la menor distancia que une a cada una de las células con su vecina, en el sentido de vecindad como cercanía en distancia euclídea a la misma. La línea negra vertical representa la distancia cota que se corresponde con ec.5.2 tal que aquellas células por encima de dicho valor, las extremales, son eliminadas del proceso de imputación.

Luego de eliminar en la matriz D aquellas células determinadas como extremales, esta es particionada en comunidades por un algoritmo espectral (ver Ápendice A). Para cada una de estas comunidades se calcularán los parámetros para el ajuste de la distribución mixta sobre las expresiones no nulas de cada gen.

En consecuencia, si el gen en esa población de células tiene una única expresión por encima del cero, la distribución normal colapsa a una distribución del tipo delta de Dirac y entonces en ec.5.3 al calcular la función distribución mixta para dicho gen tenemos infinito para un cierto valor de x. Para regular las iteraciones del método del ajuste de los parámetros por maximización de la expectación (EM) se pide $\epsilon = \sum_{j=1}^{N}{}_k(log10(f_{x_{ij}}^{k+1}) - log10(f_{x_{ij}}^{k})) < 0{,}5$, i es el gen y k la iteración. Para el caso de sólo una expresión por encima del cero, ϵ es infinito y el método no converge. Estos genes son los que llamamos inválidos para la comunidad k pues sus expresiones en dicha comunidad no pueden ser ajustadas a la distribución mixta y sobre ellos, no se va a poder realizar la imputación.

Para aquellos genes cuyas expresiones en la comunidad son modelables por el modelo mixto se calculan los cinco parámetros de la misma $\lambda_i^{(k)}, \alpha_i^{(k)}, \beta_i^{(k)}, \mu_i^{(k)}$ y $\sigma_i^{(k)}$. El modelo calcula la probabilidad de *dropouts* como el parámetro $\lambda_i^{(k)}$. Como estamos trabajando sobre una *matriz de expresión* que es un modelo, se espera que $\lambda_i^{(k)}$ se asemeje a la fracción de elementos nulos que impusimos para cada gen en la comunidad k siendo la fracción dependiente de la expresión media observada del gen.

En la Fig.5.3 se muestra el valor de $\lambda_i^{(k=1)}$ obtenido para genes que presentan un número de mediciones nulas variable dentro de la comunidad $k = 1$. Graficamos este parámetro correspondiente a cada uno de los genes en función de la fracción de elementos nulos que presentan

estos en la comunidad. Podemos ver que existe una relacion lineal entre ambos observables, con una tendencia a sobre-estimar $\lambda_i^{(k=1)}$ para fracciones altas de dropout (observamos que para las cinco comunidades restantes un comportamiento semejante). Notamos además que a pesar que la gran mayoría de las expresiones medidas son nulas (95 %), existen genes que presentan un número de dropout sensiblemente más bajo pues la probabilidad de dropout es dependiente de la media observada de las expresiones del gen.

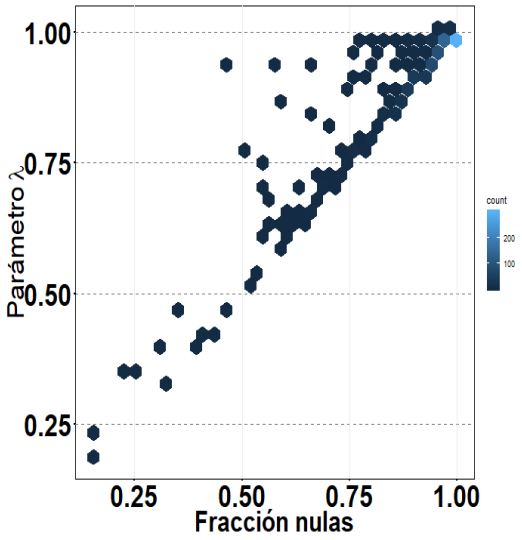


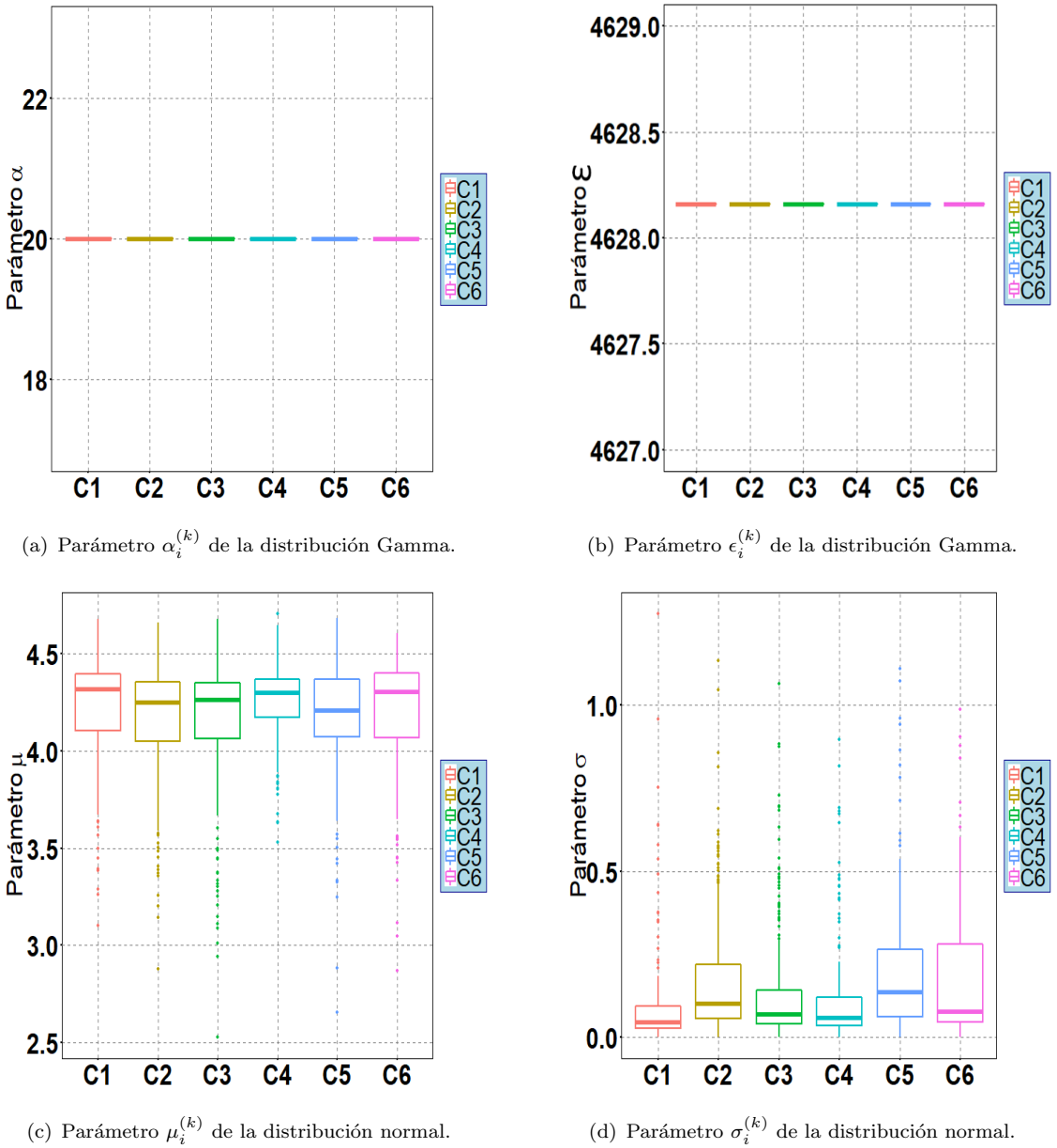
FIGURA 5.3: Las probabilidades de *dropouts* de cada gen calculada por la distribución de probabilidad mixta del modelo para la comunidad 1, $\lambda_i^{(1)}$, en función de la fracción de expresiones nulas de dichos genes en las células de la comunidad. El color representa la densidad de puntos.

Por otro lado graficamos la distribución del resto de los parámetros del ajuste (para todos los genes) para cada comunidad en un gráfico de caja (Fig. 5.4). Se observa en la parte superior (paneles A y B) que ambos parámetros que describen a la distribución Gamma son independientes de la comunidad. Estas distribuciones de probabilidad presentan altos valores del parámetro de proporción ϵ, compatibles con una distribución del tipo delta centrada en la expresión nula correspondiente con un evento dropout. La baja variabilidad en ambos parámetros se debe a que la distribución impuesta de *dropouts* es una Bernoulli con probabilidades más o menos similares.

Por el contrario, los parámetros ajustados de la distribución normal (que modela a las expresiones que no son *dropouts*) que se muestran en la parte inferior en los paneles C y D presenta una distribución no trivial. En la *matriz de expresión* modelamos dos grupos de genes que aumentan su expresión y otros grupos que la decrecen, con la misma parametrización de forma que la mediana del $\mu_i^{(k)}$ no cambia significativamente entre comunidades (ver Cap.3). Aunque hay genes que modelados como persistentes y de ruido blanco cuyas expresiones se alejan de $\mu_i^{(k)}$, están presentes en igual proporción en cada comunidad.

La razón de la diferencia en las distribuciones entre comunidades es cómo es la evolución dinámica. Las dos comunidades del medio son aquellas en las que se dan los puntos intermedios de

las curvas que disminuyen y aumentan respectivamente y donde estas se intersectan. En consecuencia, el $\sigma_i^{(k)}$ es menor. Luego, las curvas vuelven a separarse y aumenta $\sigma_i^{(k)}$. La razón por la que para C1 también la mediana es baja y hay poca variabilidad es atribuida a los genes que son expresados por más de un grupo dinámico. Es posible que la expresión tome una forma tal que en las células de C1 $\sigma_i^{(k)}$ sea chico.



(a) Parámetro $\alpha_i^{(k)}$ de la distribución Gamma.

(b) Parámetro $\epsilon_i^{(k)}$ de la distribución Gamma.

(c) Parámetro $\mu_i^{(k)}$ de la distribución normal.

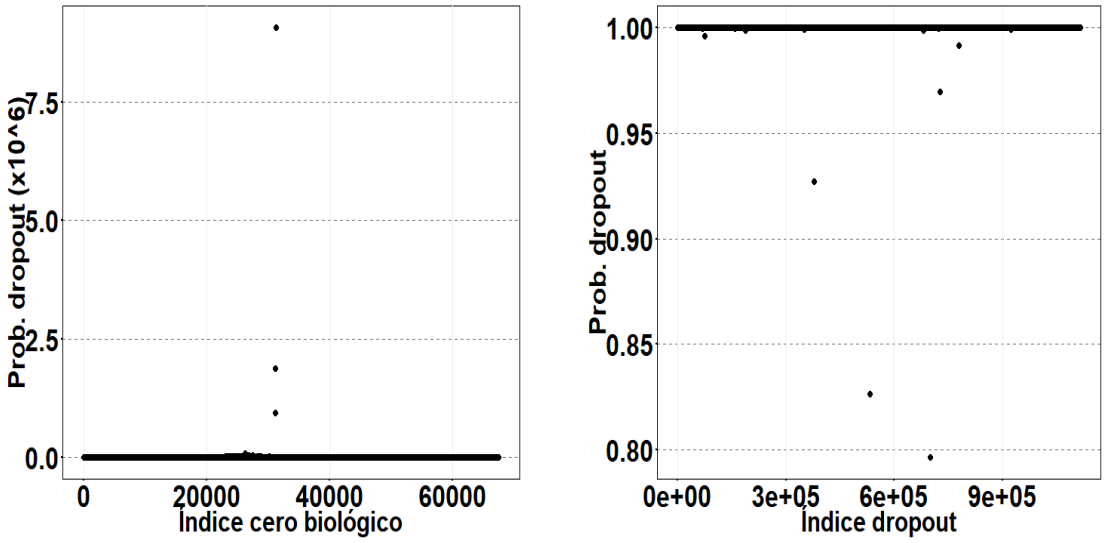
(d) Parámetro $\sigma_i^{(k)}$ de la distribución normal.

FIGURA 5.4: Las distribuciones de los parámetros de la distribución de probabilidad mixta para cada comunidad señalas tanto en el eje de abscisas como con los colores. La parte superior (paneles A y B) se corresponde a los respectivos de la distribución gamma y la inferior a la normal (paneles C y D). Se obtienen por medio de un algoritmo de maximización de expectación (EM) sobre los valores de expresión para cada gen, seleccionados como válidos, sobre las células de la respectiva comunidad.

A partir de la distribución de probabilidad mixta se realiza, para cada célula j de una dada comunidad, la distinción de genes que parcerían haber sufrido dropout técnico (grupo de genes A) de los que no lo hacen (grupo de genes B). Estos últimos representan los llamados genes confiables que no se van a imputar sino que se utilizarán como información en el proceso de imputación de los otros genes que son lo que llamamos genes a imputar, A_j. En concreto, si la probabilidad inferida de dropout d_{ij} (ver ec.5.5) resulta menor a una cota, el gen resulta confiable, mientras que si la supera será un gen cuyo valor de expresón será imputado.

Para entender cómo determinar esta cota graficamos la probabilidad de dropout en función de un índice de la expresión en Fig. 5.5. En el panel A se muestra la probabilidad asociada a expresiones que son nulas por la dinámica simulada, aquellas que representan *ceros biológicos*. Observamos que la probabilidad se acumula en el orden de 10^{-6} compatible con que representan expresiones no afectadas por *dropout*. Por otro lado, en el panel B se muestran las probabilidades asociadas a expresiones que simulamos como eventos *dropout* de forma que las probabilidades son muy altas tal que se acumulan alrededor del 1.

Esta probabilidad nos permite reconocer dos comportamientos, uno asociado a una probabilidad del orden de 10^{-6} y otro de una probabilidad muy cercana a 1. Estos dos casos se corresponden con las expresiones nulas que modelamos: ceros biológicos que se corresponden con dinámicas simuladas y los *dropouts* que impusimos según la distribución de Bernoulli.



(a) Expresiones nulas biológicas

(b) Evetos dropout

FIGURA 5.5: Distribuciones de la probabilidad de dropout de las expresiones simuladas graficadas en función de índices que representan estos niveles. Se diferencia en cada panel si son expresiones que son ceros simulados por la dinámica que sigue el gen o si se corresponde con el modelado de dropout.

En consecuencia, para el modelo con el que estamos trabajando la cota de 0.5 permitiría reconocer claramente ambos comportamientos y se corresponde con lo que proponen los autores de la técnica (Li and Jessica Li, 2018) que afirman que ese valor es suficiente para distinguir de forma clara aquellas expresiones que deben ser imputadas de las que no.

5.3. Imputación por Scimpute

En esta subsección estudiaremos la performance de la imputación variando dos parámetros que caracterizan a la matriz de cuentas del modelo: la cantidad de células y la fracción de elementos nulos de la matriz. Para ello dejamos constante uno y variamos el otro para un valor bajo y otro alto como se muestra en Tabla.5.1 que describe los tres tipos de experimentos numéricos realizados. También, como ya vimos, el algoritmo de imputación es dependiente de cómo es particionado el grafo de células que resulta de la *matriz de expresión*. Por eso, estudiamos también cómo cambia la partición para el caso de una fracción de *dropouts* alta y baja.

Este trabajo lo realizamos sobre el modelo de matriz sintética II (desarrollado en Cap.3) compuesta por 1110 genes cuya evolución dinámica es parametrizada por el tiempo de simulación. La cantidad de genes es una constante en este trabajo mientras que la cantidad de células no, podemos variar la granularidad. El vector del tiempo de simulación toma valores entre 0 y 2 y le vamos a ir cambiando la cantidad de pasos N.

Objeto de estudio	**Parámetro fijo**	**Parámetro variable**	**Bajo**	**Alto**
Cantidad células muestreadas	Fracc. elementos nulos $p = e^{-0,1<g>}$ Total = 0.95	Cantidad células	125	1250
Fracc. dropouts	Cant. células = 1250	Fracc. elementos nulos	$p = 0.4$ $Total = 0.49$	$p = e^{-0,1<g>}$ $Total = 0.95$
Partición en comunidades	Cant. células = 1250	Fracc. elementos nulos	$p = 0.4$ $Total = 0.49$	$p = e^{-0,1<g>}$ $Total = 0.95$

Tabla 5.1: El parámetro fijo y el variable (con el respectivo valor bajo y alto) para cada uno de los objetos de estudio. Para el caso de variar la fracc. elementos nulos, se varía la probabilidad p y se indica la fracción total de dropouts modelado.

Para poder determinar si el algoritmo es exitoso o no sobre el modelo estudiado desarrollamos métricas. Por un lado cuantificamos qué tan similares son las expresiones imputadas a las expresiones reales del modelo definiendo un observable para cada gen llamado diferencia cuadrado según ec.5.8 donde D_i es la diferencia cuadrado para el gen i-ésimo. El exponente M representa la expresión en el modelo e I la imputada. La sumatoria es hasta M, la cantidad de células modelables para ese gen.

$$D_i^2 = \frac{1}{M}\sum_{j=1}^{M}(E_{ij}^M - E_{ij}^I)^2 \tag{5.8}$$

También cuantificamos que tanto se ve restituida la correlación de Pearson entre las expresiones de genes de un mismo grupo, los crecientes A y C y los decrecientes B y D respectivamente, en comparación con el valor en el modelo. Calculamos el p valor asociado a la correlación con una distribución t de Student. Lo que se espera es que los *dropouts* disminuyan la correlación adentro de cada grupo y que luego de la imputación, las correlaciones sean más o menos reestablecidas a cómo son en el modelo.

La forma de simular los eventos *dropouts* es, tal como explicamos en Secc.5.2, a partir de una distribución de probabilidad de *dropouts* que es modelada como una Bernoulli caracterizada por una probabilidad p. El caso que más se asemeja a lo visto en experimentos es un fracción total ~ 0.95 y una dependencia exponencial de la fracción de estos eventos en la expresión de un gen con el valor medio de su expresión. En efecto, se toma $p = e^{-0,1<g>}$ (Welch et al., 2016). Por otro lado, para estudiar cómo es la imputación para una tasa baja de *dropouts* tomamos de forma uniforme a todos los genes $p = 0.4$ de forma que la fracción total es de 0.49.

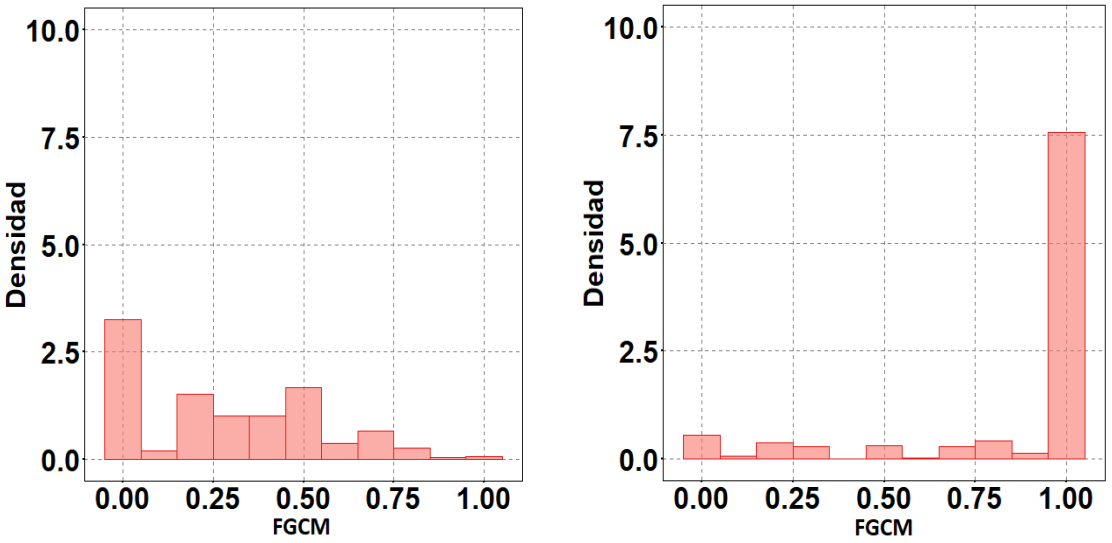
5.3.1. Cantidad de células

Como ya vimos, el algoritmo trabaja a nivel de comunidades de células y calcula la probabilidad de *dropout* a partir de una distribución de probabilidad mixta para las expresiones de cada gen en cada una de las mismas. Es así que, como primer paso, para cada célula se reconocen a los genes modelables que llamamos válidos y a los que no lo son. Estos son genes cuyas expresiones en la comunidad se puede modelar con el modelo de densidad de probabilidad mixto, en el sentido de que el algoritmo EM converge como discutimos previamente.

Estas son las expresiones que serán o bien imputadas o permanecerán igual al ser reconocidas como observables al tener asociada baja probabilidad de dropout. Por otro, las expresiones no modelables no serán ni imputadas ni observadas y será posible reconocerlas, luego de la imputación. Es así como luego de la imputación de toda la *matriz de expresión* , es posible reconocer por gen (en vez de por célula) aquellas expresiones que fueron imputadas, observadas y no fueron modelables. Definimos la fracción por gen de células modelables (FGCM) que

cuantifica sobre cuántas expresiones del gen en las células muestreadas han podido ser o bien imputadas u observadas.

Graficamos un histograma de la distribución de FGCM (Fig.5.6) para el caso de 125 (panel A) y para el de 1250 células muestreadas (panel B). Para el caso de 125 células, la densidad se concentra para fracciones por gen de células modelables bajas. Por el contrario, al aumentar la cantidad de células a 1250, se concentra para una fracción de 1. Concluimos que la FGCM, para estos valores de *dropouts* estudiados, va a depender de la cantidad de células muestreadas en el experimento. En particular, afirmamos que para 1250 células se consiguen valores de FGCM altos de forma que se imputa u observa la mayoría de las expresiones.



(a) 125 células

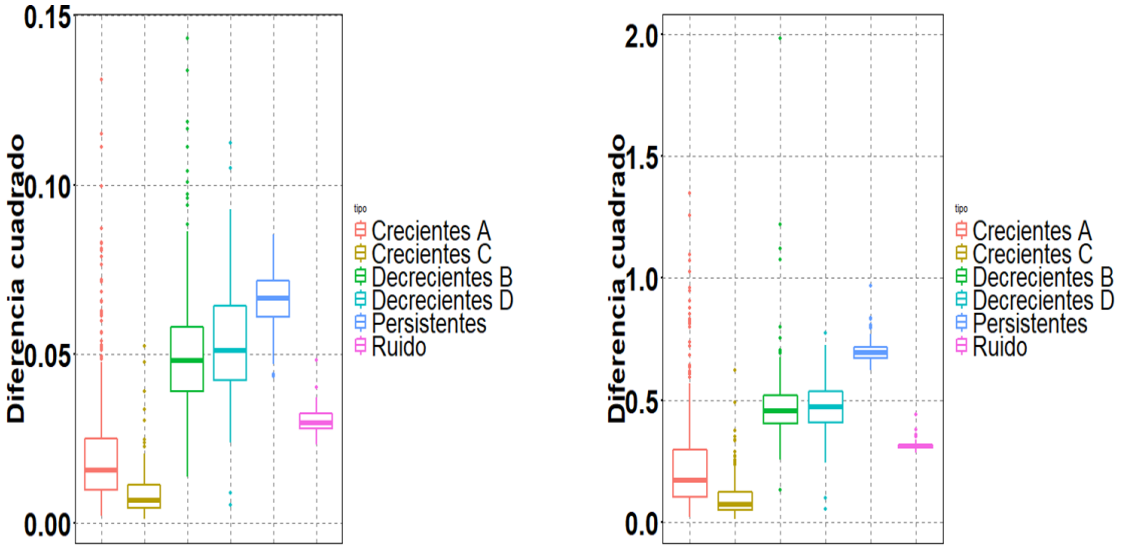
(b) 1250 células

FIGURA 5.6: Distribuciones de la fracción por gen de células modelables para distinta cantidad las células muestreadas indicada en cada panel. Nótese que la cantidad representa granularidad pues se simula a un vector de tiempo de N valores, con N cantidad de células.

Por otro lado, cuantificamos la imputación por medio de la diferencia cuadrado D_i^2 considerando las expresiones de las células válidas, tanto aquellas que fueron imputadas como las que fueron observadas. Graficamos su distribución en gráficos de caja y comparamos el caso para 125 y 1250 células (Fig. 5.7) para cada grupo de dinámica de genes (Ver Secc.3.3) que son referenciados al costado de las figuras y por sus colores.

En primer lugar notamos que las distribuciones tienen una varianza significativa. El motivo de ello es que los genes pertenecientes al mismo grupo dinámico siguen una evolución parametrizada por el tiempo a la que le sumamos un ruido gaussiano de $\sigma^2 = 0.02$. Observamos que la imputación que se realiza para el caso de una mayor cantidad de células es peor que en el caso de menor células en el sentido de que se obtienen valores de D_i^2 de un orden de magnitud

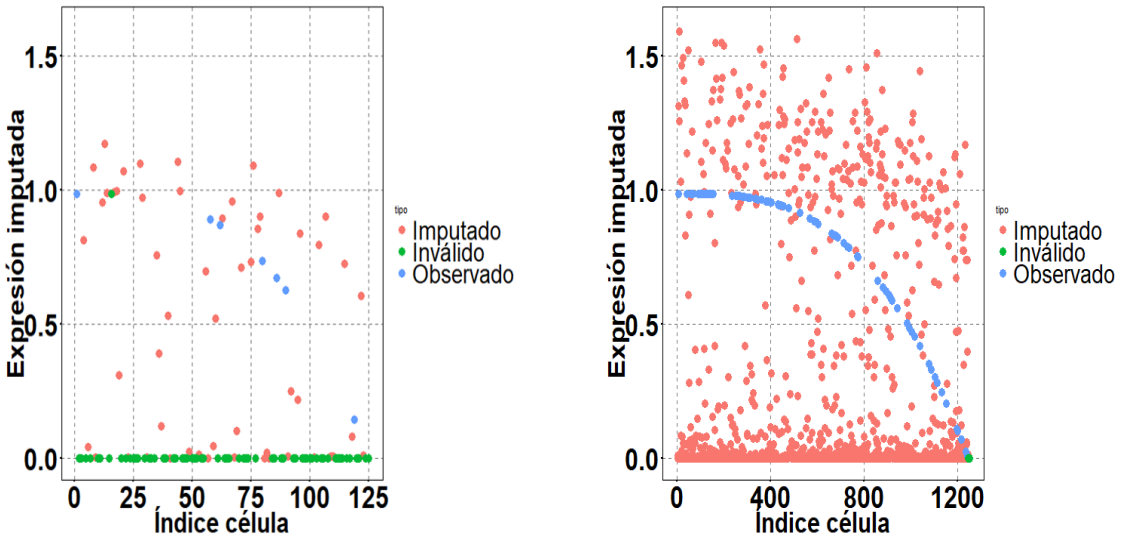
menor. Afirmamos que aunque una mayor granularidad provee de una mayor fracción de células modelables de forma que una mayor cantidad de expresiones se imputan, esta imputación significativamente peor.



(a) 125 células (b) 1250 células

FIGURA 5.7: Gráficos de caja para las distribuciones de la diferencia cuadrado D_i^2 sobre cada grupo de genes respectivamente señalados con el color. Se trata de una medida normalizada por la cantidad de células modelables en cada caso respectivamente. Notar que las escalas son distintas con el fin de que se pueda visualizar correctamente.

Este fenómeno que al aumentar la cantidad de células se incrementa la fracción de células imputables pero, la imputación es estas es peor en el sentido de la diferencia cuadrado lo podemos ver para un caso particular de un gen que es decreciente. Se grafica la expresión para este gen a lo largo de las células (indexadas en el orden del tiempo de simulación) y se diferencia con el color a aquellas expresiones que fueron inválidas, observadas e imputadas (Fig. 5.8). Se observa que para el caso de 125 células son muchas las consideradas como inválidas y éstas son, en su mayoría, valores nulos. Para el caso de 1250 la cantidad de células inválidas es reducida y vemos que la imputación (color salmón), en su mayoría, sobre-estima o sub-estima el valor del modelo.



(a) 125 células

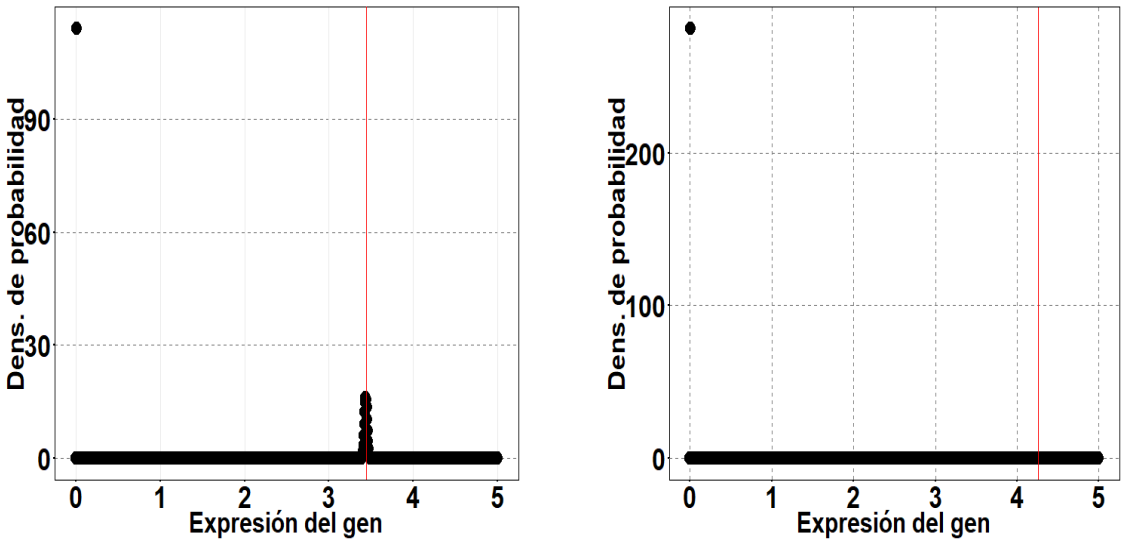
(b) 1250 células

FIGURA 5.8: Evolución reconstruida por la imputación de la expresión de un gen decreciente. Las células se encuentran indexadas y ordenadas en el sentido dado por la evolución del tiempo de simulación. El color indica si se trata de un expresión que fue imputada (salmón), observada (celeste) e inválido (verde).

5.3.2. Fracción de *dropouts*

El método de la imputación es dependiente también de la fracción de expresiones nulas en la matriz. Estudiamos su dependencia con este parámetro modelando la distribución de *dropouts* sobre la matriz como una distribución de Bernoulli con una probabilidad $p = e^{-0,1<g>}$ (Pierson and Yau, 2015) para una alta tasa de *dropouts* y para una baja, tomamos de forma uniforme a todos los genes $p = 0.4$ de forma que la fracción total es de 0.49.

Como primer acercamiento para entender cómo puede influir esta diferencia de la tasa de *dropouts* sobre la imputación, veamos para una dada comunidad y un dado gen, cómo cambia el ajuste del modelo de la distribución mixta al reducir la fracción de elementos nulos. Para ello, graficamos ambas densidad de probabilidad en Fig. 5.9.



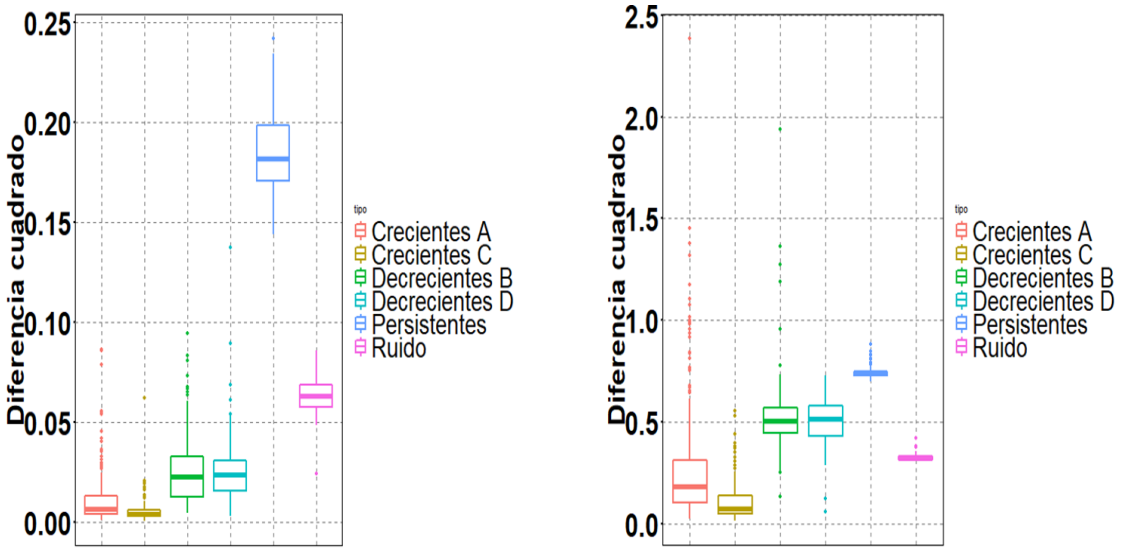
(a) 0.49 de la matriz son elementos nulos. Los parámetros del ajuste: $\alpha = 20$, $\beta = 4628.16$, $\lambda = 0.40$, $\mu = 3.43$ y $sd = 0.015$

(b) 0.95 de la matriz son elementos nulos. Los parámetros del ajuste: $\alpha = 20$, $\beta = 4628.16$, $\lambda = 0.997$, $\mu = 4.24$ y $sd = 0.009$

FIGURA 5.9: Los parámetros de la distribución son ajustados para la expresión de los genes considerados como válidos para cada una de las comunidades y se parametriza por la expresión de cada gen en las células de la comunidad. La línea roja señala el valor de expresión medio (μ) en que se ajusta la distribución normal. La distribución gamma se encuentra concentrada en un punto que se corresponde con las expresiones nulas normalizadas (log10(1.01)).

Observamos que aunque la función Gamma de la distribución no cambia (parámetros en el pie de Fig.5.9), sí lo hace significativamente la distribución normal. Mientras que al ser la fracción de nulas tan grande (panel B) no se distingue la distribución normal, al disminuir la cantidad de elementos nulos (panel A) se la visualiza no sólo por tener asociado un $sd = 0.15$ significativo sino también porque el λ aumenta su valor de 0.4. hasta 0.997.

Por otro lado, al considerar una distribución de Bernoulli con una probabilidad $p = 0.4$ se obtienen muy buenos resultados desde el punto de vista de la diferencia cuadrado. Observamos las distribuciones asociadas a cada conjunto de dinámica de genes en Fig. 5.10.



(a) 0.49 de la matriz son elementos nulos

(b) 0.95 de la matriz son elementos nulos

FIGURA 5.10: Gráficos de caja para las distribuciones de la diferencia cuadrado D_i^2 sobre cada grupo de genes respectivamente señalados con el color. Se trata de una medida normalizada por la cantidad de células modelables en cada caso respectivamente. Notar que las escalas son distintas con el fin de que se pueda visualizar correctamente.

Observamos que se consigue reconstruir la trayectoria de los genes con mayor precisión para una menor fracción de *dropouts* ya que D_i^2 toma valores de hasta un orden de magnitud menor que para el caso de 0.95 fracción de ceros.

Otra forma de visualizar la diferencia en perfomance de imputación entre un régimen de bajo y otro de alto nivel de dropout es a partir de las expresiones en la *matriz de expresión* génica. En Fig. 5.11 se muestra para el modelo simulado sin *dropouts*. Se reconocen en el eje vertical los distintos grupos de genes que obedecen una dinámica particular. El color representa el valor de expresión de cada elemento de la matriz cuya escala de indica al costado de la figura.

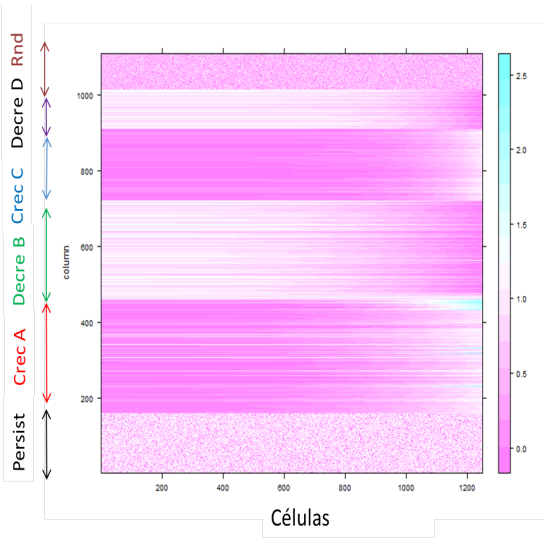
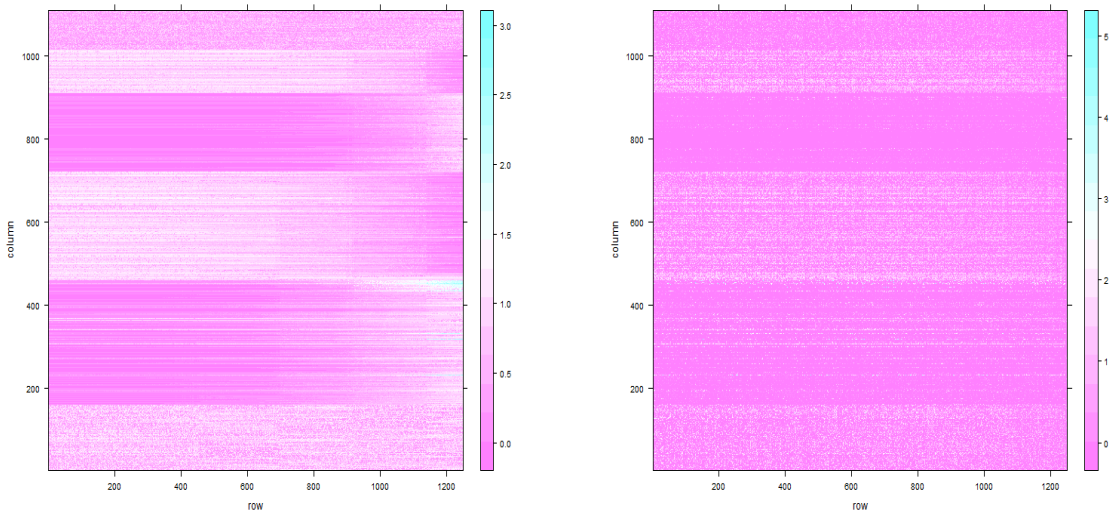


FIGURA 5.11: Gráfico de nivel para la *matriz de expresión* del modelo de 1250 células. En el eje de abscisas se grafican las células y en la ordenada los genes indicando cada grupo dinámico. El color representa el valor de expresión. Los colores de una gama rosada se corresponde con valores bajos, el blaco a un valor alto y el celeste a uno de saturación que se corresponde con aquellos genes que son doblemente expresados por dinámicas crecientes.

Se pueden distinguir claramente los distintos grupos de genes simulados y, en particular, se reconocen las distintas velocidades con que decrecen o decrecen respectivamente las expresiones de los genes dinámicos. Por ejemplo, se ve que el grupo de Decrec D decrece más lento que el grupo Decrec A. Se espera que una buena imputación sea aquella capaz de reconstruir las trayectorias de forma tal que se pueda reconocer estos grupos de genes en la matriz imputada y la dinámica que tienen asociada. Se comparan los casos para la imputación sobre una con una fracción de elementos nulos de 0.49 y sobre una con 0.95 en los paneles de Fig.5.12.



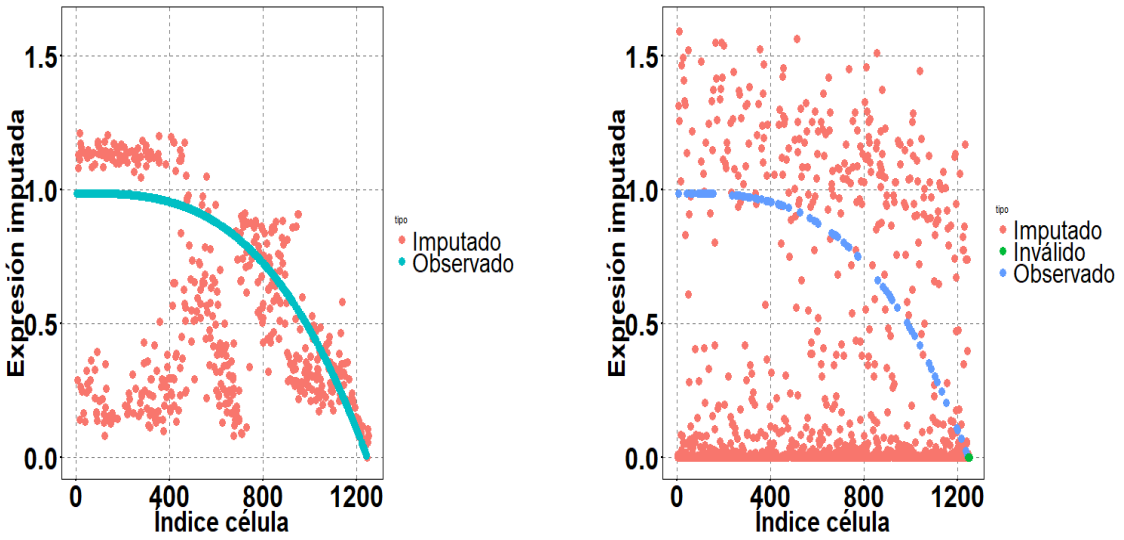
(a) 0.49 de la matriz son elementos nulos

(b) 0.95 de la matriz son elementos nulos

FIGURA 5.12: Comparación de la imputación sobre la matriz que se muestra en Fig. 5.11 para los casos en que un 0.49 de la misma son elementos nulos y para el de un 0.95.

Se ve que mientras que para el caso de una fracción de 0.49 (panel A) se reconstruye de forma tal de poder distinguir los grupos de genes, para el caso de 0.95 no (panel B). Se observa en el caso de la imputación para la matriz con 0.49 de *dropouts* que dentro de un mismo grupo no se da la heterogeneidad que se encuentra en el modelo (Fig.5.11) pero, sin embargo, los grupos de genes son bien delimitados y reconocidos.

Sin embargo, hacemos notar que aún para el caso de bajos niveles de dropout pueden aparecer situaciones problemáticas. El panel (A) de la Fig.5.13 se muestra el resultado de la imputación para el mismo gen que estudiamos en Fig.5.8 al considerar un número reducido droputs simulados.



(a) 0.49 de la matriz son elementos nulos

(b) 0.95 de la matriz son elementos nulos

FIGURA 5.13: Evolución reconstruida por la imputación de la expresión de un gen decreciente. Las células se encuentran indexadas y ordenadas en el sentido dado por la evolución del tiempo de simulación. El color indica si se trata de un expresión que fue imputada (salmón), observada (verde claro en en el panel A y celeste en el panel B e inválido (verde).

A partir de los análisis previos afirmamos que la imputación de una matriz con una fracción alta de *dropouts* no es satisfactoria porque los valores imputados se alejan de los modelados como cuatificamos con D_i^2. También, observamos que en la matriz imputada no se logra identificar los grupos de genes según sus dinámicas. En consecuencia, en la siguiente sección estudiaremos los motivos de por qué la imputación no es buena para datos con dropout alto.

5.4. Estudio de la imputación para fracción alta de *dropouts*

El algoritmo realiza la imputación de la expresión de un gen en una célula proyectando la expresión de dicho gen en las vecinas pesando por los coeficientes β que cuantifican la similitud entre las células. En consecuencia, para entener los motivos del por qué del mal funcionamiento

del algoritmo al trabajar con matrices de fracción alta de elementos nulos, vamos a comparar, para una célula en particular, las distribuciones de los coeficientes β para el caso de una tasa alta (0.95) y una baja de dropout (0.49).

Para eso, graficamos las distribuciones en un gráfico violín y de caja en Fig.5.14 de los coeficientes entre una célula que se corresponde a un estadío inicial de la simulación temporal para con las demás pertenecientes a la comunidad.

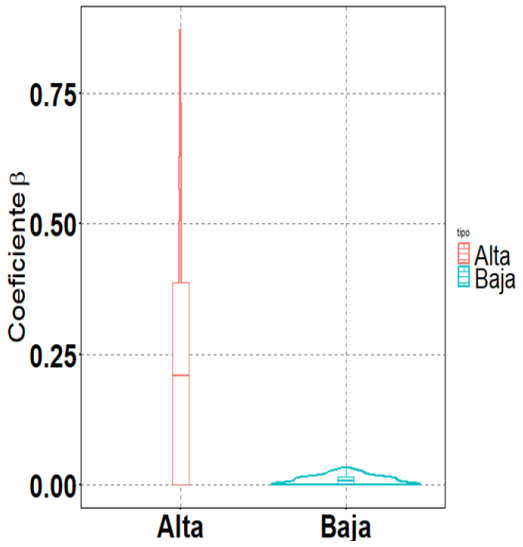


FIGURA 5.14: Distribución de los coeficientes β aprendidos para la imputación de una célula inicial en la evolución (índice 321). Se muestra para una matriz a imputar con alta fracción elementos nulos (salmón) y para otra con baja (celeste). Se indica la densidad con el gráfico de violín insertado con gráfico de caja.

Vemos que mientras que para el caso de una fracción alta, la distribución toma valores en un amplio rango, para una baja fracción se observa que la distribución se encuentra acotada para valores bajos. Esto indica que la similitud que se les da a las células vecinas es bastante uniforme. Por el contrario, para una tasa alta, el valor imputado dependerá mayormente de la expresión correspondiente al valor de β más alto en vez de depender uniformente de varias expresiones.

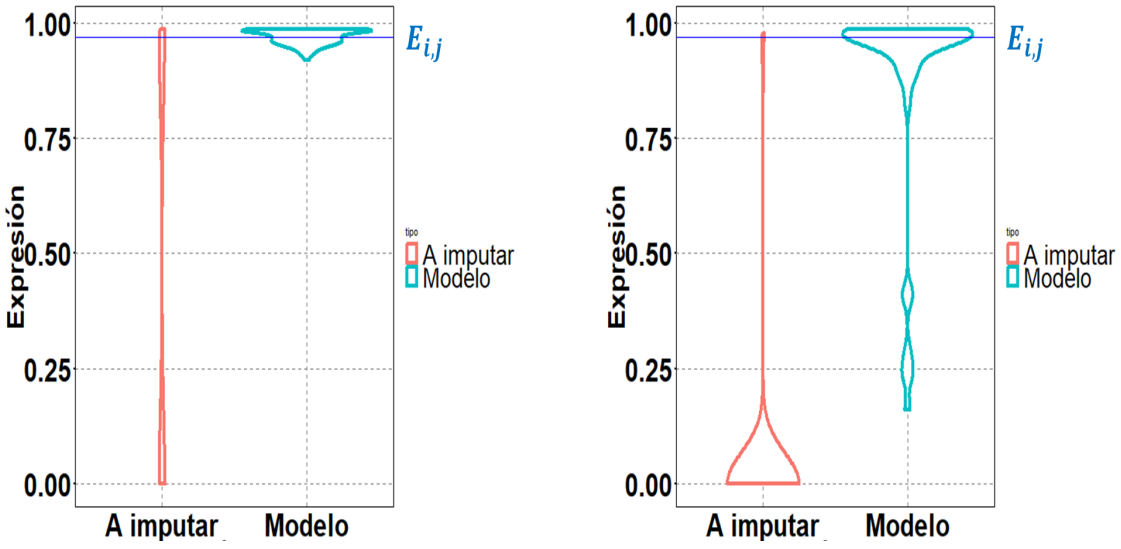
Una vez cuantificada la similitud entre células, se imputa la expresión de un gen en la célula proyectando la expresión de este en las vecinas pesando por β. La probabilidad de que dada una expresión a imputar E_{ij} (nula), la expresión del gen en una vecina (k) de la célula j sea también nula es la probabilidad conjunta de los sucesos $P(E_{i,j} = 0 \cap E_{i,k} = 0) = P(E_{i,j} = 0)P(E_{i,k} = 0) = p^2$, donde p es la fracción de elementos nulos para el gen i-ésimo en las células muestreadas.

Para el caso en que simulamos un 95 % de expresiones nulas, la probabilidad p depende de la expresión media del gen de forma que $p \in [0.8, 0.95]$. En consecuencia, la probabilidad conjunta $P(E_{i,j} = 0 \cap E_{i,j+1} = 0) \in [0.64, 0.90]$. Por otro lado, para el caso de menor cantidad de elementos nulos, la probabilidad es independiente del gen ($p = 0.4$) de forma que $P(E_{i,j} = 0 \cap E_{i,k} = 0) = 0.16$. Es notoria la alta probabilidad conjunta correspondiente al simular una

tasa alta de dropout en comparación con una tasa baja. En consecuencia, la probabilidad de que se utilice otro evento dropout para la imputación en una matriz dominada por los *dropouts* es muy alta.

Como vimos en Fig.5.14, para una tasa alta de *dropouts* simulados, la hetergeneidad de coeficientes de similaridad inter-celulares β es alta. Por lo tanto, para estos casos es de esperar que los resultados de la imputación sean fuertemente dependientes del nivel de expresión de pocas células utilizadas para imputar. Con un número limitado de vecinos, y un nivel tan alto para la probabilidad conjunta [0.64,0.90] de encontrar valores nulos entre pares de células, la probabilidad de que la imputación sea un valor bajo y por consiguiente, muy distinto al modelado, es alta. Por el contrario, para una tasa de *dropouts* simulados baja, al ser una suma más uniforme de expresiones y la probabilidad de que sean dropout más baja, la probabilidad de que el valor imputado sea bajo y por tanto, mal imputado, disminuye.

Para ver en detalle cuáles son las expresiones utilizadas para imputar (asociadas a $\beta \neq 0$) la expresión del gen en la célula con los que venímos trabajando, graficamos su distribución en gráficos de violín para las dos fracciones de *dropouts* simulados. Además es interesante comparar a estas expresiones con las que corresponden en el modelo. Se muestran ambas distribuciones en mismos gráficos en Fig.5.15 y se indica como E_{ij} la expresión en el modelo correspondiente a la que estamos imputando.



(a) 0.49 de la matriz son elementos nulos.

(b) 0.95 de la matriz son elementos nulos.

FIGURA 5.15: Distribución de expresiones en las células vecinas similares ($\beta \neq 0$) del gen a imputar. Se muestra la distribución en la matriz a imputar (salmón) y en la del modelo (celeste). Se señala en la línea horizontal azul el valor del modelo $E_{i,j}$ del gen en esa célula donde se imputa.

Como se observa en el panel A, para el caso de menor cantidad de expresiones nulas, la distribución en las expresiones utilizadas para imputar toma valores concentrados en el cero y por encima del valor en el modelo ($E_{i,j}$). Nótese que las expresiones intermedias tienen asociada densidad nula que se representa como una recta. Por otro lado, en el modelo la densidad se concentra alrededor de $E_{i,j}$ indicando que se seleccionaron de manera correcta las células vecinas.

Por el contrario, vemos en el panel B que al aumentar los elementos nulos, la distribución en la matriz a imputar se concentra para valores bajos y nulos. En consecuencia, el aporte que hace cada célula a la imputación es más o menos uniforme porque β es también acotado en valores bajos (como discutimos en Fig.5.14). En consecuencia, el valor imputado es bajo en comparación con el que se corresponde en el modelo. Por otro lado, vemos que para el modelo ahora la distribución toma valores más bajos de expresión que antes, con menos *dropouts*. Esto estaría indicando que las vecinades de células no fueron bien seleccionadas.

Estas vecindades vienen dadas por la partición en comunidades de la *matriz de expresión*. En particular nosotros utilizamos el algoritmo espectral que recibe como parámetro la cantidad de comunidades (ver Ápendice A). Como estamos trabajando con una matriz de un modelo construido por nosotros, esperamos que sean seis las comunidades de células que se correspoden con la dinámica que modelamos. Imponemos este valor al particionar la matriz del modelo y también los datos en que simulamos los droputs.

Al modificar la distribución de los elementos nulos sobre el modelo, se contruye una matriz diferente que es particionada. En consecuencia, las comunidades de cada participación pueden, en principio, diferir. Es así como las vecindades de cada célula pueden variar y por eso, las expresiones de los vecinos en el modelo son distintas que en la matriz a imputar como vimos para un gen en particular en Fig.5.15.

Un paso fundamental en el modelado estadístico de la ocurrencia de *dropouts* incluye reconocer comunidades de células similares a partir del perfil de expresión de sus genes. Se espera que si la partición es buena, los genes tengan asociado un coeficiente de variación (CV) intra-comunidad menor que en la totalidad de las células. Es decir que la variación de la expresión entre células similares (agrupadas en comunidades) sea menor que la variación del gen sobre toda la población de células.

Se observa en Tabla.5.2 la mediana de las distribuciones de los CV de las expresiones de los genes en cada comunidad y también en la población total de células siendo estos iguales para la partición sobre la matriz del modelo sin *dropouts* y sobre la matriz a imputar con una fracción de 0.49 de elementos nulos. Se observa que en todas las comunidades la mediana del CV es menor que en el que se corresponde a la población total de células. En consecuencia, afirmamos que la partición en comunidades, con y sin *dropouts*, es buena en el sentido de la variabilidad en comunidades de la expresión de genes.

Comunidad	**CV modelo/a imputar**
1	4.66
2	4.76
3	4.64
4	4.64
5	4.79
6	4.8
Total	4.86

TABLA 5.2: Valores de la mediana de la distribución del coeficiente de variaciación (CV) de los genes en las células de cada respectiva comunidad. Se obtuvieron los mismos valores para la partición sobre la matriz del modelo y la matriz a imputar (con 0.49 de elementos nulos). Se compara el valor que se obtuvo para cada comunidad con el que se obtiene sobre la totalidad de células.

Nos planteamos el interrogante de qué tan distinta será la partición en comunidades para los perfiles a imputar, al ir aumentando la fracción de elemento nulo, en comparación con la que se corresponde con el modelo. Una forma de cuantificar qué tan similares son dos particiones es calculando la información mutua media (MI).

Para particiones cuyos vectores de pertenencia de los nodos son X e Y siendo sus elementos las comunidades a las que pertenecen los nodos, la MI se define según ec.5.9 donde n es la cantidad de comunidades en la partición X y m la correspondiente a la Y. La probabilidad $P(x_i, y_j)$ es la probabilida conjunta de los elementos de los vectores X e Y que es calculada a partir del elemento n_{ij} de la matriz de concurrencia que cuantifica la cantidad de nodos que en la partición Y se encuentran en la comunidad j se corresponden a la comunidad i de la X tal que $P(x_i, y_j) = \frac{n_{ij}}{N}$ (N es la cantidad de nodos total).

$$I(X,Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i, y_j) log \frac{P(x_i, y_j)}{P(x_i)P(y_j)} \tag{5.9}$$

Graficamos la información mutua media entre la partición del modelo y la correspondiente al simular *dropouts* en función de la fracción total de estos eventos en Fig.5.16. También calculamos la MI que se correponde a un modelo nulo que preserva la cantidad de nodos por comunidad cuyo valor es (0.011 ± 0.003). Aunque vemos que la MI correspondiente a todas las fracciones supera a la del modelo nulo, se reconce un salto en la curva tal que para fracciones por encima de 0.5 disminuye significativamente la información mutua entre el modelo y los datos con dropout, alcanzando el mínimo para 0.9.

Esta disminución indica que al simular expresiones nulas con una tasa mayor a dicho valor, las comunidades halladas comparten una baja información sobre las comunidades que fueron modeladas. Por consiguiente, afirmamos que para fracciones por encima de la cota de 0.5, se pierde la similitud entre las células. Aunque la MI no es menor que la del modelo nulo, esta pérdida de información hace que las expresiones utilizadas en la imputación no sean lo suficientemente similares para obtener una imputación consistente con el modelo.

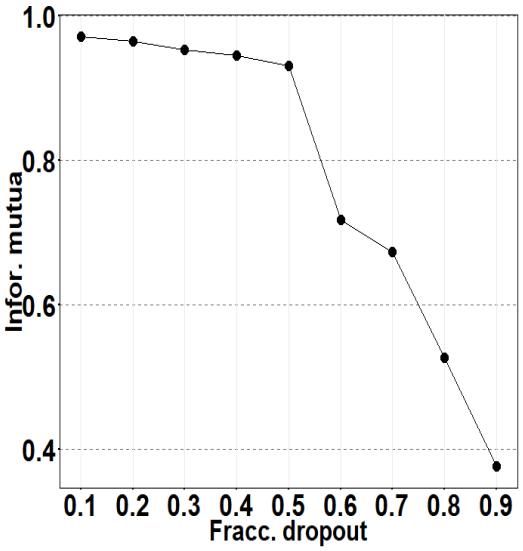


FIGURA 5.16: Información mutua entre la partición del modelo y la de la matriz con *dropouts* en función de la fracción total de estos eventos.

Correlaciones

Otra forma de cuantificar qué tan bien fueron imputadas las expresiones es calculando la correlación entre expresiones de genes pertenecientes al mismo grupo dinámico. Esperamos que esta correlación sea alta en el modelo pues son trayectorias que siguen la misma dependencia en el tiempo de simulación y que la inclusión de *dropouts* la deteriore. Sin embargo, una buena imputación de expresiones debería poder recomponer, aunque sea parcialmente, los valores modelados. Para ver esto calculamos el p-valor asociado a la correlación de Pearson entre todo par de genes de un mismo grupo dinámico (Ver Secc.3.3).

En particular calculamos $-log(pvalor)$ y calculamos la mediana correspondiente a la distribución para cada grupo. Comparamos las correlaciones en el modelo (sin dropout) con las correspondientes al simular dropout y también con las que se obtienen una vez imputada la matriz en Tabla.5.3. Nótese que calculamos $-log(pvalor)$ de forma que los p valores nulos son eliminados para el cálculo.

	Modelo	49 %		95 %	
	No dropout	Con dropout	Imputada	Con dropout	Imputada
Crec A	3.12	1.51	2.18	0.24	0.82
Crec C	256.54	13.77	166.29	0.10	0.61
Decr B	17.45	2.09	2.80	0.34	0.75
Decr D	NA (p=0)	1.81	85.50	0.32	0.73

TABLA 5.3: Medianas de la distribución del $-log(pvalor)$ de correlación entre las expresiones de los genes modelados en cada grupo dinámico. Se muestran valores correspondientes al modelo y al simular una tasa de 49 % y de 90 % de dropout indicando los valores en la matriz con dropout (a imputar) y en la imputada.

Observamos que en el modelo la mediana del $-log(pvalor)$ asociado depende del grupo y en particular, aquellos que presentan la mayor correlación son el Crec C y el Decr B. Al simular *dropouts*, en ambas fracciones, encontramos que la correlación disminuye pues $-log(pvalor)$ decrece significativamente. Esto indica que los *dropouts* difuminan también la similitud entre las expresiones de los genes. Como esperabamos, al ser mayor la cantidad de elementos nulos, la correlación es menor siendo esta de un orden de magnitud menor para el caso de 95 % de dropout.

Al imputar la matriz con una cantidad baja de *dropouts* (por debajo de la cota de 50 % previamente hallada) se consigue aumentar los valores $-log(pvalor)$ de forma tal que la correlación modelada es reestablecida. También vemos que, en particular, para los grupos Crec D y Decr D se obtienen los p valores más bajos en correspondencia a lo que vimos en el modelo. Por otro lado, al imputar con una tasa alta de dropout, aunque la correlación se incrementa, se encuentra por debajo de 0.1 y por debajo de los correspondientes en el modelo sin dropout.

En consecuencia, afirmamos que al simular una fracción alta de *dropouts*, por encima de la cota que encontramos en el análisis de las comunidades de células (50 %), la imputación no logra corregir los valores de correlación entre las expresiones de genes que siguen la misma dinámica. Por el contrario, para un nivel menor encontramos que sí se reestablecen estos valores que fueron degrados por el dropout, de forma que la imputación es compatible con el modelo.

Concluimos que para tasa de dropout por encima de 50 % la imputación no permite reconocer la similitud entre células ni tampoco recuperar los valores de correlación entre los genes que evolucionan según las mismas funciones parametrizadas por el tiempo.

A modo de resumen, en este capítulo modelamos los datos recopilados de un experimento scRNASeq. Para ello, simulamos los *dropouts* a partir de una distribucion de Bernoulli sobre el modelo sintético II (explicado en Cap.3) y estudiamos la técnica de imputación Scimpute. Para ello, desarrollamos observables para cuantificar la efectividad de la técnica. También, estudiamos la dependencia de la imputación con dos parámetros que caracterizan a una matriz de cuentas

de un experimento scRNASeq: dantidad de células y fracción de *dropouts*. Concluimos que una mayor granularidad en el muestreo, al incrementar de 125 a 1250 células, aumenta significativamente la fracción de células modelables. Sin embargo, las imputaciones no se aproximan a los valores en el modelo pues D_i^2 aumenta en un orden de magnitud.

Con respecto a la fracción de *dropouts*, vimos que al disminuir la fracción tomando $p = 0.4$, en vez de $p = e^{-0,1<g>}$, la imputación mejora significativamente. Se reconstruye una evolución en el tiempo de simulación por gen heterogénea y también se reconocen poblaciones de genes y su respectiva evolución colectiva. Por otro lado, nos propusimos entender en profundidad por qué la imputación no es buena para el caso de alta fracción de *dropouts*. Estudiamos cómo es el aprendizaje de los coeficientes β que cuantifican la similitud y cómo es la distribución de las expresiones. Sin embargo, un primer paso para la creación de las vecinades de células a las cuales se les asignará los coeficientes β es la partición en comunidades.

Por ello, estudiamos cómo es la partición en las matrices a imputar y la comparamos con la que se corresponde con el modelo. Calculamos la mediana del CV de las expresiones de los genes. Afirmamos que en el modelo y al simular una tasa baja, los CV de las comunidades son menores que el global indicando una partición compatible con el modelado realizado. También comparamos a qué comunidad pertenece cada célula en la partición del modelo y en la de las matrices a imputar. Afirmamos que al modelar *dropouts* se difuminan las vecindades modeladas pues las comunidades del modelo se ven divididas en más comunidades de forma que se agrupan células que no fueron modeladas como similares.

Para hallar una cota de la fracción de dropout tal que por encima de esta no se particiona en concordancia con lo modelado, calculamos la información mutua media entre la partición del modelo y para las matrices a imputar en función de la fracción de dropout. Afirmamos que para valores por encima de 0.5 la información mutua disminuye abruptamente tal que se pierde la similitud simulada entre células. Y por último, estudiamos la correlación de Pearson entre grupos de genes modelados por la misma dinámica. Afirmamos que mientras que para una tasa baja, la correlación es reestituida por la imputación, para una tasa alta esto no es así.

Capítulo 6

Conclusión

Los experimentos de secuenciación de célula única (scRNAseq) introducen un nivel de detalle que trae múltiples desafíos relacionados con el volumen y complejidad de los datos relevados. Para enfrentarlos, asumimos como hipótesis de trabajo que los genes no se expresan independientemente, sino que su expresión se coordina de forma que el estado celular relevante puede ser descripto sobre una variedad de células de mucha menor dimensión, la VBR.

En un mismo experimento scRNASeq se miden poblaciones célulares que podrían incluir muestras en distintas étapas de desarrollo. Esto permite detectar procesos y evoluciones temporales de manera indirecta que han permitido ganar conocimiento sobre aspectos moleculares de la biología del desarrollo de procesos variados.

La forma de reconocer diferentes evoluciones temporales es reconstruyendo trayectorias que ocurren a lo largo del *pseudo-tiempo* sobre la VBR. En esta tesis modelamos a esta variedad como un grafo de muestras scRNASeq cuyas aristas se construyen por la similitud euclidea de los transcriptomas medidos. La reducción de dimensionalidad y el posterior reconocimiento de caminos desde un estadío inicial sobre la VBR depende fuertemente de los genes seleccionados como de interés.

En el Cap.2 estudiamos la técnica de Slicer sobre un modelo sintético, al simular ciertas evoluciones: para grupos de genes que siguen una dinámica, parametrizando por el tiempo más un ruido gaussiano, y para genes aislados que no siguen una dinámica, asignando un valor aleatorio.

Dado que hay genes que son expresados por más de un grupo dinámico, en el Cap.3 desarrollamos un modelo más robusto en el cual construimos los grupos de genes a partir de un conocimiento de en qué procesos biológicos participan según la ontología GO.

Sobre ambos modelos realizamos la selección de genes según el criterio de Slicer y cuantificamos la efectividad de esta selección según la idea de una reconstrucción de un orden de las células en

el grafo que funciona como una aproximación a la VBR. También determinamos la efectividad de la selección respecto a si filtra a aquellos genes que no presentan una dinámica.

Por otro lado, en el Cap.4 trabajamos con datos experimentales recopilados en el marco de un proyecto de Linnarson Lab (Hochgerner et al., 2018) orientado a la comprensión del desarrollo distintos tipos celulares en el giro dentado del hipocampo en ratones. Realizamos un análisis de calidad para seleccionar células y genes sobre los realizamos la selección por el criterio de Slicer.

Encontramos que por la técnica de Slicer se selecciona una muy baja fracción de genes (4 %). Esto, en principio, implica una pérdida muy grande de información. En consecuencia, introducimos un criterio estadístico, resultante de comparar la variabilidad local de la expresión de los genes sobre el grafo con la de un modelo nulo en el que se enlazan los nodos de forma aleatoria. Por otro lado, vimos que la tasa de *dropouts* en estos datos experimentales es extremadamente alta (90 %). Por esta razón, desarrollamos los criterios de Slicer y del modelo nulo sin considerar *dropouts* para el cálculo de las varianzas locales y globales con la finalidad de disminuir los efectos de ésta limitación técnica.

Sin embargo, concluimos que ninguno de estos criterios de selección resulta lo suficientemente confiable. Esto nos llevó en el Cap.5 a estudiar una técnica específica de imputación de eventos *dropouts* en experimentos scRNASeq llamada Scimpute. Estudiamos su funcionamiento y limitaciones. Para ello, aplicamos el algoritmo sobre el modelo de la *matriz de expresión* que construimos en Cap.3 al que le simulamos estos eventos.

Medimos la calidad de la imputación con respecto a dos parámetros que caracterizan a la *matriz de expresión*: la cantidad de células y la tasa de *dropouts*. Con respecto al primer parámetro observamos que se requiere una cantidad mínima de células para obtener una fracción razonable de células cuyas expresiones pueden parametrizarse por un modelo de probabilidad mixta que modela los eventos *dropouts* y los que no lo son.

En lo que respecta a la tasa de eventos *dropouts*, observamos que para un valor alto (95 %) no se logra reconstruir la evolución temporal ni tampoco se pueden reconocer los distintos grupos de dinámica genética. Por el contrario, para una tasa baja (49 %) observamos que resultó posible. Afirmamos que la imposibilidad de imputar correctamente, para una fracción alta de *dropouts*, se debe a la probabilidad alta de que un evento dropout tenga como vecino a otro evento de este tipo siendo esta expresión utilizada para la imputación.

Por otro lado, también vimos que la presencia alta de dropout hace que la partición en comunidades no se corresponda con la similitud entre los perfiles de las células que modelamos. En consecuencia, las células vecinas, cuyas expresiones se utilizan para imputar, no son semejantes y la imputación resulta siendo mala. Encontramos que 0.5 de fracción de *dropouts* es un valor cota tal que por encima de este, no es posible reconocer las comunidades modeladas y por tanto, el algoritmo Scimpute no funciona de forma satisfactoria

Apéndice A

Apéndice

La técnica de partición en comunidades espectral (Ng et al., 2002) es una técnica para encontrar comunidades entre N vectores de dimensión M $E = \{E_1, E_2, ..., E_N\}$ en el cual se debe especificar la cantidad de comunidades K en que se quiere particionar los datos. En primer lugar se calcula la matriz de afinidad (o similaridad) A cuyo elemento A_{ij} cuantifica la similitud entre E_i y E_j por medio de una función kernel. En particular, nosotros usamos la radial basis function kernel que computa la similitud de la siguiente forma:

$$K(x, y) = e^{-\gamma||x-y||^2} \tag{A.1}$$

donde $||x - y||$ es la distancia euclideana y γ un parámetro libre.

A partir de la matriz de afinidad A, se calcula el laplaciano L como $L = D - A$ donde D es la matriz diagonal $D_{ii} = \sum_j A_{ij}$. Se calculan los K (cantidad de comunidades que pasamos como parámetro) autovectores de L, $U = \{u_1, u_2, ..., u_K\}$. Se construye la matriz Y cuyas filas son los elementos de U normalizados.

Esta matriz Y de tamaño $K \times M$ es la que particionaremos en comunidades. La idea intuitiva detrás de esto es que esta representación mejora las propiedades de las comunidades de los datos (el conjunto E) y así podremos reconocer comunidades de forma más certera.

La matriz Y es particionada en comunidades por el algoritmo k-means (Jain, 2010) y encontramos las K comunidades $\{C_1, C_2, ..., C_K\}$ de Y. Una vez que ya hemos encontrado las K comunidades, nos resta asignar a cada elementos de E_i a la la comunidad C_j. Decimos que E_i pertence a la comunidad C_j si $\{Y_{i,j}, \forall j\}$ (fila i de la matriz Y) pertenece a C_j.

El algoritmo de partición k-means particiona al set $\{Y_1, Y_2, ..., Y_M\}$ en comunidades que minimicen la suma cuadrados inter-comunidad (i.e varianza) según ec.A.2 donde μ_i es el valor medio de los Y_i pertenecientes a la comunidad C_i

$$\arg\min_{\mathbf{C}} \sum_{i=1}^{k} \sum_{\mathbf{Y} \in C_i} \|\mathbf{Y} - \boldsymbol{\mu}_i\|^2 \tag{A.2}$$

Bibliografía

Campbell, J. N., E. Z. Macosko, H. Fenselau, T. H. Pers, A. Lyubetskaya, D. Tenen, M. Goldman, A. M. Verstegen, J. M. Resch, S. A. McCarroll, et al. (2017). A molecular census of arcuate hypothalamus and median eminence cell types. *Nature neuroscience*@(3), 484.

Cannoodt, R., W. Saelens, D. Sichien, S. Tavernier, S. Janssens, M. Guilliams, B. N. Lambrecht, K. De Preter, and Y. Saeys (2016). Scorpius improves trajectory inference and identifies novel modules in dendritic cell development. *bioRxiv*, 079509.

Chambers, J. M., T. J. Hastie, et al. (1992). *Statistical models in S*, Volume 251. Wadsworth & Brooks/Cole Advanced Books & Software Pacific Grove, CA.

Dal Molin, A. and B. Di Camillo (2018). How to design a single-cell rna-sequencing experiment: pitfalls, challenges and perspectives. *Briefings in bioinformatics*.

Dijk, D. v., J. Nainys, R. Sharma, P. Kaithail, A. J. Carr, K. R. Moon, L. Mazutis, G. Wolf, S. Krishnaswamy, and D. Pe'er (2017). Magic: A diffusion-based imputation method reveals gene-gene interactions in single-cell rna-sequencing data. *bioRxiv*.

Eberwine, J., J.-Y. Sul, T. Bartfai, and J. Kim (2014). The promise of single-cell sequencing. *Nature methods*@(1), 25.

Grün, D., L. Kester, and A. van Oudenaarden (2014). Validation of noise models for single-cell transcriptomics. *Nature Methods 11, 637–640.*

Hochgerner, H., A. Zeisel, P. Lönnerberg, and S. Linnarsson (2018). Conserved properties of dentate gyrus neurogenesis across postnatal development revealed by single-cell rna sequencing. Nature neuroscience*@(2), 290.*

Huang, M., J. Wang, E. Torre, H. Dueck, S. Shaffer, R. Bonasio, J. Murray, A. Raj, M. Li, and N. R. Zhang (2017). Gene expression recovery for single cell rna sequencing.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des alpes et des jura. Bull Soc Vaudoise Sci Nat 37, 547–579.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern recognition letters*@(8), 651–666.

Jaitin, D. A., E. Kenigsberg, H. Keren-Shaul, N. Elefant, F. Paul, I. Zaretsky, A. Mildner, N. Cohen, S. Jung, A. Tanay, et al. (2014). Massively parallel single-cell rna-seq for marker-free decomposition of tissues into cell types. *Science*@(6172), 776–779.

K., P. (1901). Liii on lines and planes of closest fit to systems of points in space. *Philos Mag Series@*(2), 559–572.

Kolodziejczyk, A. A., J. K. Kim, J. C. Tsang, T. Ilicic, J. Henriksson, K. N. Natarajan, A. C. Tuck, X. Gao, M. Bühler, P. Liu, et al. (2015). Single cell rna-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell stem cell@*(4), 471–485.

Li, W. V. and J. Jessica Li (2018, 12). An accurate and robust imputation method scimpute for single-cell rna-seq data. *Nature Communications 9.*

Macosko, E. Z., A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell*@(5), 1202–1214.*

Nawy, T. (2013). Single-cell sequencing. Nature methods*@(1), 18.*

Ng, A. Y., M. I. Jordan, and Y. Weiss (2001). On spectral clustering: Analysis and an algorithm. In ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS*, pp. 849–856. MIT Press.*

Ng, A. Y., M. I. Jordan, and Y. Weiss (2002). On spectral clustering: Analysis and an algorithm. In Advances in neural information processing systems*, pp. 849–856.*

Pierson, E. and C. Yau (2015). Zifa: Dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome biology*@(1), 241.*

Rizvi, A. H., P. G. Camara, E. K. Kandror, T. J. Roberts, I. Schieren, T. Maniatis, and R. Rabadan (2017). Single-cell topological rna-seq analysis reveals insights into cellular differentiation and development. Nature biotechnology*@(6), 551.*

Rostom, R., V. Svensson, S. A. Teichmann, and G. Kar (2017). Computational approaches for interpreting scrna-seq data. FEBS letters*@(15), 2213–2225.*

Science, T. D. (2018). Adversarial learning with local coordinate coding.

Tang, F., C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, et al. (2009). mrna-seq whole-transcriptome analysis of a single cell. Nature methods*@(5), 377.*

Van Dijk, E. L., H. Auger, Y. Jaszczyszyn, and C. Thermes (2014). Ten years of next-generation sequencing technology. Trends in genetics*@(9), 418–426.*

Welch, J. D., A. J. Hartemink, and J. F. Prins (2016). Slicer: inferring branched, nonlinear cellular trajectories from single cell rna-seq data. Genome biology*@(1), 106.*

West, D. B. et al. (1996). Introduction to graph theory*, Volume 2. Prentice hall Upper Saddle River, NJ.*