Máster en Ingeniería y Ciencia de Datos

Trabajo Fin de Máster

Desarrollo de un sistema basado en aprendizaje automático para el reconocimiento de patrones de expresión asociados a mutaciones a partir de secuenciación de ARNm en células individuales

---

Autor

Víctor Rodríguez Bouza

Tutores

Xose Antón Suárez Puente

Rafael Pastor Vargas

Curso 2021–2022 - Convocatoria extraordinaria

Universidad de Oviedo

UNED

Resumen

En este trabajo se ha estudiado la posibilidad de construir modelos de aprendizaje automático que identifiquen la presencia de mutaciones en secuenciaciones individuales de ARN mensajero (*single cell RNA-sequencing*) utilizando datos de pacientes de leucemia linfática crónica (LLC). Para ello, se han creado simulaciones de Montecarlo que permitieron replicar secuenciaciones de *scRNA-seq*. Los resultados confirman que existen diferencias que permiten identificar estas mutaciones en secuenciaciones individuales de ARNm. Sin embargo, debido al reducido número de datos no es posible obtener un modelo final estable.

Palabras clave — aprendizaje automático, ARNm, mutaciones, cáncer

Índice general

Índice de figuras

Abreviaturas, siglas y acrónimos

A Adenina. 3, 4, 10, 12

ADN Ácido desoxirribonucleico. 3–7, 9–12

ADNc Ácido desoxirribonucleico copia. 9–11

ARN Ácido ribonucleico. 1, 3, 4, 6–9, 56

ARNm Ácido ribonucleico mensajero. 1, 3–5, 9–12, 16, 18, 19, 57

AUC *Area under the curve*. 22, 23, 29, 38, 47, 51, 53, 54, 56, 57

C Citosina. 3, 4

DGE *Differential gene expression*. 19

G Guanina. 3, 4

GLM *Generalised linear model*. 19, 57

IGHV *Inmunoglobulin heavy chain variable region*. 14, 15

LLC Leucemia linfótica crónica. 1, 14–16, 20, 57

LOF *Local outlier factor*. 23

MBL *Monoclonal B lymphocytosis*. 14

NGS *Next generation sequencing*. 6–9

OS *Overall survival*. 15

PCA *Principal component analysis*. 22, 23, 28, 29, 38, 51, 54

RNA-seq. RNA-*sequencing*. 10, 11, 16, 20

ROC *Receiver operating characteristic*. 22, 23, 29, 47, 51

scRNA-seq. *single cell* RNA-*sequencing*. 11, 54

SVM *Support vector machine*. 23, 53

T Timina. 3, 4, 10

TTFT *Time to first treatment*. 15

U Uracilo. 4

Introducción

El objetivo de este Trabajo Fin de Máster es, en primer lugar, explorar la posibilidad de desarrollar un modelo de aprendizaje automático para clasificar patrones de expresión en datos de secuenciación de ARN mensajero de célula única (*single cell RNA sequencing*) asociados a mutaciones concretas en cáncer. Con este fin, se tendrán datos procedentes de secuenciación común (*bulk RNA sequencing*), a partir de los cuales se deberá extrapolar la situación con secuenciación individual. Este texto constituye una memoria de los esfuerzos dedicados a estos objetivos.

La estructura de este texto es la siguiente. El primer capítulo (precedido por un resumen del trabajo al completo y esta introducción) muestra una aproximación a la leucemia linfática crónica (LLC), que es la enfermedad que da lugar a los tumores de cuyas muestras obtenemos los datos de secuenciación para el trabajo. También se ofrece un resumen sobre la secuenciación de ARNm en general, y en particular sobre la secuenciación individual de ARNm.

El segundo capítulo está dedicado a la descripción de la metodología seguida y a la explicación del desarrollo del trabajo para la consecución de sus objetivos. La descripción de los datos se detalla en su primera sección, a lo que sigue la explicación sobre la estrategia seguida en el mismo.

El siguiente capítulo recoge todos los resultados y discusión sobre los modelos que se entrenaron y los distintos resultados obtenidos. Y, finalmente, el último capítulo contiene las conclusiones de todo el trabajo, a las que sigue la bibliografía.

1 Contexto biomédico

En este capítulo nos introduciremos en los prolegómenos biológicos y médicos necesarios para poder comprender el trabajo detallado en los epígrafes siguientes de este documento. Comenzaremos viendo los elementos biológicos mínimos necesarios para poder comprender la secuenciación de ARN mensajero de forma superficial, y en particular, la técnica en torno a la cual este TFM gira, que es la secuenciación individual de ARNm. Después, haremos una visión somera a la enfermedad que afecta a los datos que usaremos, la LLC: un tipo de tumor hematológico (cáncer de la sangre).

1.1. La información genética y su transcripción

(Lodish, 2016) La molécula bioquímica más conocida por el público general probablemente sea el ácido desoxirribonucleico, comúnmente conocido por las siglas ADN (o también «DNA», del inglés *deoxyribonucleic acid*). Esta molécula tiene como función esencial ser la guardiana del conocimiento para el funcionamiento de la célula, y permitir a las células transmitir esta información a su descendencia.

Químicamente, la molécula es un polímero (es decir, una unión de moléculas más simples) de nucleótidos, que son moléculas comparativamente mucho más pequeñas conformadas por un grupo fosfato (i.e. una molécula derivada del ácido fosfórico), un azúcar (desoxirribosa) y una base nitrogenada. Las bases nitrogenadas pueden ser cuatro en el ADN: adenina (usualmente representamos a esta base y a su nucleótido con la letra A), timina (T), citosina (C) y guanina (G). La estructura del ADN en los organismos vivos, propuesta por Watson y Crick en 1953, es de dos de estos polímeros, hechos de secuencias de nucleótidos (e.g. ...AATGGCA...), con estos

apuntando hacia el interior, y disponiéndose como pares A-T y C-G. Las bases opuestas (A con T y C con G) se unen entre sí a través de puentes de hidrógeno, de forma que una cadena es complementaria de la otra, como muestra la figura 1.1.

El ADN se organiza en cromosomas: grandes estructuras que constan de este polímero muy enrollado en torno a sí mismo y a múltiples proteínas. En el caso de la especie humana, poseemos 23 pares de cromosomas que suman un total de 3200 millones de pares de bases de ADN con del orden de 20.000 genes.

Figura 1.1: (Lodish, 2016) Representación de la estructura de la molécula de ADN.

La secuencia en la que se ordenan las bases del ADN es lo que constituye la información genética. Llamamos gen a un segmento de esa secuencia que contiene la información que codifica la secuencia de una proteína, y por lo tanto contribuye al funcionamiento celular. Las proteínas son polímeros, como el ADN, pero formados por una secuencia de moléculas llamadas aminoácidos. Las proteínas son muy diversas y realizan múltiples funciones, determinadas por su secuencia y la estructura tridimensional que adopten. En general, se encargan de «llevar a cabo» las funciones codificadas en la información guardada en el ADN.

Aunque el ADN codifica la información para la síntesis de proteínas, el ADN no participa directamente en el proceso de síntesis de proteínas, sino que se requiere la participación de una molécula intermediaria que lleve el mensaje: el ARN mensajero (ARNm). El ARNm se sintetiza usando como molde el ADN mediante el proceso conocido como transcripción.. Esto se hace con ayuda de una proteína (llamada ARN polimerasa) que transcribe el orden de nucleótidos del ADN (tomándolo como molde) al orden de nucleótidos del ARN: una diferencia más es que el contenido de nucleótidos no es el mismo. En el ADN teníamos bases A, T, G y C, pero en el caso del ARN no existe timina (T), sino uracilo (U), que ocupa su posición. Otra diferencia más con el ADN es que el ARN solo tiene una cadena, no dos entrelazadas como en el caso del ADN.

En el caso de las células eucariotas (como las de los seres humanos), este ARN inicial se transforma en otro llamado mensajero (ARNm, o mRNA), más corto, que se encarga de llevar la información genética desde el núcleo hasta el citoplasma, donde se encuentran unas comple-

jas máquinas moleculares llamadas ribosomas. Es ahí donde se lleva a cabo la traducción de este mensaje codificado por 4 bases distintas a la secuencia de una proteína codificada por 20 aminoácidos diferentes. La traducción se hace a través del llamado código genético: las reglas que codifican los aminoácidos en función a los codones, que son grupos de tres bases nitrogenadas del ARNm que codifican un aminoácido. Por ejemplo, el codón CGA codifica la arginina (un aminoácido). En la figura 1.2 se puede ver un esquema de ambos procesos (transcripción y traducción).

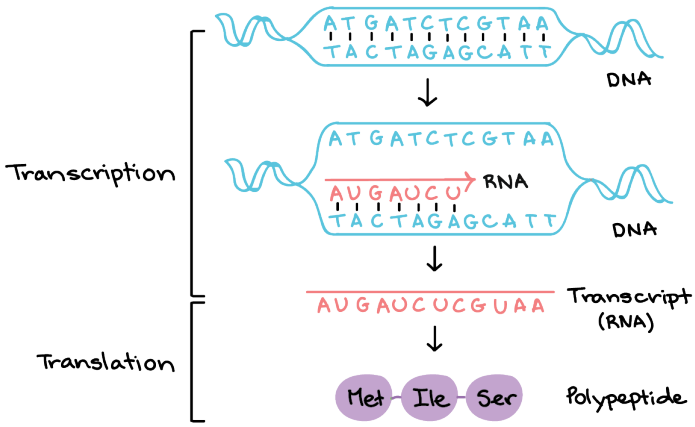


Figura 1.2: (*Khan Academy*, s.f.) Esquema de los procesos de transcripción y traducción de la información genética.

La molécula de ADN tiene una relevancia capital en el funcionamiento de las células, y por ello, esa información se protege. Las células poseen sistemas de corrección de errores en la célula que puedan darse, por ejemplo, a la hora de replicar esa información para la reproducción celular. Una alteración en la cadena de ADN se llama mutación, y sus efectos pueden tener importantes afectaciones al desarrollo celular. Aquellas que se dan en un organismo multicelular (como los seres humanos) y no se propagan a la descendencia se denominan mutaciones somáticas, para diferenciarlas de las que se dan en la línea germinal.

Los orígenes de las mutaciones también son diversos. La exposición a mutágenos (agentes de cualquier tipo capaces de modificar la secuencia de ADN) como radiación ionizante que puede afectar directamente al ADN interaccionando con él es uno de ellos, pero también pueden darse mutaciones espontáneas, como errores a la hora de replicar el ADN.

Finalmente, sus consecuencias también son múltiples: desde modificar el gen sobre el que se dan para otorgarle una nueva función (que puede constituir un nuevo avance en la evolución de la especie), o por el contrario si afectan a genes clave en el funcionamiento celular pueden tener

consecuencias muy negativas, pudiendo ser letales. Debido a esto, el conocimiento de la secuencia genética del ADN, así como el estudio de las posibles mutaciones tiene un alto interés, puesto que permite comprender cómo se desarrolla la transferencia y modificación de la información que guarda esta molécula de especial importancia para la vida. Y, desde un punto de vista más práctico, nos permite comprender cómo se generan enfermedades genéticas, como por ejemplo el cáncer.

1.2. La secuenciación genómica

1.2.1. Secuenciación de ADN

(Shendure y col., 2017) La obtención de la secuencia (i.e. secuenciación) de las bases nitrogenadas en el ADN tuvo sus primeros pasos a mediados del siglo XX. Fred Sanger se dedicó al esfuerzo de conocer la secuencia primaria que guarda esa información, y gracias a sus esfuerzos, realizó a principio de los años 50 la primera secuenciación de una proteína: la insulina. El siguiente gran paso se dio con la secuenciación de la primera cadena de ARN, en los años 60: cinco personas trabajando durante tres años fueron necesarias para secuenciar una cadena de 76 nucleótidos.

Finalmente, las primeras secuenciaciones de ADN se dieron en torno a 1970, aunque solo se lograron descifrar cadenas de pocas bases. Por ejemplo, en 1973 se lograron secuenciar 24 bases de ADN a la rapidez de una base por mes de media (Gilbert & Maxam, 1973). Lo que revolucionó el campo fue el desarrollo en torno a 1976 de nuevos procedimientos experimentales que permitían secuenciar cientos de bases en horas. El desarrollo científico-tecnológico continuó y en torno a 1987 se conseguían secuenciar mil bases por día. El avance en los métodos permitió nuevas investigaciones, como el ambicioso Proyecto Genoma Humano, que tenía como objetivo determinar la secuencia de ADN de nuestra especie, empresa culminada en el año 2003.

La investigación en nuevas técnicas de secuenciación genómica dio sus frutos condensados en lo que se denomina secuenciación de ADN de nueva generación (NGS, del inglés *next-generation sequencing*). Las primeras máquinas de secuenciación salieron al mercado en 2005, permitiendo alcanzar cotas muchísimo más altas de secuenciación que anteriormente y sobre todo, reduciendo

el coste de secuenciación por base: entre 2007 y 2012, por cuatro órdenes de magnitud. Esta tecnología es en líneas generales la predominante hoy en día, y permite secuenciar en cuestión de días un genoma humano completo por un coste inferior a 600€.

El procedimiento resumido de NGS (existen varias tecnologías desarrolladas por diferentes empresas) es el siguiente. Primero, la molécula de ADN a secuenciar se fragmenta en múltiples pedazos: ha de notarse que estos pedazos no son estrictamente disjuntos entre sí, sino que se superponen, ya que se fragmentan miles o millones de células. Estas subpartes son amplificadas después, esto es, se multiplican las pequeñas cadenas de ADN. A continuación, se produce la secuenciación *per se* de estos fragmentos de ADN en paralelo, secuenciando muchas en paralelo (esta es la diferencia esencial de NGS con las técnicas anteriores). En el caso de una de las tecnologías existentes (y más popular), esto se hace acoplando marcadores fluorescentes a los nucleótidos y sacando una imagen con una cámara: cada marcador emite luz a una frecuencia diferente, que permite identificar, para cada una de las múltiples pequeñas cadenas de ADN, qué nucleótido se ve en su extremo, en particular en el llamado 3'[1]. Después, se repite este proceso avanzando un nucleótido en la cadena, y así hasta que se termina.

Con este procedimiento (del cual un esquema se puede ver en la figura 1.3) no se obtiene una sola secuencia de ADN, sino millones de ellas. La información de estas pequeñas secuencias, llamadas lecturas (comúnmente *reads*), ha de ser tratada para poder obtener la información de la molécula completa que queríamos secuenciar. El proceso por el cual se organizan y ordenan estos pequeños fragmentos para obtener la secuencia original se denomina ensamblaje, y es un proceso computacionalmente complejo y costoso. El hecho de haber fragmentado el ADN y luego haberlo amplificado permite que los algoritmos de ensamblaje hallen segmentos en común para ir enlazando las secuencias. Como se secuencia de manera redundante, al estar las lecturas repetidas esto permite tener una superposición de ellas sobre las secuencia que se está ensamblando. Esto ayuda a prevenir potenciales errores en la secuenciación, al poder exigir en el ensamblaje una determinada «cobertura» (es decir, que la secuencia ensamblada tenga superpuesta tantos *contigs* individuales a lo largo de la misma).

En el caso de que exista un genoma de referencia (como en el caso humano), el proceso es

[1]Toda cadenas de ADN o ARN tienen dos extremos llamados 3' y 5', viniendo estos nombres de los carbonos que sirven de engarce para las moléculas de azúcar de los nucleótidos. De esta forma, el extremo 3' se corresponde con el último nucleótido que tiene el carbono 3' libre, y de forma análoga el 5'. Esto es relevante también para establecer sentidos a la hora de recorrer la cadena, por ejemplo, 5'→3' es el sentido en el que se da la síntesis real de ADN.

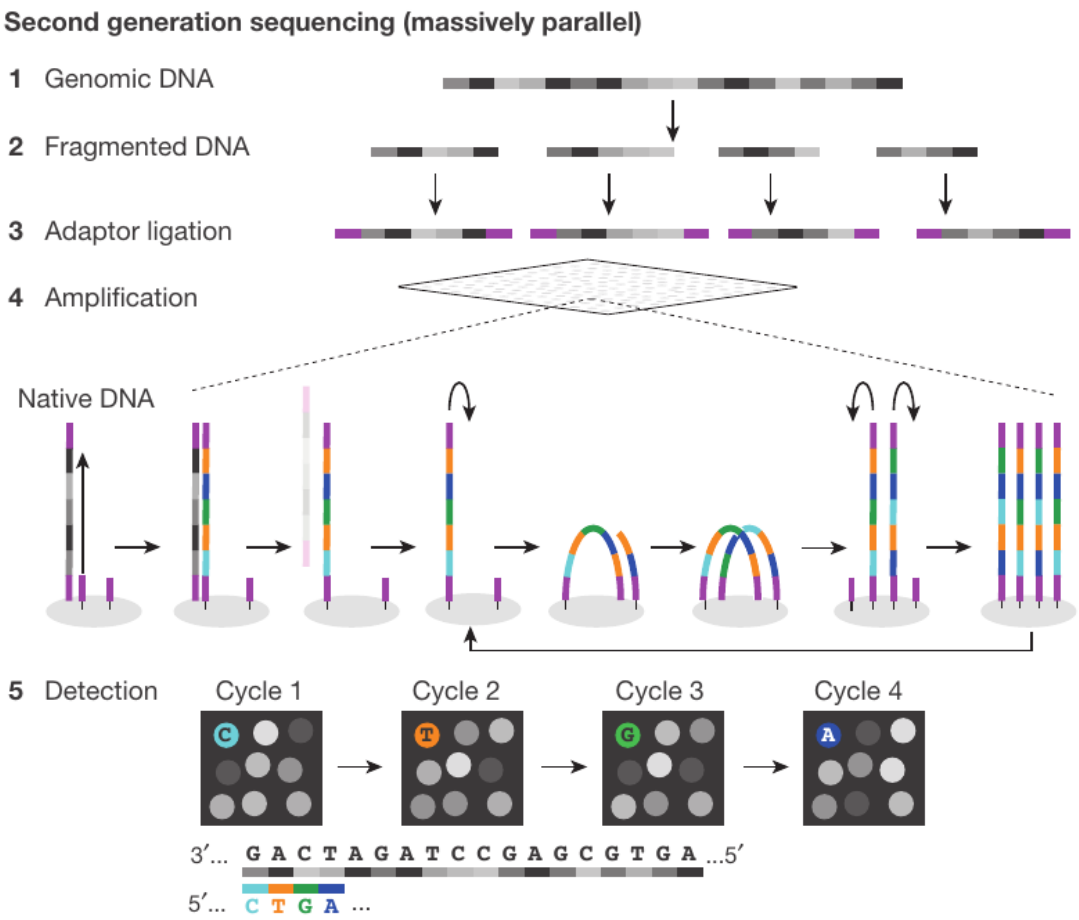


Figura 1.3: (Shendure y col., 2017) Diagrama de los procesos de secuenciación de segunda generación o NGS.

algo más sencillo puesto que se puede hacer un alineamiento de las lecturas con el genoma de referencia. Si no se tiene, se dice que la secuenciación (y el ensamblaje) es *de novo*.

En los últimos años han aparecido nuevas tecnologías que conforman lo que se llama ya secuenciación de tercera generación, en contraposición con NGS, que se denomina también secuenciación de segunda generación. Estas permiten secuenciar una molécula sola (sin necesidad de trocear y amplificar ni re-ensamblar la información después) de forma continuada en tiempo real. No profundizaremos en ellas, puesto que se alejan del tema que concierne a este trabajo.

1.2.2. Secuenciación de ARNm

Como mencionamos con anterioridad, el ARN mensajero es al final quien traslada la información de un gen al ribosoma para que pueda sintetizarse la proteína que el gen codifica para la

función correspondiente. Una noción importante, si bien aparentemente trivial, es que no todos los genes se expresan en una célula, ni los genes que se expresan lo hacen al mismo nivel. Las principales diferencias entre las células de diferentes tejidos (e.g. una neurona y una célula muscular) se debe a que ambas expresan genes distintos, y solo algunos comunes. El conocimiento de qué genes se expresan más y cuáles menos (y en qué circunstancias, y de qué células) es la principal motivación tras la secuenciación de ARNm: su mayor o menor presencia en una célula está directamente relacionada con la mayor o menor expresión de un gen determinado, debido al rol de «transportista» que ocupa en la transcripción y traducción de la información genética. Este conocimiento se denomina expresión diferencial genética (DGE, del inglés *differential gene expression*).

Afortunadamente, no hay necesidad de crear una técnica de cero para hacer esta secuenciación como con la del ADN: podemos servirnos de los desarrollos de esta última, si hacemos el paso inverso a la transcripción. Esta técnica es posible gracias a la transcriptasa inversa (o retrotranscriptasa), que es una enzima presente en los retrovirus que permite sintetizar la cadena complementaria de ADN a partir de una molécula de ARN. Esta cadena se denomina ADN copia (ADNc o cDNA en inglés).

Así, podríamos partir de nuestras moléculas de ARNm, obtener su análogo en ADNc, hacer una secuenciación de ADN usando las técnicas de NGS descritas en la sección anterior, y obtener el resultado en bruto usual de una secuenciación de ADN. Con una diferencia: hemos secuenciado múltiples moléculas de ARNm, así que el resultado corresponde a todas ellas. La presencia de lecturas correspondientes a un mismo gen será proporcional a la cantidad de moléculas de ARNm que hayamos introducido, y por lo tanto, a su expresión celular. Esto nos permite obtener mapas comparativos de la expresión génica.

Hoy en día, la secuenciación siguiendo este proceso es el estándar *de facto*. Un detalle relevante a hacer notar es que en la secuenciación común de ARNm no es posible secuenciar tan solo las moléculas de ARNm correspondientes a una sola célula: se hacen de conjuntos, muestras de tejidos, de forma que la expresión de los genes que se observa es siempre promedio entre todas las células que se usaron.

Los últimos desarrollos en las técnicas de secuenciación de ARNm permiten por ejemplo la secuenciación de cadenas más largas (dando lugar a lecturas más largas, *long-read RNA sequen-*

cing) de ARNm, sin necesidad de usar fragmentos pequeños como en la secuenciación común (*short-read RNA sequencing*). Otros, aprovechándose de las tecnologías de tercera generación, permiten hacer una lectura completa de la cadena, sin necesidad de fragmentar el ARNm. Y en la última década se ha ido posibilitando la secuenciación de ARNm en células individuales (*single cell RNA sequencing* o *scRNA sequencing*) diferenciada de la secuenciación «en bloque» común (*bulk RNA sequencing*).

1.2.2.1. Secuenciación de ARNm en células individuales

La motivación principal de secuenciar el ARNm de una sola célula es eliminar el promediado que se hace en la secuenciación usual (*bulk RNA-seq.*) sobre los patrones de expresión de todas las células cuyos ARNm se secuencian. Si en una muestra de tejido o de tumor existen dos poblaciones de células, unas que expresan un gen y otras que no lo expresan, cuando se analiza la expresión mediante *bulk RNA-seq.* se ve que hay cierta expresión del gen (promedio), mientras que si se analizan mediante *single cell RNA-seq.*, se podrá observar que hay dos poblaciones, ya que se estudia cada célula por independiente. Esto permite un estudio mucho más fino y mediciones exactas sobre qué células (o tejidos celulares) expresan qué genes en qué proporción.

En 2009 (Tang y col., 2009) se consiguió por vez primera secuenciar ARNm de una única célula, para lo cual tuvo que ser aislada manualmente primero. Desde entonces, han aparecido nuevos métodos (Svensson y col., 2018) que automatizan este proceso de aislamiento o separación de las células y permiten paralelizarlo, para realizar muchas lecturas simultáneas.

La estrategia general de secuenciación de ARNm (obtener ADNc, y secuenciar este) se sigue aplicando aquí, pero hay diferencias relevantes con respecto al *bulk RNA-seq.*. Tomemos para ejemplificarlo uno de los métodos usados en la actualidad, en el cual las células se atrapan en gotas empleando sistemas de microfluídica. Después, dentro de esas gotas, se introducen pequeñas esferas que llevan pegadas en el exterior de su superficie muchas secuencias pequeñas de ADN con las que las moléculas de ARNm de la célula tienden a enlazarse, pues la secuencia contiene una cadena larga de timinas (una cadena de poly(T)). Por construcción, todas las cadenas de ARNm poseen una larga cadena de adeninas en uno de sus extremos. Debido a la complementariedad A-T, los puentes de hidrógeno permiten «pescar» con esas cadenas de poly(T) a los ARNm de la célula. Sin embargo, esa cadena de poly(T) no es lo único que tienen esas cadenas: también

poseen un «código de barras», que es una pequeña secuencia artificial de nucleótidos, diferente para cada esfera. Puesto que en cada gota entra una célula y una esfera, este código de barras resulta diferente para cada célula, y permite después identificar todos los ARNm de una misma célula. Después, los contenidos de todas las gotas (múltiples células se procesan a la vez) se juntan y las esferas son eliminadas. El ADNc unido a cada ARNm se puede usar como partida para terminar de retrotranscribir todo el ARNm, y después secuenciarlo como con cualquier cadena de ADN. Un esquema general del proceso entero se puede ver en la figura 1.4.

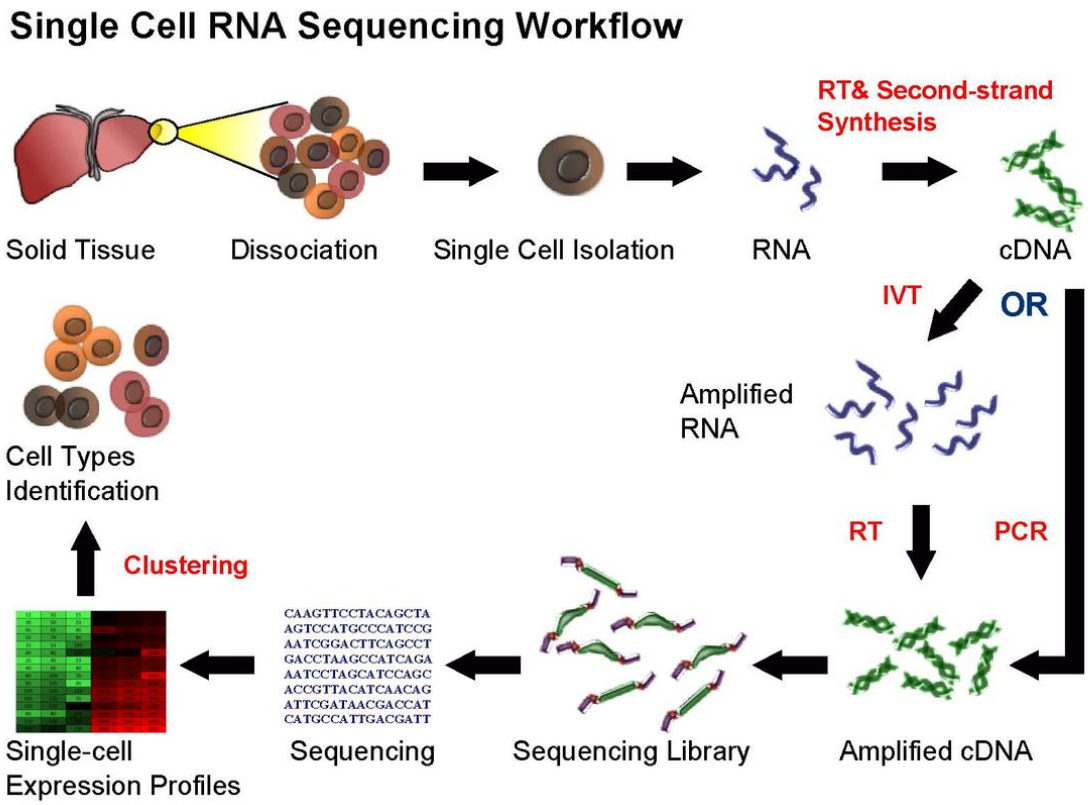


Figura 1.4: (*Wikipedia, the free encyclopedia*, s.f.) Esquema del proceso de secuenciación de secuenciación individual de ARNm. Esta imagen incluye un paso adicional por el cual es posible identificar, con los datos de la secuenciación, qué tipos de célula es cada una (en función a su expresión).

Las diferencias con la secuenciación de ARNm común empiezan por el número de cuentas totales (es decir, de *lecturas* que tenemos en los datos en bruto) que se obtienen. Mientras que en el *bulk RNA-seq.*. trabajamos con del orden de $50 \cdot 10^6$ lecturas, en *scRNA-seq.*. esta cifra se reduce a $20 - 60 \cdot 10^3$ lecturas. Esto tiene una implicación relevante: al tener menos lecturas es menos probable que secuenciemos ARNm de aquellos genes cuya expresión es más baja.

Otra diferencia clara con este método es que las secuencias que podemos asociar directamente

con una u otra célula con este método tan solo nos permiten secuenciar un fragmento de la cadena original de ARNm cercano al extremo con la cola de poly(A), puesto que es la que contiene el código de barras que nos garantiza el origen. Esto puede dificultar más la vinculación con los genes correspondientes.

A pesar de estas dificultades añadidas para el estudio de los datos, las posibilidades de investigación de la secuenciación individual de ARNm son muy relevantes y es un campo en actual desarrollo.

1.3. El cáncer

(Carlberg & Velleuer, 2021) El cáncer es un conjunto diverso de enfermedades cuyo nexo de unión es el crecimiento celular descontrolado. Debido al incremento global en esperanza de vida durante el pasado siglo, se ha ido incrementando su relevancia como causa de muerte. Durante el siglo XX, las ratios de muerte por cáncer se incrementaron significativamente hasta 1991 (con 215 muertes por 100.000 personas), fecha desde la que comenzó a descender.

Un tumor puede no conllevar el desarrollo de un cáncer necesariamente, y en tal caso se dice que es benigno, en contraposición con los casos en los que el crecimiento es descontrolado, cuando decimos que es maligno y pasa a constituir un cáncer. El origen de los tumores se debe a la acumulación de mutaciones en el ADN de nuestras células, que modifican el comportamiento de las mismas y fomentan su división. Esto se pudo establecer a mediados del siglo pasado, gracias al avance de la investigación genética y también al desarrollo de las técnicas experimentales (como con la secuenciación genética) en el campo. Ello permitió lograr la identificación del primer gen que fomentaba la multiplicación y el desarrollo celular en un tumor maligno en los años setenta. En 1969 se acuñó el término «oncogén» para designar a ese tipo de genes. Más tarde se descubrieron otro tipo de genes, también mutados en cáncer, y cuya función en las células normales es evitar la transformación tumoral, por lo que se denominaron «genes supresores tumorales».

En las últimas décadas del siglo XX y a principios del XXI, se avanzó en el descubrimiento y caracterización de estos genes conductores del cáncer (*driver genes*). Estos serían aquellos genes

que, tras verse afectados por una mutación, facultan a que la célula donde esta ocurra (i) resista la apoptosis (i.e. la muerte celular programada), (ii) fomente la proliferación celular, (iii) evada los supresores de crecimiento celular, (iv) inicie la invasión de otros espacios y la metástasis celular, (v) permita la replicación celular infinita (normalmente las células se reproducen un número limitado de veces), (vi) fomente la angiogénesis (la creación de nuevos vasos sanguíneos), (vii) permita la desregulación del metabolismo celular o (viii) evite la destrucción por parte del sistema inmune. Las mutaciones que otorgan a los genes alguna, o varias de estas capacidades se denominan mutaciones conductoras (*driver mutations*), en contraposición con las mutaciones «pasajeras»: estas serían mutaciones somáticas (i.e. que no se propagan a la descendencia del individuo) del tumor, al igual que las conductoras, pero que no modifican la capacidad de la célula para crecer.

La comprensión de qué genes conductores se encuentran tras el desarrollo de qué tumores permite comprender cómo evolucionan estos (y en general, mejorar el conocimiento al respecto de los tumores), y con ello llevar a potenciales nuevos desarrollos clínicos. La idea tras ella es que los tumores siguen una evolución natural (darwiniana), según la cual las células del tumor, desde su origen, se desarrollan a través de la variabilidad (debida a nuevas mutaciones) y la selección (que permite que ciertas mutaciones otorguen ventajas a algunas células sobre las demás). Esta evolución se suele dar en torno a los genes conductores.

Los tumores pueden desarrollarse en diversos órganos o tejidos, y en función a cuál sea el origen de un cáncer, estos se diferencian en grupos: los carcinomas (la inmensa mayoría,∼80–90 %) surgen de células epiteliales, los blastomas (∼1 %) de células precursoras o tejidos embrionarios, los sarcomas (∼1 %) del tejido conectivo, etcétera. Los tumores «líquidos» se originan a partir del tejido hematopoyético, que es el responsable de la producción de las células sanguíneas. Los cánceres derivados de estos tumores se denominan leucemias y constituyen un 5 % del total de cánceres en adultos. En función al ratio de crecimiento se agrupan en agudas (las que tienen un crecimiento celular y desarrollo cancerígeno rápido) y crónicas (de desarrollo más pausado). Y, en función a sobre qué células sanguíneas en particular afecta el tumor, se clasifican en dos grandes grupos: mieloides y linfáticas (también linfocíticas o linfoides). En el primer caso, son cánceres conformados por tumores que afectan a las células cuyo desarrollo se da en la médula ósea, como los eritrocitos (glóbulos rojos) o los monocitos (glóbulos blancos o leucocitos, por ejemplo). En el caso de las leucemias linfáticas, están hechas de tumores de células cuyo desarrollo no concluye en la médula ósea, estos son los linfocitos T, B y NK.

1.4. Leucemia linfática crónica

(Bosch & Dalla-Favera, 2019; Nadeu y col., 2018) La leucemia linfática crónica (LLC o también del inglés, CLL) es, como veníamos diciendo, un tipo de cáncer líquido (i.e. afecta a la sangre) cuyo desarrollo no es necesariamente rápido (crónico o indolente) y que afecta a los linfocitos B. Es la leucemia más frecuente en Europa y Estados Unidos, con una incidencia estimada de 4.7 casos por 100.000 personas, con mayor predominancia en hombres.

1.4.1. Desarrollo

Es una enfermedad heterogénea, con desarrollos diferentes en función a la evolución del tumor. La célula origen de los tumores que conforman LLC (a partir de la cual, vía mutaciones, un tumor de LLC se originaría) no ha sido establecida con seguridad, pero se hipotetiza que sean las células madres hematopoyéticas (i.e. del tejido de la médula ósea). Después, esa célula daría lugar a un conjunto de progenitores de linfocitos (o células) B: estas células son glóbulos blancos, componentes esenciales del sistema inmunitario, que se encargan de producir anticuerpos. Ese conjunto sería clonal.

Una de las clasificaciones de los tumores de LLC se da en función a si sus células presentan mutaciones en los genes que se encargan de producir inmunoglobulina (las moléculas que dan lugar a los anticuerpos). Así, podemos hablar de tumores que poseen mutaciones en una región del genoma donde se hallan esos genes denominada IGHV (*inmunoglobulin heavy chain variable region*), o no. Varias características del desarrollo de la enfermedad dependen de la presencia de mutaciones o no en la región IGHV.

El conjunto clonal de células B mutadas dará lugar eventualmente a una linfocitosis, que es la presencia anormalmente alta de linfocitos en la sangre. En este caso particular, esta se denomina linfocitosis monoclonal de células B (MBL, *monoclonal B lymphocytosis*). Finalmente, mutaciones adicionales convierten el tumor en maligno y se da el paso de MBL a LLC.

1.4.2. Mutaciones recurrentes

El avance de las técnicas de secuenciación genómica ha permitido comprender y estudiar cómo las mutaciones comúnmente presentes en las LLC afectan al genoma de las células tumorales. Los estudios muestran que la presencia o no de algunas de estas mutaciones (y su correlación con la mutación en IGHV) pueden afectar significativamente a la prognosis de la enfermedad. Aquí siguen algunas de las alteraciones más frecuentes que aparecen en LLC.

del13q14 La deleción del brazo largo (q) del cromosoma trece en su decimocuarta región es la afectación genética más predominante en LLC, presente en un 50-60 % de los pacientes. Suele estar correlacionada con mutaciones en IGHV, mejor tiempo hasta el primer tratamiento (TTFT, *time to first treatment*) y mejor supervivencia general (OS, *overall survival*).

Mutaciones en *NOTCH1* *NOTCH1* es un oncogén que también tiene relevancia en otras leucemias. Las mutaciones en este gen aparecen en un ∼12 % de los pacientes diagnosticados con LLC, y la frecuencia de estas se incrementa cuanto más avanzada la enfermedad está y también si está presente la trisomía del cromosoma 12. De aquellos pacientes que tienen mutaciones en *NOTCH1*, el 80 % aproximadamente no tienen mutaciones en la región IGHV. La presencia de mutaciones en *NOTCH1* está relacionada con peor OS.

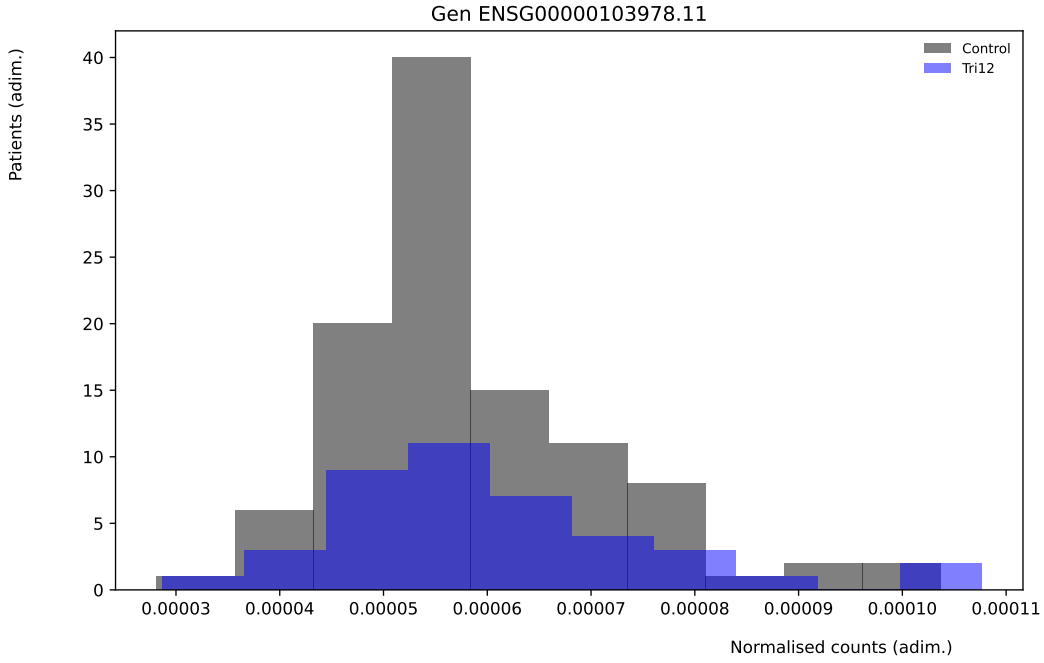
tri12 La trisomía del cromosoma 12 aparece en aproximadamente el 15 % de los pacientes de LLC, y se ha vinculado con modificar la morfología de los linfocitos. Suele ocurrir a la vez que aparecen mutaciones en el oncogén *NOTCH1*.

2 Metodología y desarrollo

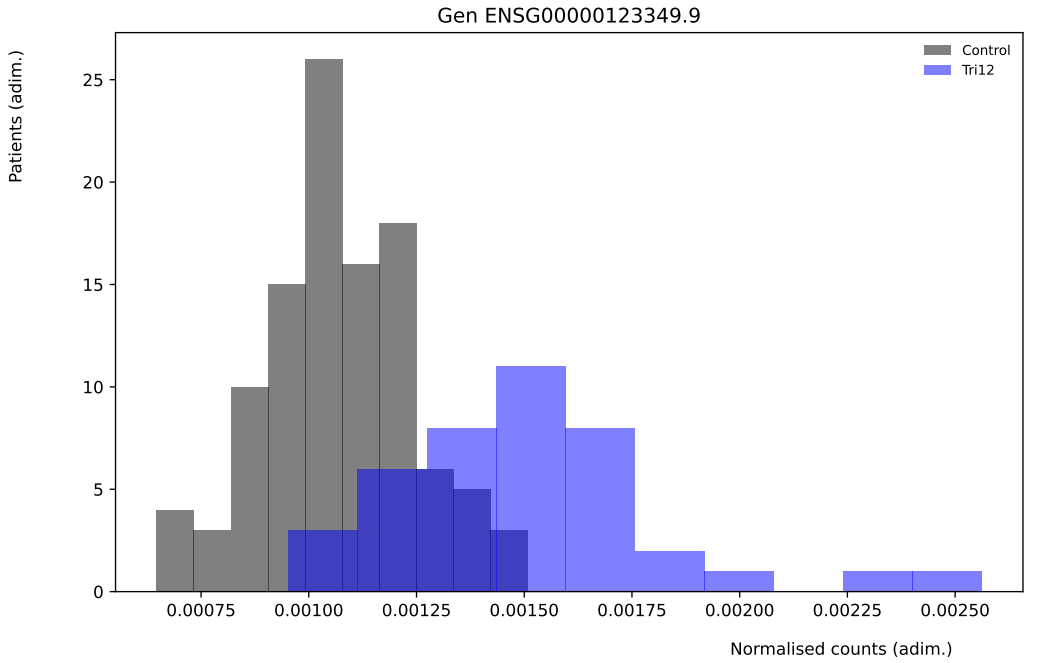
Tras la introducción del contexto necesario, dedicaremos este capítulo a detallar tanto los datos como el desarrollo del trabajo en sí, con el fin de cumplir el objetivo del mismo: construir un modelo de aprendizaje automático que nos permita identificar la presencia (o no) de una mutación determinada en información extraída de secuenciaciones individuales de ARNm (*single cell* RNA-seq.). Para ello, usaremos datos que provienen de secuenciación en bloque, que describimos en la sección 2.1, con los que hicimos los desarrollos explicados en la sección 2.2.

2.1. Datos

El conjunto de datos a emplear consiste en cuentas de lecturas de genes procedentes de secuenciación en bloque (*bulk RNA-seq.*) de tumores procedentes de pacientes que sufren LLC. Están ya anonimizados (cada fila de cuentas, correspondiente a un paciente, lleva asignado un número natural como identificador) y ofrecen también, para cada entrada, cierta información extra relacionada con la información de los genes (como si se comprobó o no la presencia de ciertas mutaciones en las células tumorales del susodicho paciente), o detalles sobre el paciente (e.g. sexo). El número total de entradas que tenemos son 273, y en conjunto conforman una matriz de cuentas, con el número de lecturas de 57820 genes humanos. De la mayor parte, no tenemos información (no tenemos ninguna cuenta): hay 15782 genes de los que al menos hay una lectura en un paciente, cada uno de ellos con un identificador de la forma `ENSGXXXXXXXXXXX.X` (donde X identifica un número cualesquiera). En la figura 2.1 se pueden observar dos histogramas con este conjunto de datos, para dos genes donde se pueden ver separadas las contribuciones de aquellos casos en los que se da una mutación (trisomía del cromosoma 12) y en los que no.



(a)



(b)

Figura 2.1: Histogramas de los datos totales separados según tengan la mutación de la trisomía del cromosoma 12 o no para los genes ENSG00000103978.11 y ENSG00000123349.9.

A la vista de esta descripción está claro que nuestro conjunto de datos posee una enorme cantidad de características ($O(10^4)$) comparada con la cantidad de datos que poseemos ($O(10^2)$). Esto tendrá que ser tenido en cuenta en el desarrollo de los modelos.

2.1.1. Consideraciones experimentales

Como explicamos en el capítulo anterior, el proceso de secuenciación de ARNm consta de bastantes partes que han de ser repetidas para cada secuenciación de una muestra. Las variaciones debido a factores puramente experimentales dan lugar a que, en verdad, cada secuenciación sea «única» en el sentido de que es imposible hacer una secuenciación de otra muestra con exactamente las mismas condiciones.

Una de las consecuencias, es que mismamente el número total de lecturas (también llamado el tamaño de la librería) es diferente para cada secuenciación, a veces de forma significativa. Esto hace pertinente que, para ser capaces de poder usar todos los datos de forma conjunta, debemos tener en cuenta estas diferencias. Una de las formas de afrontar esta situación es aplicar una normalización a los datos, de forma que tras ello sean comparables. La normalización más directa e intuitiva, que sería normalizar cada secuencia al número total de cuentas, no es muchas veces la preferida en el campo, optándose por otras derivadas como TMM (*trimmed mean of M values*), o RLE (*relative log estimate*), más complejas. Algunas comparaciones entre varias normalizaciones se pueden ver en (Abrams y col., 2019; Li y col., 2020).

2.1.2. Consideraciones estadísticas

El proceso de lecturas o cuentas a partir de la secuenciación de ARNm puede considerarse como un proceso de Poisson y, por lo tanto, *a priori* cabría esperar que la distribución de las mismas siguiera una distribución de Poisson. Sin embargo, los resultados empíricos muestran que esto no es así (Auer & Doerge, 2011; Robinson & Smyth, 2007): se muestra una tendencia a la llamada «sobredispersión» que existiría respecto a los valores de la distribución de Poisson. A partir de la función de probabilidad de esta distribución,

$$P(x;\lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!}, \tag{2.1}$$

se deriva fácilmente una propiedad intrínseca a la misma: su esperanza y su varianza son iguales, i.e. $E(x) = \mu = \lambda = Var(x)$, fijándose así la dispersión de una Poisson toda vez su media poblacional lo está.

Las secuenciaciones de ARNm muestran que la distribución no ajusta bien si la desviación típica se fija a $\sqrt{\mu}$. No se conoce con certeza el motivo de este fenómeno, si bien se suelen aducir motivos experimentales (efectos del proceso de secuenciación, «ruido» introducido al tratar las muestras, etc.) para apuntar a su origen. También se teoriza con que la sobredispersión tendría raíces termodinámicas y de cinética química de las reacciones intracelulares que sintetizan en última instancia el ARNm (Konishi, 2005): teniendo en cuenta tal caso, la distribución que seguirían las cuentas sería en efecto una distribución de Poisson, pero la cantidad de ARNm seguiría una lognormal (distribución definida de forma que su logaritmo sigue una distribución normal). De esta forma, nos encontraríamos con una mezcla de distribuciones, que permiten ya sí una dispersión mayor que la de la poissoniana. La relación entre la desviación típica y media muestrales (que nos permite evaluar la sobredispersión o no con respecto a lo esperable de la Poisson) se puede encontrar en el entorno de $\frac{\sigma}{\mu} \approx 0{,}1$ (Gierliński y col., 2015).

De forma fenomenológica, con el fin de estudiar los datos y (sobre todo) realizar test de hipótesis para evaluar la expresión diferencial de los genes (una de las principales motivaciones para hacer secuenciación de ARNm), se han ido empleando sustitutos o derivados de la distribución de Poisson para modelizar los datos flexibilizando el requisito de dispersión de esta última. Para el caso que describíamos de la distribución de cuentas que sigue un solo gen, se usa sobre todo la distribución binomial negativa (un caso particular de la distribución gamma, al igual que la Poisson), así como la distribución normal o la ya mencionada mezcla de Poisson y lognormal.

A la hora de generalizar el caso a un conjunto de genes (que, como mencionamos antes, pueden llegar a ser $O(10^{3-4})$), e intentar modelizar completamente los datos, debemos enfrentarnos a la necesidad de estimar bastantes parámetros (expresión, correlaciones, dispersión) para potencialmente un gran número de genes. Se emplean en el campo modelos lineales generalizados (GLM) que se apoyan en algunas distribuciones de las mencionadas anteriormente para modelizar los datos y realizar test de hipótesis (tests de Wald) de cara a evaluar DGE (McCarthy y col.,

2012). También existen alternativas, como modelos multivariable basados en la mezcla Poisson y lognormal (Zhang y col., 2015).

2.1.3. Preprocesado

Antes de comenzar con la preparación de los modelos, se hace un preprocesado general a los mismos, en los que se combina la matriz de cuentas con la información extra (metadatos). En el caso de los datos originales, están en formato `csv`. Los metadatos provienen en formato de hoja de cálculo (`xlsx`). Después se arreglan algunas entradas marcadas como `nan`, sustituyéndolas por ceros, puesto que provienen de errores al transcribir desde los formatos originales a `DataFrame` de `pandas`. Finalmente, se eliminan todos los genes que no poseen al menos una cuenta.

2.2. Estrategia y desarrollo

2.2.1. Modelo de aprendizaje automático

Para comprobar si será posible detectar mutaciones con un modelo de aprendizaje automático usando datos de *single cell RNA-seq.* a partir de los nuestros de *bulk RNA-seq.*, intentaremos primero entrenar un modelo haciendo simulaciones de Montecarlo que imiten cuentas de secuenciaciones hechas con *single cell RNA-seq.*. Además, nos centraremos en detectar, como mutación, la trisomía del cromosoma 12, que puesto que consiste en añadir una copia más de un cromosoma, es esperable que sus efectos sean los más grandes de las distintas mutaciones que comúnmente se contemplan en tumores de LLC.

Antes de llegar a la fase en sí de entrenamiento de un modelo, necesitamos crear pues nuestro conjunto de pseudodatos. El primer paso será quedarnos tan solo (por simplificar el proceso) con aquellos genes de los cuales podemos esperar una cuenta al menos. Para ello, primero debemos estimar, aunque sea de una forma somera, la probabilidad de que al hacer una lectura, esta sea de un gen determinado. Para ello, primero normalizamos las cuentas de todas las secuenciaciones que tenemos a la suma total de lecturas de cada una (esta será la normalización que escogeremos

para este trabajo). Podríamos asumir que los valores que ya tenemos conforman una distribución de probabilidad discreta *per se*, pero para refinar la estimación global, podemos cogernos la media muestral para cada gen con todos los datos como proporción. En este caso, una pequeña renormalización es necesaria para que todas las probabilidades sumen la unidad. Con esa distribución de probabilidad discreta, podemos estimar el número de lecturas que se podrían esperar con una cantidad de lecturas determinada, que para este trabajo consideraremos como 25000 lecturas, multiplicando la probabilidad de un gen por ese mismo valor. Con los datos que tenemos, esto nos deja un total de 3189 genes.

Con esa lista de genes, podemos generar una lista nueva de cuentas a partir de una secuenciación de Montecarlo creada a partir de 25000 números (pseudo)aleatorios que siguen la distribución de probabilidad que extrajimos en el paso anterior. En nuestro caso, hemos creado un total de 102900 pseudosecuenciaciones. Esta generación se hace paciente a paciente, usando la lista de genes extraída anteriormente como los únicos posibles, y las probabilidades que se pueden extraer del paciente en cuestión en exclusiva. De esta forma, luego podemos crear los conjuntos de prueba y entrenamiento teniendo en cuenta qué pacientes se usan en uno y otro sitio. Debido al desbalanceo entre las dos clases (mutadas con trisomía del cromosoma 12, o que no poseen esa mutación), esta división se hace manteniendo esas proporciones en ambos conjuntos.

Una apreciación importante ha de hacerse aquí: estamos ignorando cualquier tipo de correlación que pueda haber entre la expresión de los genes considerados. Las correlaciones de expresión génica son esperables y tiene sentido que existan, pero la generación de secuenciaciones de Montecarlo que las respeten sería mucho más compleja, y para el método descrito en esta subsección decidimos ignorarlas.

Esto nos ofrece ya un conjunto de pseudosecuenciaciones con el que trabajar: estas tienen los genes que hemos estimado que con 25000 cuentas totales podríamos ver. Para poder llegar a la fase de entrenar modelos, lo primero que hacemos es la separación entre conjunto de prueba y entrenamiento. Esto lo hacemos a nivel de paciente, de forma que no se mezclan las pseudosecuenciaciones de distintos pacientes. Con esos dos conjuntos, hacemos un procesado consistente en los siguientes pasos.

1. Normalizar las cuentas a la suma total: puesto que al fin y al cabo las secuenciaciones reales de *single cell RNAseq* que se puedan comparar tendrán un número variable de

lecturas totales y deberemos compararnos en un territorio común con ellas.

2. Debido a la naturaleza estadística de los datos, y con el fin de uniformarlos, hacer una estandarización (a una normal) para las cuentas de cada gen.

Tras estas acciones seguimos manteniendo una gran cantidad de variables, grande a pesar de los pseudodatos generados. Con el fin de reducir, al menos parcialmente, la dimensionalidad del conjunto de datos, intentamos aplicar un PCA (*principal component analysis*). Sin embargo, los resultados de usarlo no fueron muy satisfactorios, resultando en modelos aún poco potentes, como se muestra en resultados.

Tras este procesado, procedemos a la fase de preparar un modelo de aprendizaje automático para ser entrenado. Debido a que, con la gran cantidad de características de nuestros datos no sería descabellado sobreentrenar nuestros modelos, hemos decidido trabajar con bosques aleatorios (*random forests* (Breiman, 1998, 2001)), que por sus características (el uso de *bagging* tienden a evitar el sobreentrenamiento.

Los intentos iniciales de entrenamiento ofrecieron modelos de clasificación que ofrecían valores razonables de AUC ($\sim 0{,}8$) y precisión ($\sim 0{,}7$). Para evitar un sobreentrenamiento evidente hubo que ajustar en gran medida las capacidades de los árboles aleatorios, reduciendo su profundidad y la cantidad de características que puede usar cada árbol al máximo.

Pese a ello, los resultados eran desconcertantes. En particular, se observó cómo aparentemente los resultados del conjunto de prueba (que inicialmente se estimó con un tercio de los pseudodatos totales) eran mejores que en el caso del de entrenamiento. A pesar de hacer varias pruebas cambiando las proporciones de ambos conjuntos hasta llegar al 50 %-50 %, se seguían observando fenómenos extraños: es más, los resultados variaban significativamente al hacer este cambio (los modelos que no sobreentrenaban, pasaban a hacerlo, las curvas ROC cambiaban de forma brusca de trayectoria, etc.). Una última prueba, cambiar las semillas de generación de números pseudoaleatorios (que está fijada manualmente y se propaga adecuadamente a todas las funciones y clases que la necesite para que sea siempre la misma semilla) indicó cuál era la causa del origen. Si solo cambiábamos eso, ya se daban todas estas variabilidades.

Por consiguiente, el problema estaba en que no somos capaces de estimar correctamente un

conjunto de prueba y de entrenamiento. La cantidad de pseudodatos que tenemos ($O(10^5)$) combinado con el número de características que estamos usando (83), no hacen pensar que el problema sea el número de datos necesariamente. De hecho, teniendo en cuenta los cambios al cambiar de semilla (lo que en definitiva afecta esencialmente al cambio de conjuntos de prueba y entrenamiento y qué pseudodatos de qué pacientes los contienen) hacen pensar que la cuestión es que a pesar de generar pseudodatos que son levemente diferentes entre sí, se siguen pareciendo demasiado entre sí y a (con la diferencia del número de lecturas totales) la secuenciación de *bulk RNA seq.* real, y ese efecto no es despreciable (que entre un dato en el conjunto de prueba o entrenamiento) porque tenemos en total pocos datos de secuenciaciones reales.

En la sección 3 se muestran curvas ROC y AUC de múltiples modelos que se hicieron tratando de sortear estas dificultades y buscando mejorar la potencia discriminadora de los mismos. En particular, se incluyen estas comparaciones:

- Preselección de genes: hacer una pre-evaluación del poder discriminatorio de cada gen de forma individual (ignorando toda posible correlación) posibilita mejorar la calidad de los modelos, tomando cada gen como variable discriminadora, obteniendo por lo tanto una curva ROC de la cual extraer un AUC, y de esta forma poder clasificarlos por su poder clasificador, de forma *naive*. Esto nos permite después quedarnos tan solo con los genes cuya AUC asociada sea mayor que un valor, que hemos escogido en 0.8. Como se muestra en resultados, permite reducir el número de características de forma tal que los resultados de los modelos finales son mejores. Esta preselección se aplica antes del PCA.

- Eliminación de casos extremos: una posible explicación que habría a variaciones significativas de resultados en función a cómo hagamos los conjuntos de prueba y entrenamiento es la existencia de casos extremos u *outliers*. Existen múltiples algoritmos que permiten detectar casos extremos en conjuntos de datos: nosotros extrajimos listas con varios de ellos (*isolation forest*, *local outlier factor* o LOF, SVM de una clase) usando el conjunto de datos original (no las pseudosecuenciaciones), incluidos en `scikit-learn`. Al final nos quedamos con la lista extraída con un LOF con 10 «vecinos», puesto que encontraba un número moderado de casos extremos que se superponía más o menos con otras listas más extensas del resto de modelos. El LOF encontró un total de ocho pacientes que pueden ser considerados como casos extremos.

Los obstáculos encontrados nos forzaron a plantearnos otras formas alternativas de proceder. Se intentaron variaciones para incrementar aún más el número de pseudosecuenciaciones, como mismamente generar un conjunto mayor o antes de eso usar *bootstrap* con las mismas, pero ninguno dio resultado. Otro planteamiento que hicimos fue tratar de generar pseudosecuenciaciones más enriquecidas y variadas, quizás incluyendo la información de las correlaciones entre genes, como detallamos en el siguiente epígrafe.

2.2.2. Método alternativo: test de hipótesis

Con el fin de buscar una salida alternativa, debido a los inconvenientes encontrados durante la construcción de los modelos, dedujimos un clasificador alternativo sacando partido del conocimiento que tenemos de las distribuciones de los datos. Esto nos permitió construir un test de hipótesis con el que extraer p-valores para usar como discriminadores y también intentar incluir las correlaciones entre la expresión de los distintos genes.

Como ya mencionamos en la sección 2.1.2, los datos con los que trabajamos son cuentas que siguen una distribución de Poisson, si bien con una dispersión mayor de la misma. Una aproximación en primera instancia permitiría hacer una aproximación de los datos por distribuciones normales, a las cuales las distribuciones de Poisson tienden cuando su parámetro λ es suficientemente elevado (en nuestro caso, esto se traduce en tener suficientes cuentas).

Partiendo pues de una distribución multinormal con un vector de medias $\vec{\mu}$ y una matriz de covarianza V, definimos el test de hipótesis con el que trabajaremos como sigue.

$$\begin{aligned} H_0 &: \vec{\mu} = \vec{\mu}_{\text{tri12}}, \quad V = V_{\text{tri12}} \\ H_1 &: \vec{\mu} \neq \vec{\mu}_{\text{tri12}}, \quad V = V_{\text{tri12}} \end{aligned} \tag{2.2}$$

Si bien sería posible definir una hipótesis alternativa más concreta,

$$H_2 : \vec{\mu} = \vec{\mu}_{\text{control}}, \quad V = V_{\text{control}}, \tag{2.3}$$

esto no nos permite hacer uso del teorema de Wilk para tener una distribución asintótica del estadístico del ratio de funciones de verosimilitudes puesto que el espacio de parámetros del modelo de la hipótesis nula no es subconjunto del de la alternativa. Tratar de construir un test de hipótesis implicaría pues deducir la distribución del estadístico test (o buscar uno del cual conociéramos la misma), lo cual es bastante complejo en el contexto de distribuciones normales multivariantes.

La región crítica para nuestro test con la hipótesis alternativa H_1 la construimos con ayuda del ratio de verosimilitudes, que gracias al lema de Neyman-Pearson tendremos la garantía de que nos ofrece la mejor discriminación posible.

$$\mathrm{RC} := \{\vec{x} : -2\log\Lambda \leq k\}, \quad \Lambda := \frac{\mathcal{L}_{H_0}}{\sup_{\theta\in\Theta}\mathcal{L}_{H_1}}. \tag{2.4}$$

Al haber escrito $-2\log\Lambda$ en vez de sencillamente el cociente de funciones de verosimilitud podemos usar el teorema de Wilk por el cual $-2\log\Lambda \sim \chi^2_p$, con p siendo la diferencia de grados de libertad entre H_0 y H_1, que en nuestro caso es el número de genes que consideremos para hacer el test. La función de verosimilitud de la hipótesis nula $\mathcal{L}_{H_0}$ está completamente determinada,

$$\mathcal{L}_{H_0} = \frac{1}{\sqrt{2\pi|V_{\mathrm{tri12}}|}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu}_{\mathrm{tri12}})^T V^{-1}_{\mathrm{tri12}}(\vec{x}-\vec{\mu}_{\mathrm{tri12}})}, \tag{2.5}$$

mientras que para la hipótesis alternativa debemos maximizar la función de verosimilitud obteniendo los valores máximo verosímiles de sus parámetros θ, que consiste al final en el vector de medias $\vec{\mu}$, puesto que asumimos que la matriz de covarianza es la misma, V_{tri12}.

$$\mathcal{L}_{H_1} = \frac{1}{\sqrt{2\pi|V_{\mathrm{tri12}}|}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^T V^{-1}_{\mathrm{tri12}}(\vec{x}-\vec{\mu})}. \tag{2.6}$$

Es inmediato ver que podemos simplificar $-2\log\Lambda$ como sigue,

$$\begin{aligned} -2\log\Lambda = &-2\log\frac{\mathcal{L}_{H_0}}{\sup_{\theta\in\Theta}\mathcal{L}_{H_1}} = 2\left(\log\left(\sup_{\theta\in\Theta}\mathcal{L}_{H_1}\right) - \log\mathcal{L}_{H_0}\right) = \ldots = \\ =&(\vec{x}-\vec{\mu}_{\text{tri12}})^T V_{\text{tri12}}^{-1}(\vec{x}-\vec{\mu}_{\text{tri12}}) - (\vec{x}-\hat{\vec{\mu}})^T V_{\text{tri12}}^{-1}(\vec{x}-\hat{\vec{\mu}}), \end{aligned} \tag{2.7}$$

siendo $\hat{\vec{\mu}}$ los parámetros resultados de maximizar $\mathcal{L}_{H_1}$. Este paso se implementó minimizando

$$(\vec{x}-\vec{\mu})^T V_{\text{tri12}}^{-1}(\vec{x}-\vec{\mu}), \tag{2.8}$$

que es equivalente a la maximización de $\mathcal{L}_{H_1}$.

Con estos ingredientes, para cada nueva entrada con lecturas de genes $\vec{x}$, podemos calcular el p-valor asociado al test de hipótesis calculando el valor de nuestro estimador como aparece en la ecuación (2.7) y usando la distribución χ^2 con un número de grados de libertad igual al de genes que tenemos. Podemos usar el p-valor tal y como si fuera la probabilidad dada por un modelo de aprendizaje automático, permitiéndonos compararlo con los modelos obtenidos según lo descrito en la sección 2.2.1, así como de los que hablaremos en la sección 2.2.3.

El proceso metodológico completo para obtener este modelo será pues el siguiente.

1. Normalizar las cuentas a la suma total.

2. Realizar una división entre entrenamiento y prueba de los datos.

3. Preselección de genes, quedándonos con aquellos de mayor número de lecturas esperadas para *single cell RNA-seq.*. El número se establece con el número máximo de genes que podemos tener para ser capaces de estimar tanto las medias muestrales como la matriz de covarianza muestral del conjunto de entrenamiento.

4. Estimación de los valores poblacionales $\vec{\mu}_{\text{tri12}}$ y de V_{tri12} a través de las medias muestrales $\vec{\overline{x}}$ y de la covarianza muestral S.

Con estos pasos, ya tenemos lo necesario para hacer comparaciones con el conjunto de prueba

y estimar el error de generalización. Este método, a través de la matriz de covarianza muestral, tiene en cuenta las correlaciones en la expresión génica de nuestra señal.

2.2.3. Imperativo pragmático

Como se comentó ya en la sección 2.2.1, uno de los problemas a la hora de tratar de entrenar un modelo de aprendizaje automático con el conjunto de pseudodatos (y también con el original) es la dificultad de ser capaces de tener una muestra representativa de datos para el conjunto de entrenamiento y para el conjunto de prueba, puesto que variando sin más la división entre ambas se aprecian diferencias significativas. Es de esperar que este problema se pudiera solucionar con, o bien un conjunto de datos más homogéneo, o uno mayor. Ambas cosas son complicadas, en el primer caso por la naturaleza de los datos en sí, secuenciaciones, que son prácticamente experimentos únicos cada uno de ellos, y en el segundo por la dificultad de encontrar secuenciaciones concretas para lo que queremos.

Sería esperable que un mayor número de datos, siempre que no agregaran mucha más heterogeneidad, nos permitiesen estimar correctamente un conjunto de prueba y de entrenamiento, invariantes (hasta cierto punto) frente a diferentes particiones del conjunto total. Por consiguiente, y con ánimo ilustrativo, además de los modelos de las secciones 2.2.1 y 2.2.2, mostraremos qué sucedería si construimos sendos discriminadores con el conjunto de datos como si fuera de entrenamiento, sin modificar los hiperparámetros (en el primer caso). Esto naturalmente no permite estimar el posible error de generalización, pero cabría esperar que las diferencias debido a usar un conjunto mayor de datos se volcaran en una mejor estimación (y más realista) de todo el conjunto de datos al completo.

2.2.4. Implementación informática

Técnicamente, todo el trabajo se hizo empleando `Python`, el entorno usual de cálculo numérico y científico compuesto por `Numpy`, `Scipy` y `Pandas`. Para entrenar los modelos de aprendizaje automático, se empleó `scikit-learn`. Todo esto se hizo en un ordenador con CPU AMD FX-8350 y 12 GB de RAM bajo un SO GNU/Linux que es Manjaro Linux (basado en Arch Linux).

3 Resultados

En esta sección detallamos las principales comparaciones y pruebas que se hicieron en este trabajo. Los epígrafes siguientes las contienen, mientras que el último compara y discute las mismas.

3.1. Modelo de aprendizaje automático

3.1.1. Intento inicial

En esta sección mostramos los modelos entrenados con el preprocesamiento detallado en la sección 2.2.1. Los primeros estudios comparando los hiperparámetros (haciendo una búsqueda en malla de los mismos) mostraron que se han de tomar los hiperparámetros más laxos posibles con el fin de evitar el sobreentrenamiento, es decir:

- Profundidad de árboles (`max_depth`): 1,
- Número máximo de características por árbol (`max_features`): 1,

dejando el resto tal y como estaban, y con *bagging* activado. Estos hiperparámetros serán usados en todos los modelos expuestos en los siguientes apartados. Con ellos, entrenamos modelos conservando cuatro porcentajes diferentes de la varianza en el PCA: 25, 60, 85 y 98 % y empleamos además cuatro semillas diferentes para la generación de números pseudoaleatorios (aunque es irrelevante la semilla en sí: 34, 151, 602, 151112).

En la figura 3.1 se pueden ver las curvas ROC para las cuatro semillas manteniendo un 98 % de la varianza total en el PCA, tanto para los conjuntos de entrenamiento como de prueba (con una separación 50 %-50 %), y las AUC que se obtienen en ambos casos.

Lo primero apreciable es que el modelo entrenado no es muy potente: apenas obtenemos valores en torno al 60 % de AUC. También podemos ver que, aunque cogimos los hiperparámetros más laxos que pudimos, tenemos cierto sobreentrenamiento, que se aprecia en las diferencias entre los valores de AUC que tenemos para las curvas de prueba y entrenamiento en los cuatro casos. En el mejor, la diferencia es de un 2 % en AUC. Las figuras también son reflejo de que el mero hecho de cambiar tan solo la semilla de generación de números pseudoaleatorios hacen que obtengamos resultados bastante diferentes en función a la misma.

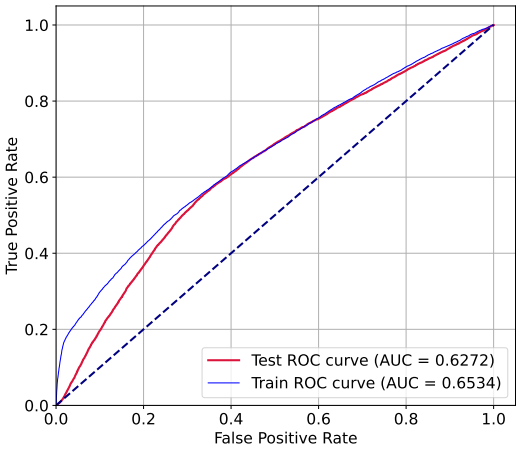
Quedándonos con un 85, 60 y 25 % obtenemos modelos con curvas ROC como las de las figuras 3.2, 3.3 y 3.4 respectivamente. Con pequeñas diferencias podemos afirmar lo mismo del caso del 98 % para estos nuevos modelos: potencia limitada, sobreentrenamiento, y diferencias significativas cuando variamos tan solo la semilla. También hay una mejora de rendimiento del modelo en términos de AUC al bajar al 25 % de varianza conservada por el PCA.

3.1.2. Eliminando casos extremos (*outliers*)

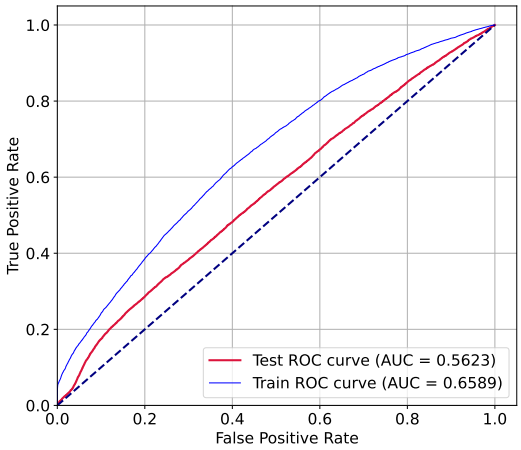
En las figuras 3.5, 3.6, 3.7 y 3.8 podemos ver el resultado de entrenar modelos tras eliminar aquellas pseudosecuenciaciones derivadas de los pacientes identificados como casos extremos, tal y como describimos en la sección 2.2.1. Los resultados son similares a los que se observan el caso anterior: variaciones notorias al cambiar la semilla, y pequeñas al cambiar la varianza preservada tras el PCA, excepto nuevamente al bajar al 25 %, donde se aprecia una mejoría de rendimiento.

3.1.3. Haciendo preselección de genes

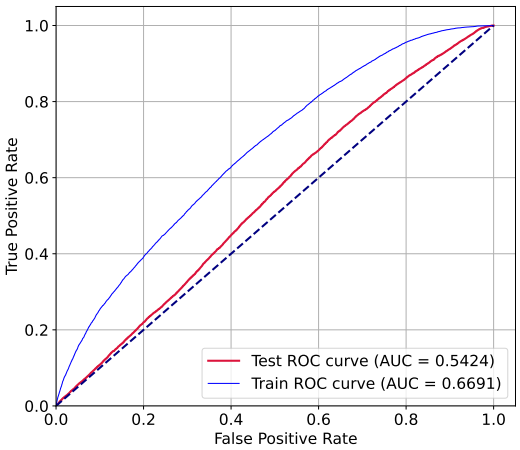
En las figuras 3.9, 3.10, 3.11 y 3.12 podemos ver el resultado de entrenar modelos tras hacer una preselección de genes en función a su poder discriminador individual tal y como describimos



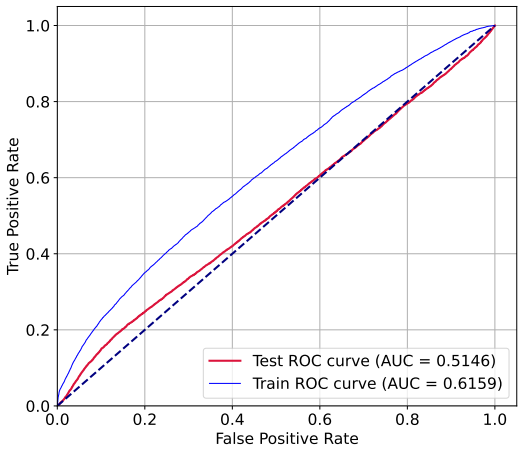
(a) Semilla 34.



(b) Semilla 151.

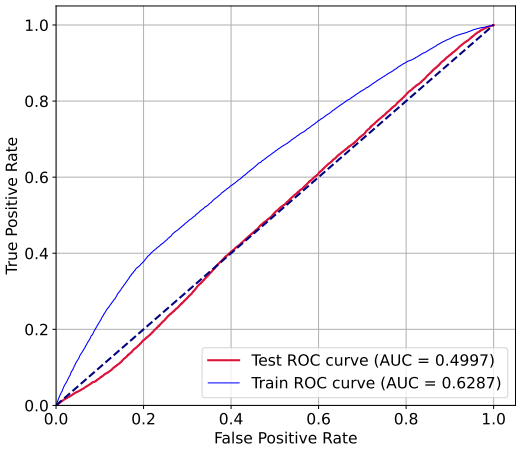


(c) Semilla 602.

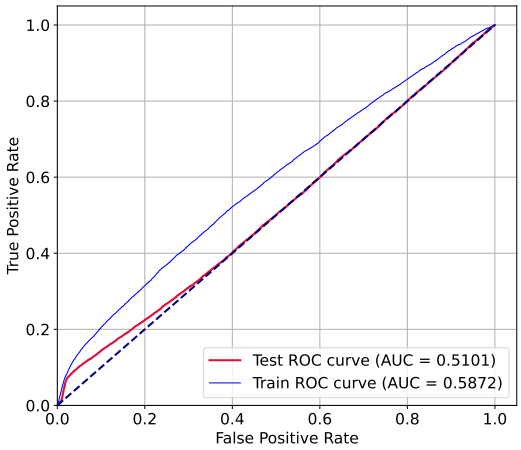


(d) Semilla 151112.

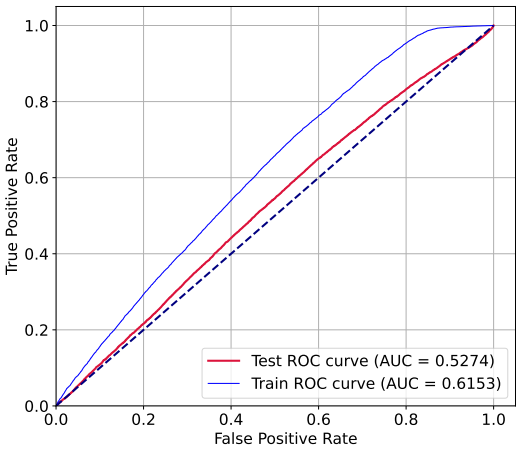
Figura 3.1: Curvas ROC para los conjuntos de prueba y entrenamiento con el modelo inicial conservando un 98 % de la varianza en el PCA para las cuatro semillas de generación de números pseudoaleatorios consideradas.



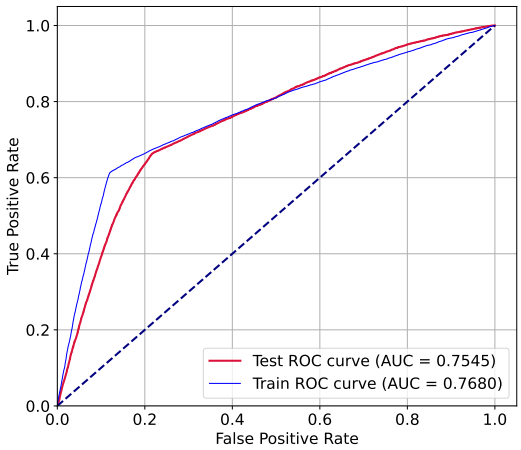
(a) Semilla 34.



(b) Semilla 151.

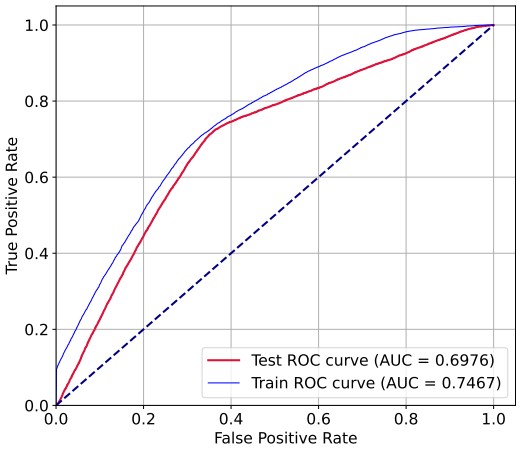


(c) Semilla 602.

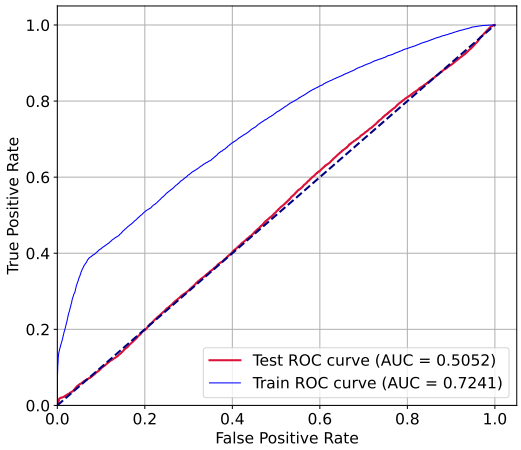


(d) Semilla 151112.

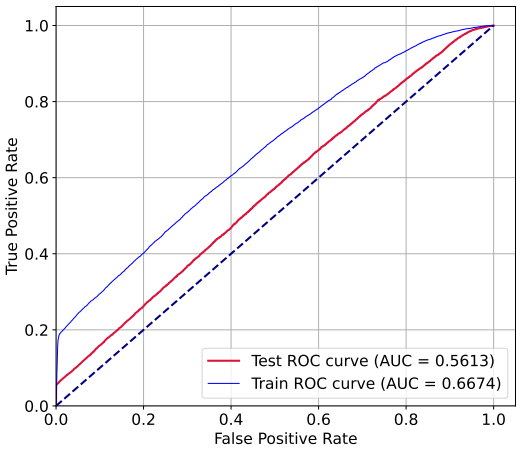
Figura 3.2: Curvas ROC para los conjuntos de prueba y entrenamiento con el modelo inicial conservando un 85 % de la varianza en el PCA para las cuatro semillas de generación de números pseudoaleatorios consideradas.



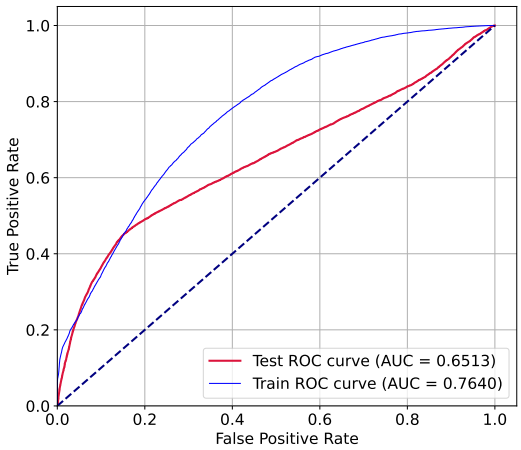
(a) Semilla 34.



(b) Semilla 151.

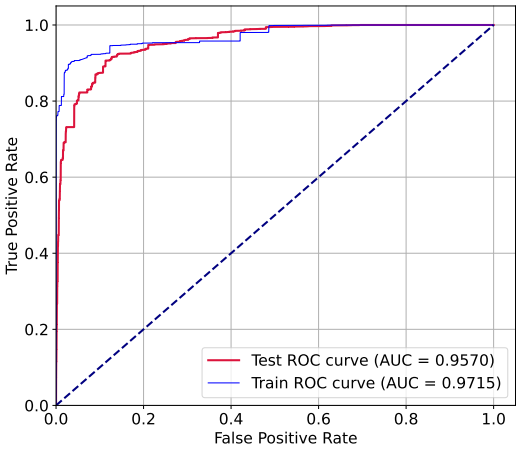


(c) Semilla 602.

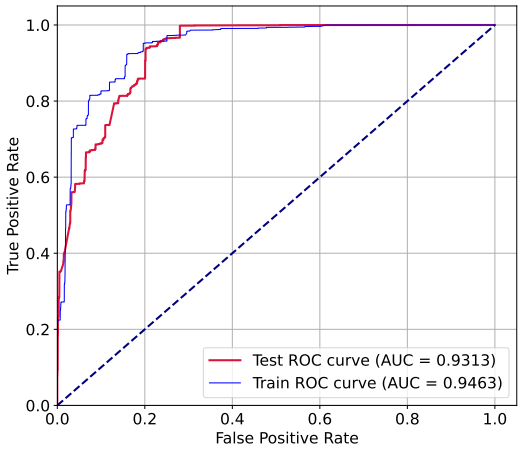


(d) Semilla 151112.

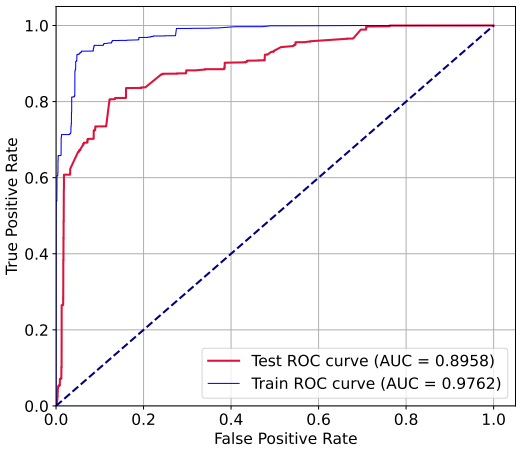
Figura 3.3: Curvas ROC para los conjuntos de prueba y entrenamiento con el modelo inicial conservando un 60 % de la varianza en el PCA para las cuatro semillas de generación de números pseudoaleatorios consideradas.



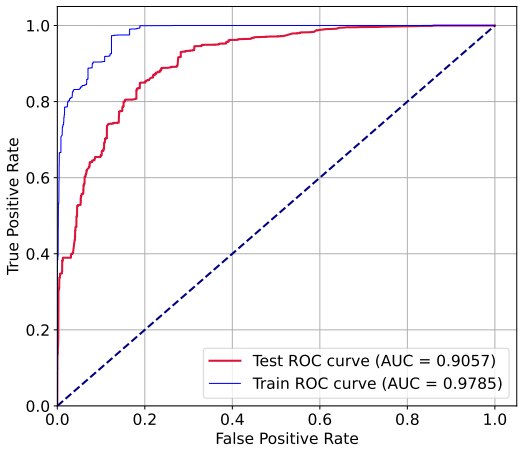
(a) Semilla 34.



(b) Semilla 151.

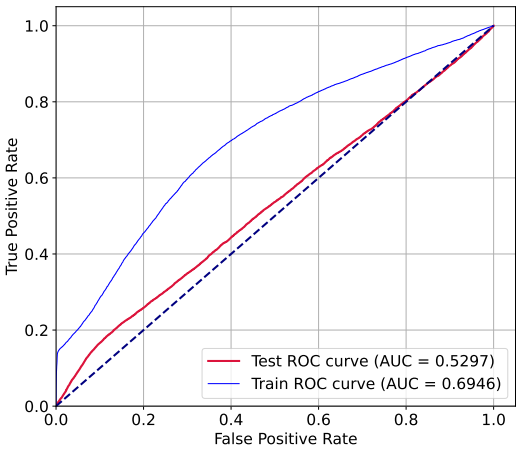


(c) Semilla 602.

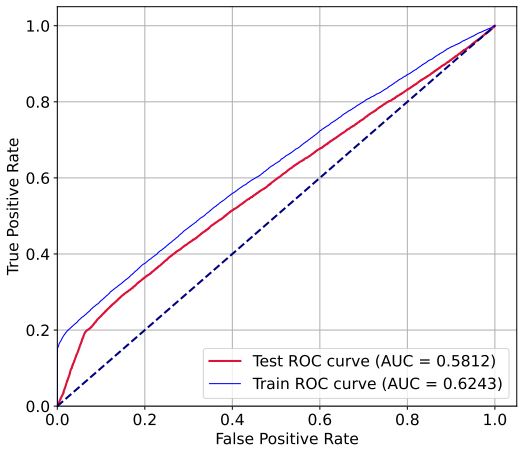


(d) Semilla 151112.

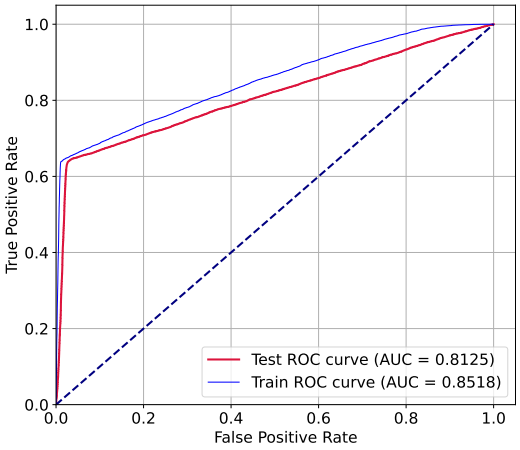
Figura 3.4: Curvas ROC para los conjuntos de prueba y entrenamiento con el modelo inicial conservando un 25 % de la varianza en el PCA para las cuatro semillas de generación de números pseudoaleatorios consideradas.



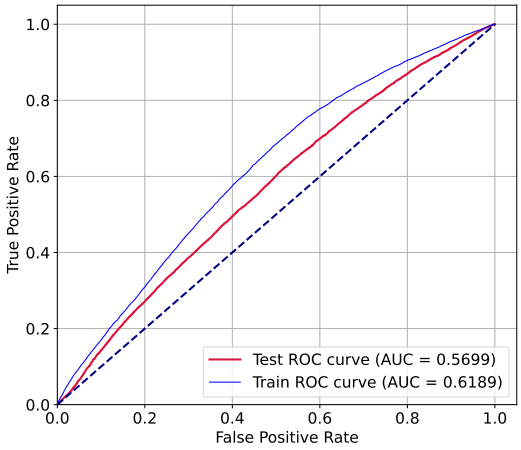
(a) Semilla 34.



(b) Semilla 151.

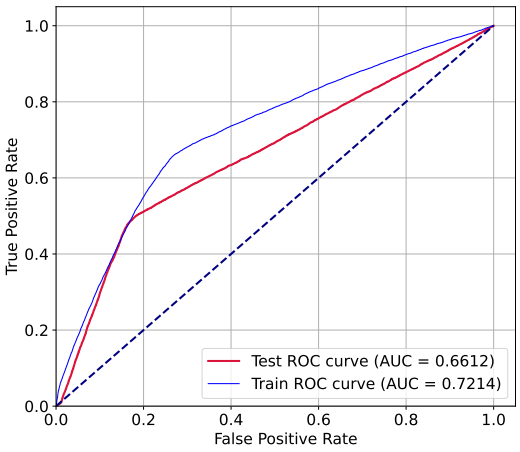


(c) Semilla 602.

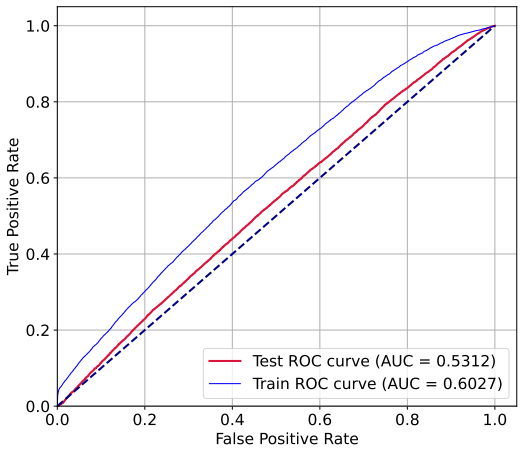


(d) Semilla 151112.

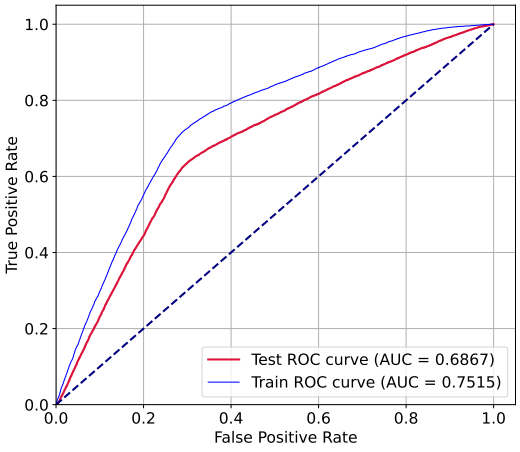
Figura 3.5: Curvas ROC para los conjuntos de prueba y entrenamiento conservando un 98 % de la varianza en el PCA para las cuatro semillas de generación de números pseudoaleatorios consideradas y tras la eliminación de las pseudosecuenciaciones asociadas a casos extremos.



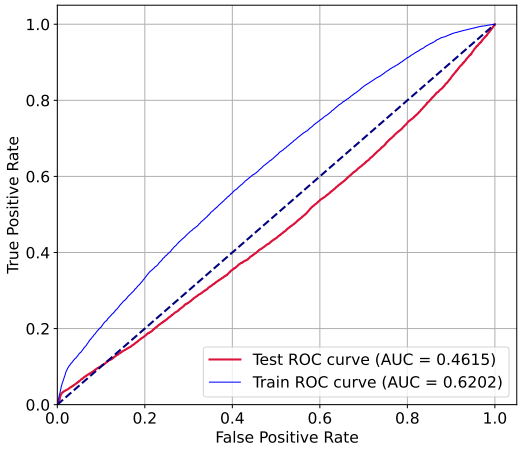
(a) Semilla 34.



(b) Semilla 151.

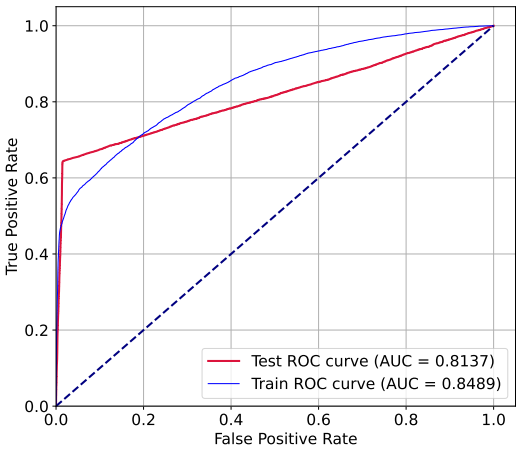


(c) Semilla 602.

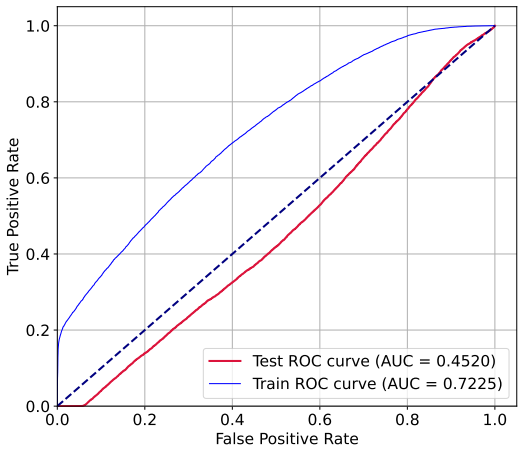


(d) Semilla 151112.

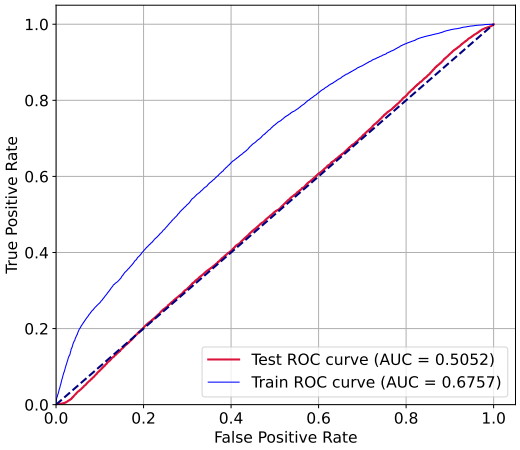
Figura 3.6: Curvas ROC para los conjuntos de prueba y entrenamiento conservando un 85 % de la varianza en el PCA para las cuatro semillas de generación de números pseudoaleatorios consideradas y tras la eliminación de las pseudosecuenciaciones asociadas a casos extremos.



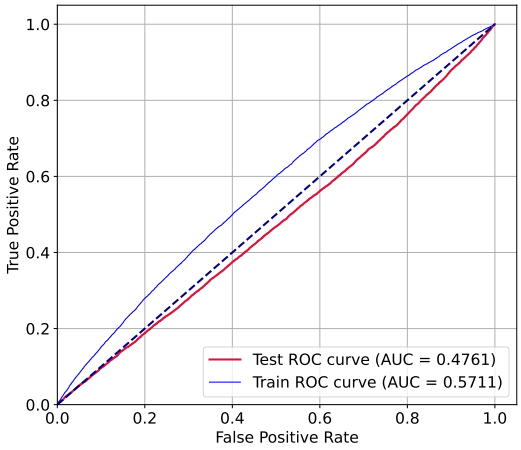
(a) Semilla 34.



(b) Semilla 151.

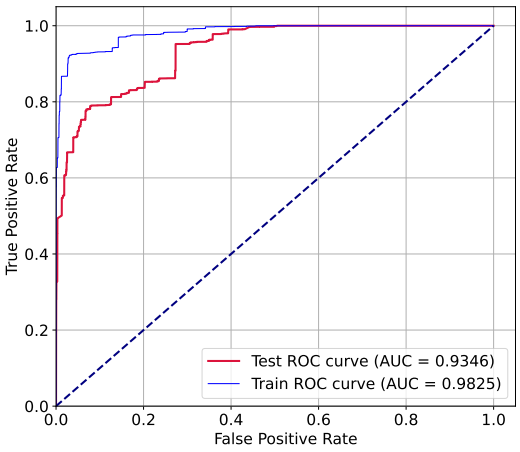


(c) Semilla 602.

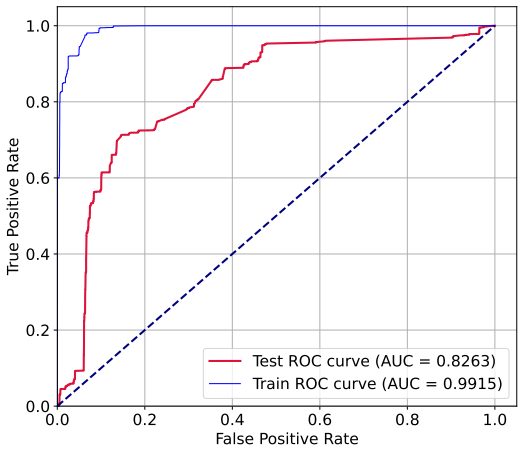


(d) Semilla 151112.

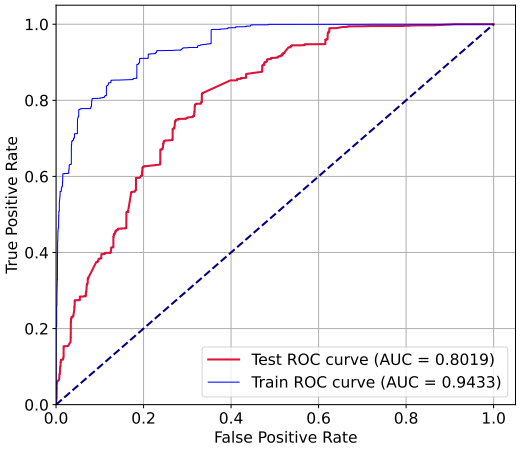
Figura 3.7: Curvas ROC para los conjuntos de prueba y entrenamiento conservando un 60 % de la varianza en el PCA para las cuatro semillas de generación de números pseudoaleatorios consideradas y tras la eliminación de las pseudosecuenciaciones asociadas a casos extremos.



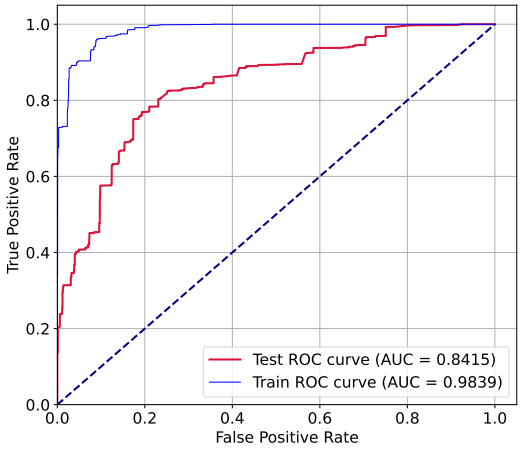
(a) Semilla 34.



(b) Semilla 151.



(c) Semilla 602.



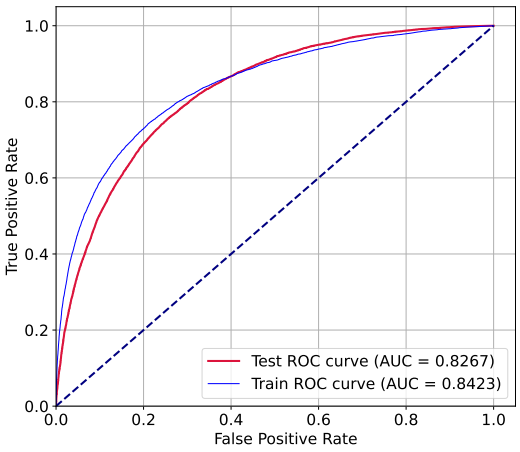
(d) Semilla 151112.

Figura 3.8: Curvas ROC para los conjuntos de prueba y entrenamiento conservando un 25 % de la varianza en el PCA para las cuatro semillas de generación de números pseudoaleatorios consideradas y tras la eliminación de las pseudosecuenciaciones asociadas a casos extremos.

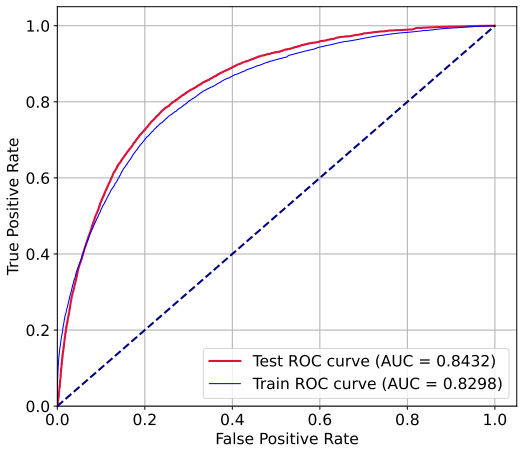
en la sección 2.2.1. Los resultados, al igual que los anteriores, muestran variabilidad en función a la semilla empleada. Sin embargo, la preselección nos permite obtener modelos bastante más potentes, con AUC ∼0.85–0.95, con muy pequeñas variaciones en función a la varianza preservada por el PCA. También observamos cierto sobreentrenamiento (e.g. con la semilla 602 en la figura 3.9), que depende de la semilla escogida. También se observan fenómenos extraños en los cuales el conjunto de prueba tiene mejor rendimiento que el de entrenamiento, lo que induce a pensar, como se mencionó en la sección 2.2.1, que no tenemos datos suficientes para tener dos estimaciones funcionales de los conjuntos de prueba ni de entrenamiento, aún usando una distribución 50 %-50 % de los datos.

3.1.4. Eliminando casos extremos y haciendo preselección de genes

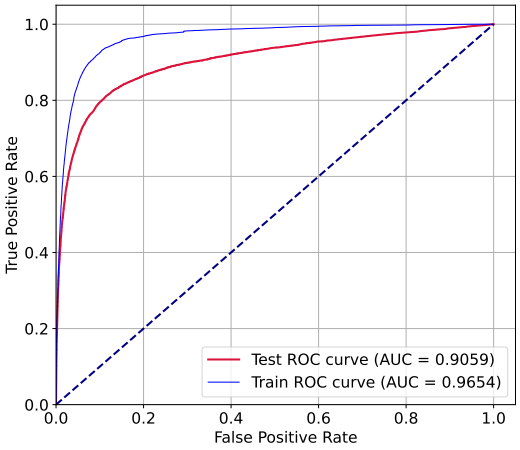
En las figuras 3.13, 3.14, 3.15 y 3.16 podemos ver el resultado de entrenar modelos tras hacer la preselección de genes de la sección anterior y además la eliminación de casos extremos. Los resultados son bastante similares a los obtenidos haciendo únicamente la preselección de genes, con pequeñas variaciones al agregar la eliminación de casos extremos.



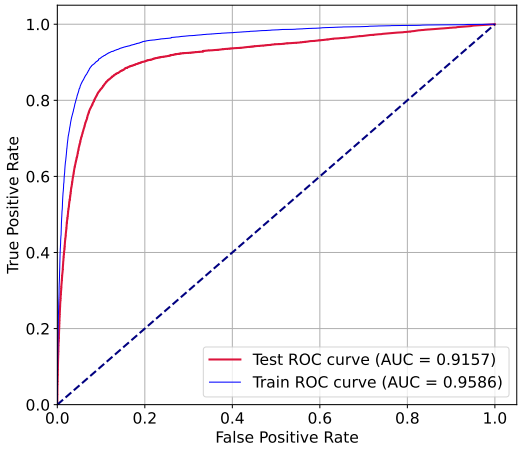
(a) Semilla 34.



(b) Semilla 151.

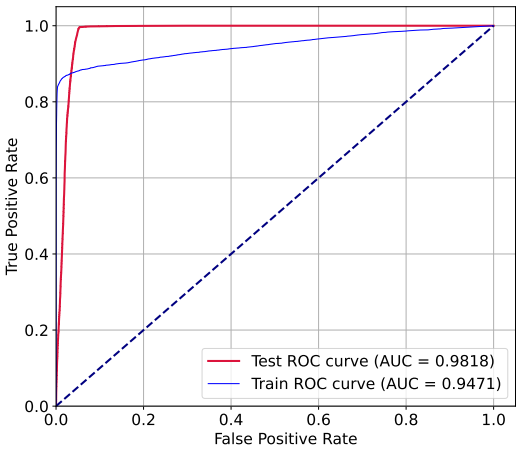


(c) Semilla 602.

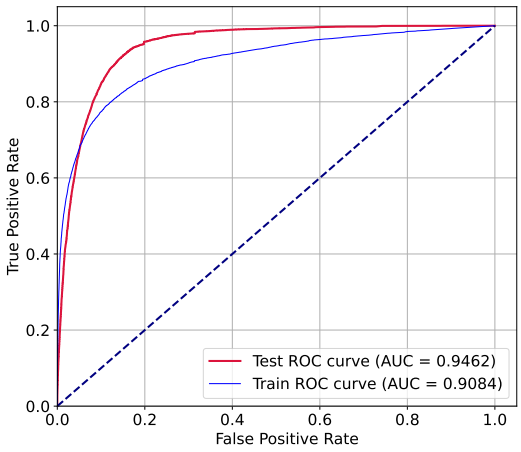


(d) Semilla 151112.

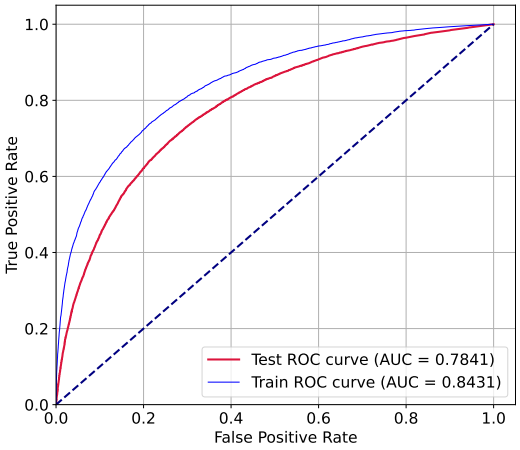
Figura 3.9: Curvas ROC para los conjuntos de prueba y entrenamiento conservando un 98 % de la varianza en el PCA para las cuatro semillas de generación de números pseudo-aleatorios consideradas y tras una preselección de genes basada en su poder discriminador individual.



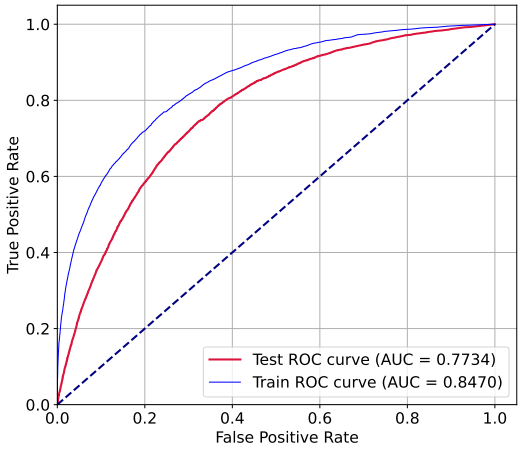
(a) Semilla 34.



(b) Semilla 151.

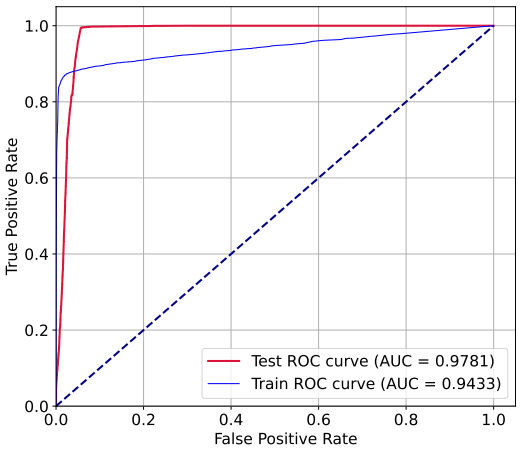


(c) Semilla 602.

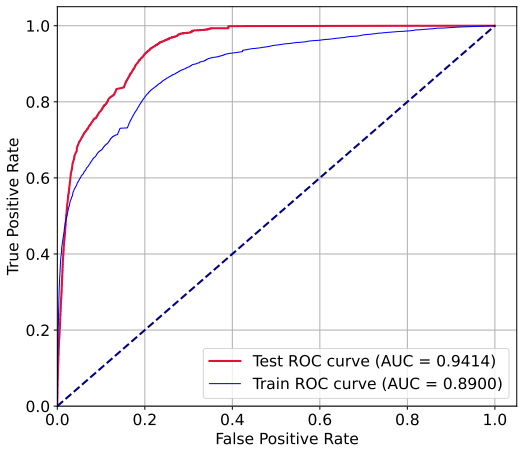


(d) Semilla 151112.

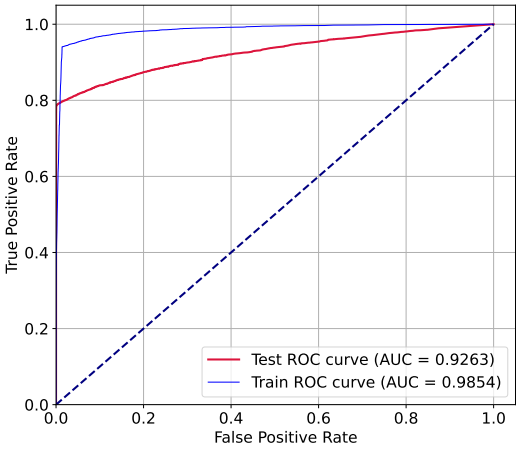
Figura 3.10: Curvas ROC para los conjuntos de prueba y entrenamiento conservando un 85 % de la varianza en el PCA para las cuatro semillas de generación de números pseudo-aleatorios consideradas y tras una preselección de genes basada en su poder discriminador individual.



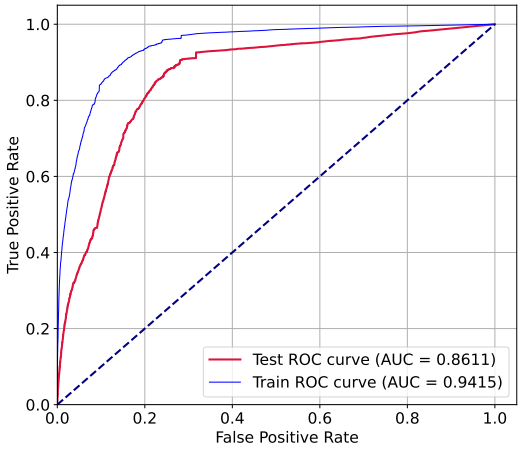
(a) Semilla 34.



(b) Semilla 151.

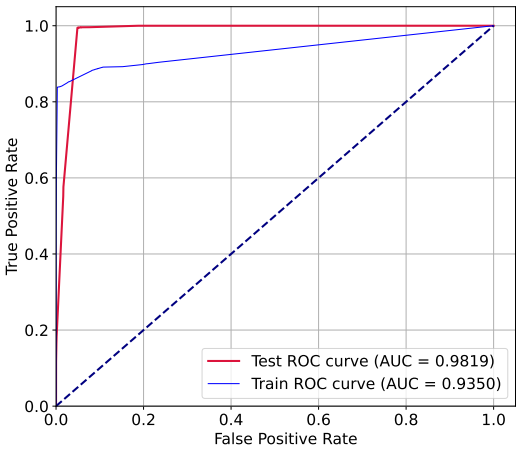


(c) Semilla 602.

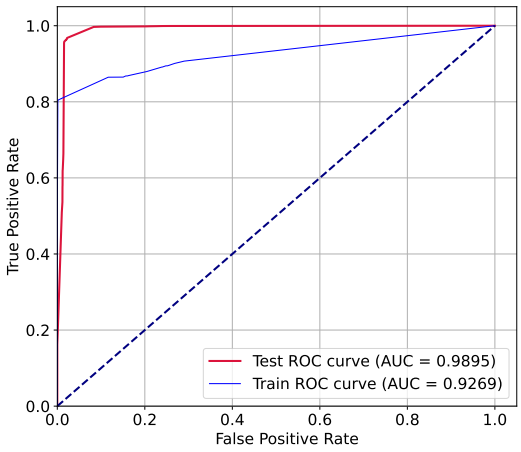


(d) Semilla 151112.

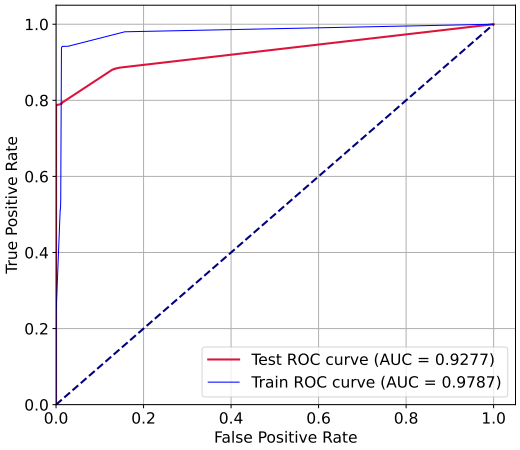
Figura 3.11: Curvas ROC para los conjuntos de prueba y entrenamiento conservando un 60 % de la varianza en el PCA para las cuatro semillas de generación de números pseudo-aleatorios consideradas y tras una preselección de genes basada en su poder discriminador individual.



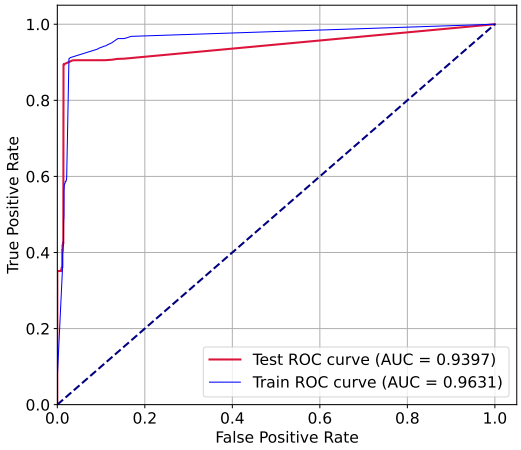
(a) Semilla 34.



(b) Semilla 151.

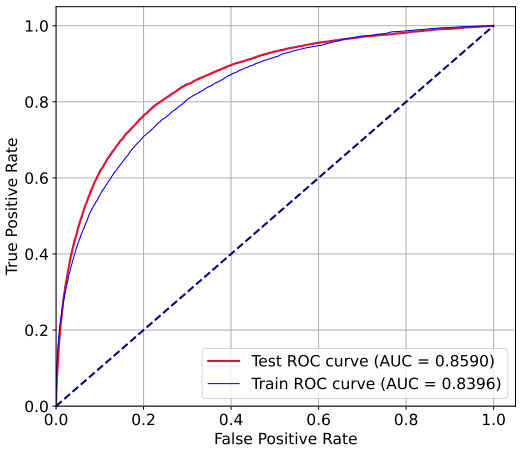


(c) Semilla 602.

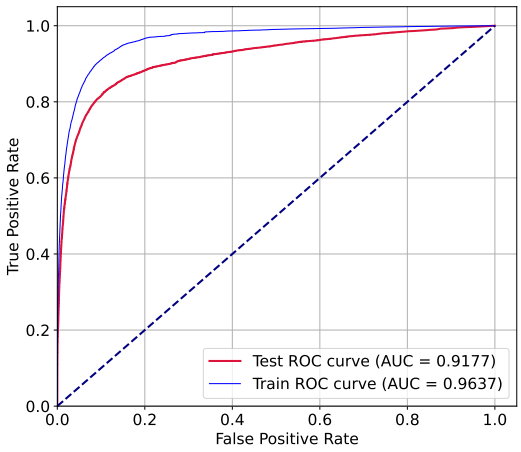


(d) Semilla 151112.

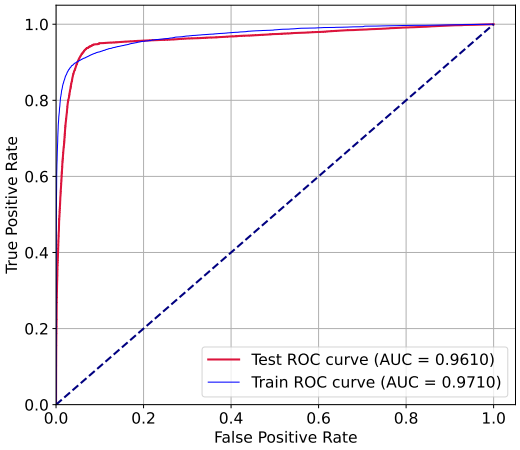
Figura 3.12: Curvas ROC para los conjuntos de prueba y entrenamiento conservando un 25 % de la varianza en el PCA para las cuatro semillas de generación de números pseudo-aleatorios consideradas y tras una preselección de genes basada en su poder discriminador individual.



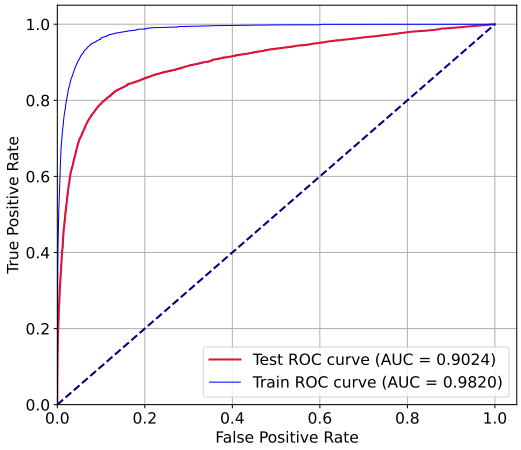
(a) Semilla 34.



(b) Semilla 151.

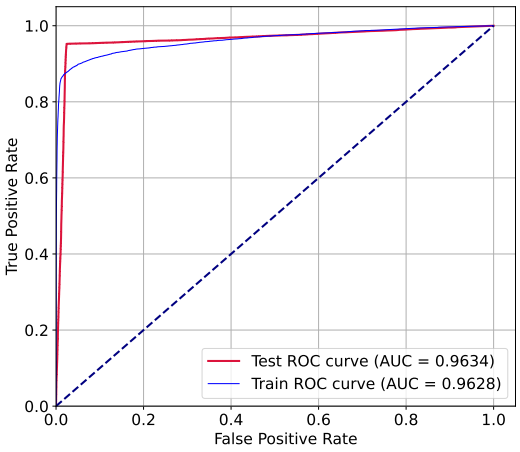


(c) Semilla 602.

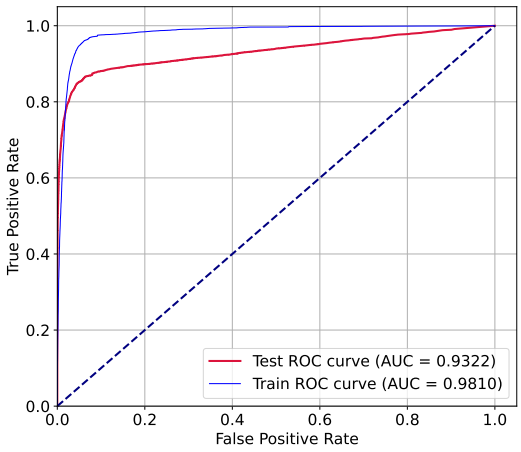


(d) Semilla 151112.

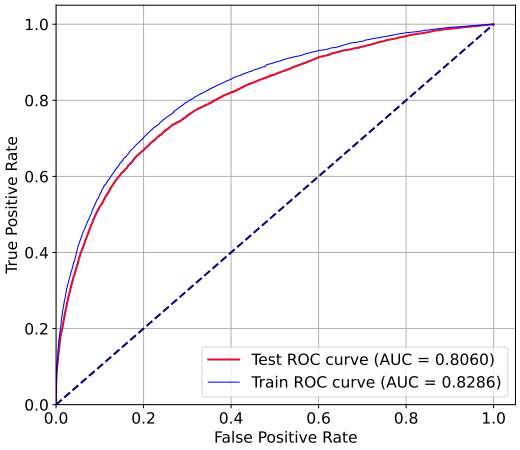
Figura 3.13: Curvas ROC para los conjuntos de prueba y entrenamiento conservando un 98 % de la varianza en el PCA para las cuatro semillas de generación de números pseudoaleatorios consideradas y tras eliminar casos extremos y una preselección de genes basada en su poder discriminador individual.



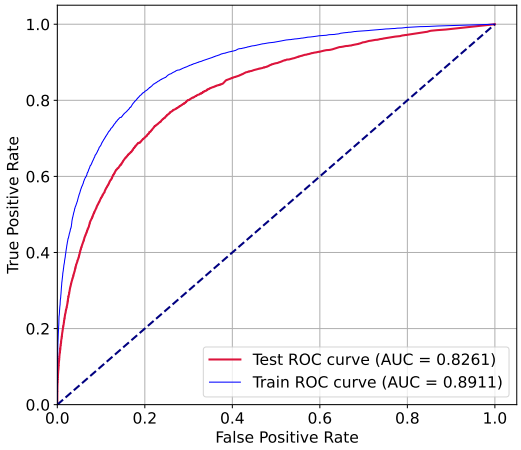
(a) Semilla 34.



(b) Semilla 151.

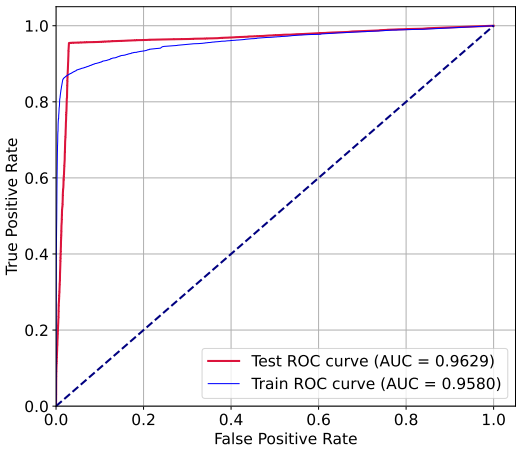


(c) Semilla 602.

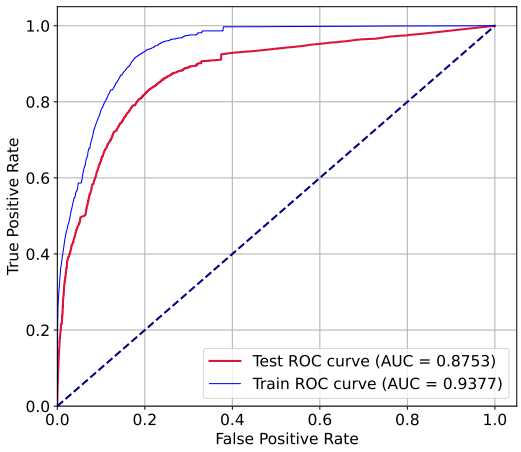


(d) Semilla 151112.

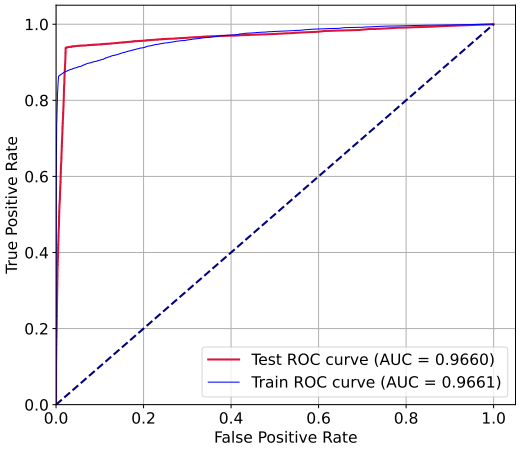
Figura 3.14: Curvas ROC para los conjuntos de prueba y entrenamiento conservando un 85 % de la varianza en el PCA para las cuatro semillas de generación de números pseudoaleatorios consideradas y tras eliminar casos extremos y una preselección de genes basada en su poder discriminador individual.



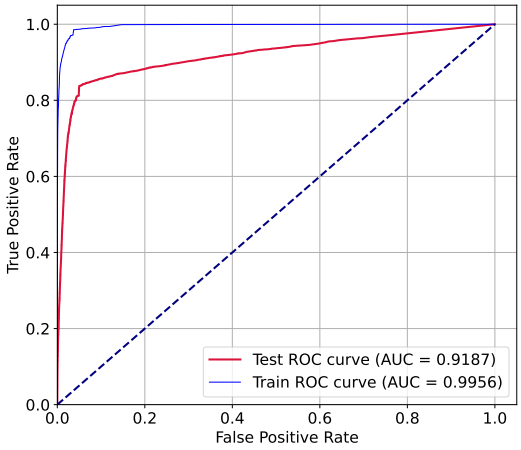
(a) Semilla 34.



(b) Semilla 151.

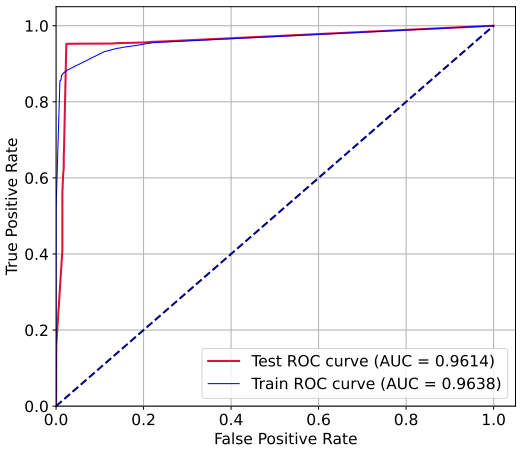


(c) Semilla 602.

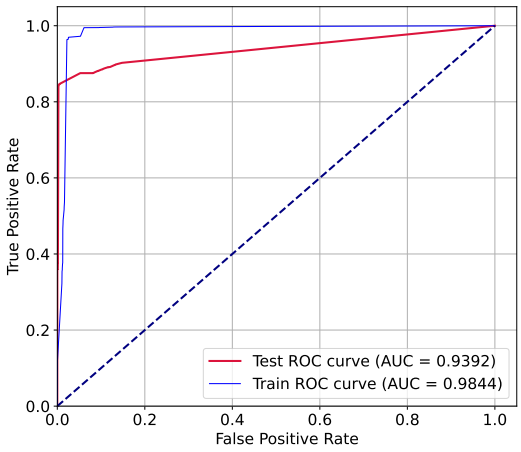


(d) Semilla 151112.

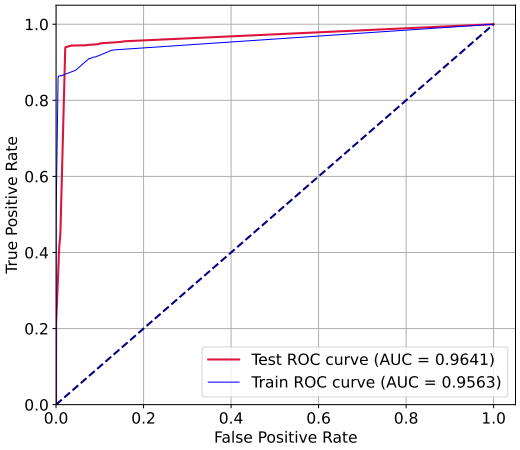
Figura 3.15: Curvas ROC para los conjuntos de prueba y entrenamiento conservando un 60 % de la varianza en el PCA para las cuatro semillas de generación de números pseudoaleatorios consideradas y tras eliminar casos extremos y una preselección de genes basada en su poder discriminador individual.



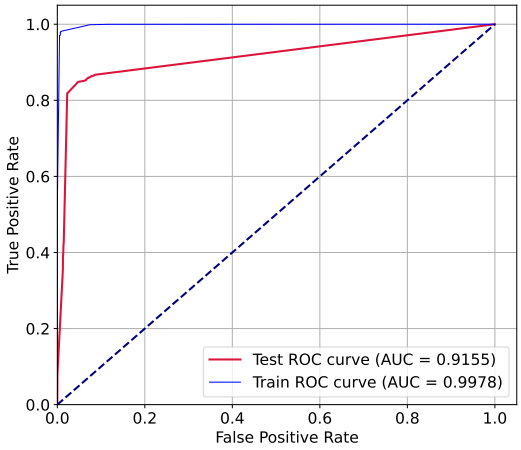
(a) Semilla 34.



(b) Semilla 151.



(c) Semilla 602.



(d) Semilla 151112.

Figura 3.16: Curvas ROC para los conjuntos de prueba y entrenamiento conservando un 25 % de la varianza en el PCA para las cuatro semillas de generación de números pseudoaleatorios consideradas y tras eliminar casos extremos y una preselección de genes basada en su poder discriminador individual.

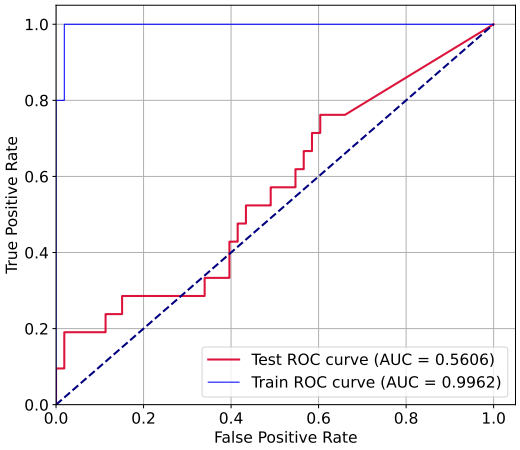
3.2. Test de hipótesis

En la figura 3.17 se pueden observar, para las cuatro semillas diferentes de generación de números pseudoaleatorios anteriores, las curvas ROC con sus AUC correspondientes para los tests de hipótesis construidos con un reparto 50-50 % de los datos en conjuntos de entrenamiento-prueba. Con el fin de reducir el número de características de forma que tengamos siempre suficientes secuenciaciones para poder estimar la matriz de covarianza, nos quedamos con aquellos genes que tienen al menos 120 lecturas, lo que nos deja con 16 genes.

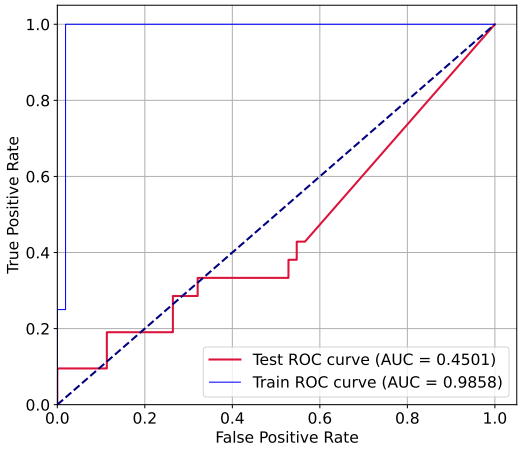
Como se observa, existe una gran diferencia entre el efecto sobre el conjunto de prueba y el de entrenamiento, habiendo claro sobreentrenamiento, mucho más que el apreciable en los modelos de aprendizaje automático. Esto es parcialmente entendible por la necesidad de modelizar completamente las distribuciones de los genes, a partir del conjunto de entrenamiento. En la figura 3.18 se puede ver lo mismo pero con un reparto 25-75 %, dándole un 75 % al conjunto de entrenamiento y un 25 % al de prueba.

Este reparto es mejor, pero sigue habiendo un alto sobreentrenamiento. También podemos apreciar, al igual que en el caso de los modelos de la figura 3.17, que hay variaciones en función a la semilla empleada, si bien quizás no sean tan significativas. Finalmente, en la figura 3.19 se muestra un caso más extremo en el que el conjunto de prueba usa un 10 % de los pacientes y el resto se usan para entrenar.

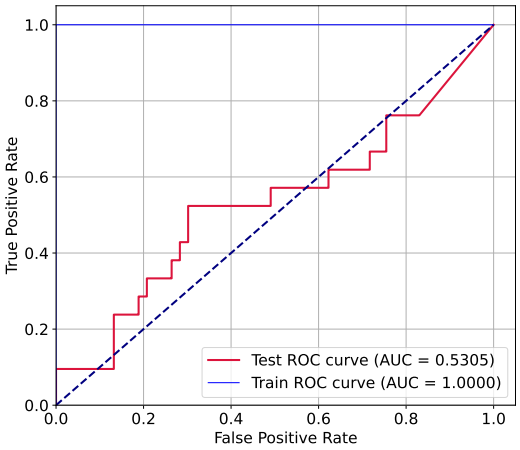
De nuevo, podemos acusar cierto sobreentrenamiento, aunque en este caso el reducido tamaño del conjunto de prueba suscita fluctuaciones grandes al cambiar las semillas. Comparando las tres figuras, podemos ver cómo de importante es el número de datos para hacer estas estimaciones, particularmente en el conjunto de entrenamiento, puesto que varía sustancialmente cuando usamos el 50 % del conjunto de datos para entrenamiento (con un sobreajuste completo) y los modelos menos sobreentrenados de usar el 75 y el 90 % de los datos. Así mismo, vemos aún diferencias en las curvas al usar distintas semillas en los tres casos considerados, aunque estas son pequeñas comparadas con lo que vemos en los modelos de aprendizaje automático, para los casos 50-50 % y 25-75 %. Las fluctuaciones del caso extremo 10-90 % se pueden asociar directamente a la baja estadística en el conjunto de prueba.



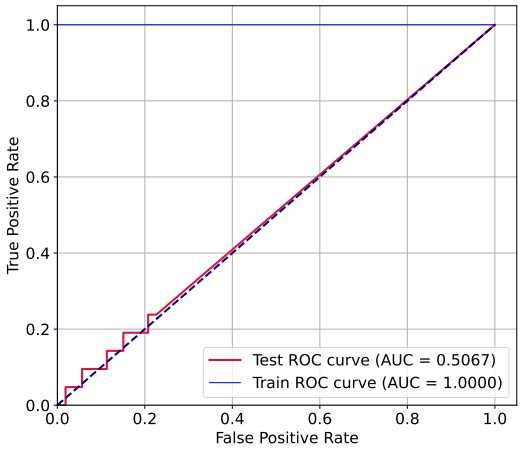
(a) Semilla 34.



(b) Semilla 151.

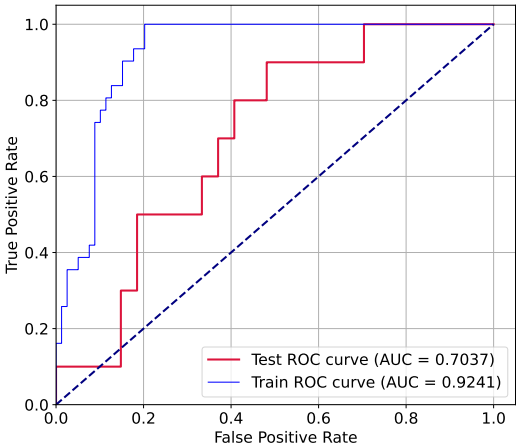


(c) Semilla 602.

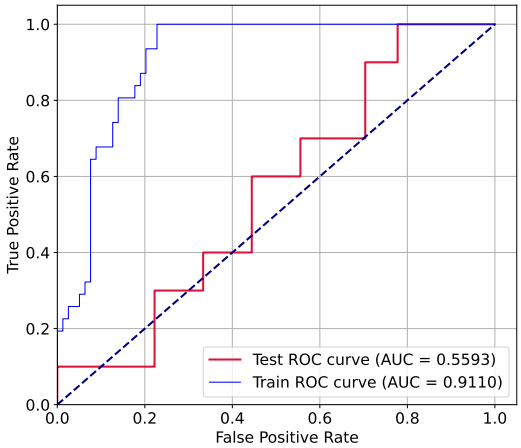


(d) Semilla 151112.

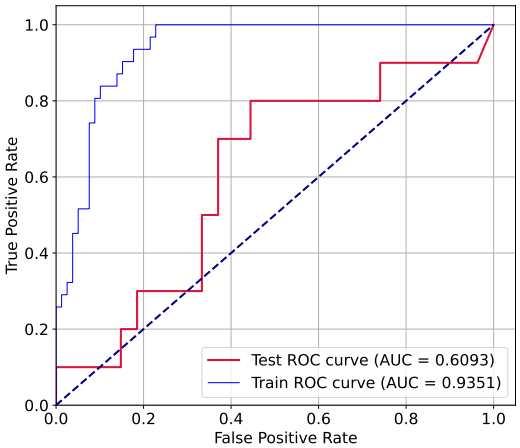
Figura 3.17: Curvas ROC para los conjuntos de prueba y entrenamiento de los tests de hipótesis usando un reparto 50-50 % (prueba-entrenamiento).



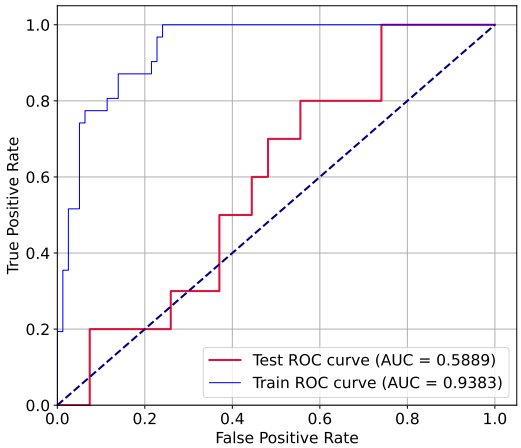
(a) Semilla 34.



(b) Semilla 151.

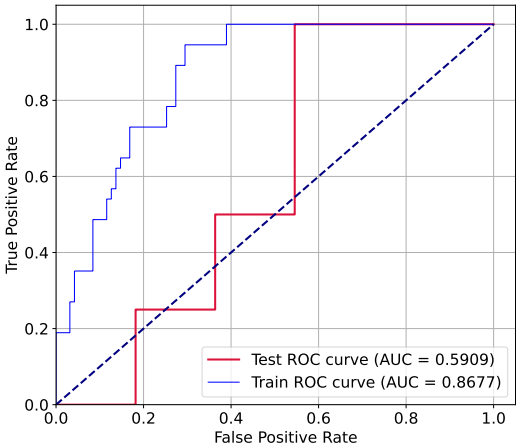


(c) Semilla 602.

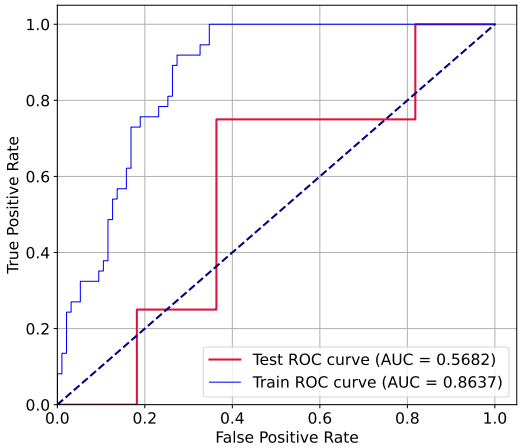


(d) Semilla 151112.

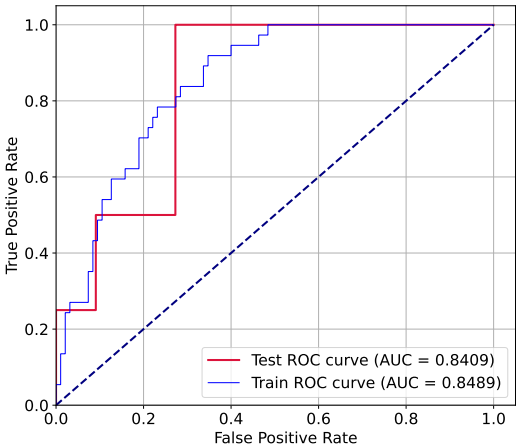
Figura 3.18: Curvas ROC para los conjuntos de prueba y entrenamiento de los tests de hipótesis usando un reparto 25-75 % (prueba-entrenamiento).



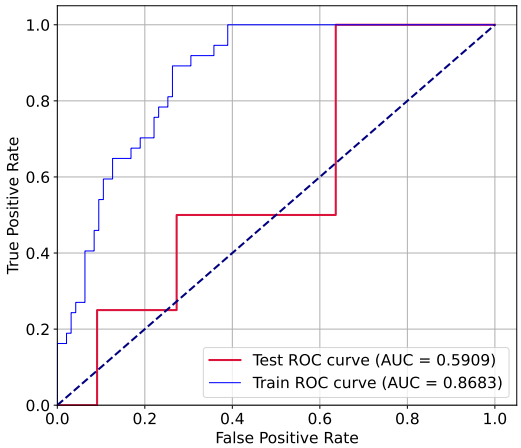
(a) Semilla 34.



(b) Semilla 151.



(c) Semilla 602.

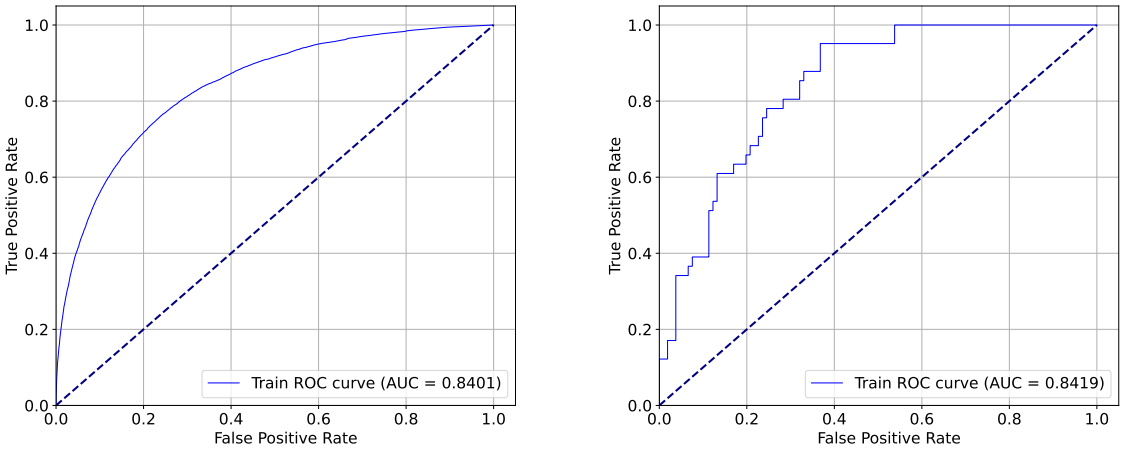


(d) Semilla 151112.

Figura 3.19: Curvas ROC para los conjuntos de prueba y entrenamiento de los tests de hipótesis usando un reparto 10-90 % (prueba-entrenamiento).

3.3. Modelos empleando todo el conjunto de datos

Como se explicó en la sección 2.2.3, se entrenó un modelo de aprendizaje automático aprovechando todos los datos posibles, y también se construyó otro test de hipótesis de esta misma manera. Las curvas ROC de ambos se pueden ver en la figura 3.20. En el caso del modelo de aprendizaje automático, construido sin eliminar *outliers* (debido a que su efecto es muy pequeño), haciendo la preselección de genes, y conservando un 98 % de la varianza en el PCA obtenemos una AUC de 0.84 y una precisión del 72 %. Para el test de hipótesis, se obtiene una AUC de 0.84 con una precisión del 78 % (tomando 0.5 como punto de corte para el p-valor).



(a) Modelo de aprendizaje automático.

(b) Test de hipótesis.

Figura 3.20: Curvas ROC para el modelo de aprendizaje automático y el test de hipótesis hechos usando el conjunto de datos total y la semilla 34.

3.4. Comparación y discusión

Las comparaciones del apartado 3.1 muestran que entrenar un modelo de aprendizaje automático, aún teniendo generadas un número elevado de pseudosecuenciaciones de Montecarlo, sigue siendo un desafío debido al tamaño reducido de datos del conjunto original. Esto se plasma en las sistemáticas diferencias que se observan, desde los intentos iniciales, entre considerar unas u otras semillas para separar el conjunto de datos en prueba y entrenamiento. A pesar de tener una separación 50-50 %, esto sigue sin ser suficiente para tener una representación correcta y estable de los datos en ambos subconjuntos.

Otra derivada del tamaño pequeño de los datos y la alta dimensionalidad es el sobreentrenamiento que está presente en casi todas las pruebas. Tan solo algunos casos muestran un sobreajuste pequeño, pero la gran fluctuabilidad ya mencionada en función a cómo se construyen los conjuntos de prueba y entrenamiento hace que no podamos fiarnos de ellos. Esta doble condición (baja cantidad de datos, y muchas características), que nos forzó a establecer hiperparámetros del modelo lo menos detallados posibles, nos niega el uso de modelos más complejos (como redes neuronales). Otros modelos, como SVM o clasificadores ingenuos de Bayes, también fueron probados con peores o similares resultados que los aquí mostrados: en el caso del *naive Bayes classifier* se obtenían AUC entre 0.8 y 0.9 con un muy notable sobreentrenamiento (mayor que con *random forest*) y con la variabilidad que vemos dependiendo de la semilla. Para el SVM (entrenado usando el hiperparámetro *probability*, que agrega una validación cruzada), los resultados son similares a los mejores que obtenemos con otros modelos enseñados, con AUC que llegan más allá de 0.9 incluso. Sin embargo, los mismos problemas que con el clasificador ingenuo de Bayes y los modelos enseñados anteriormente de sobreentrenamiento y de variabilidad con la semilla siguen presentes.

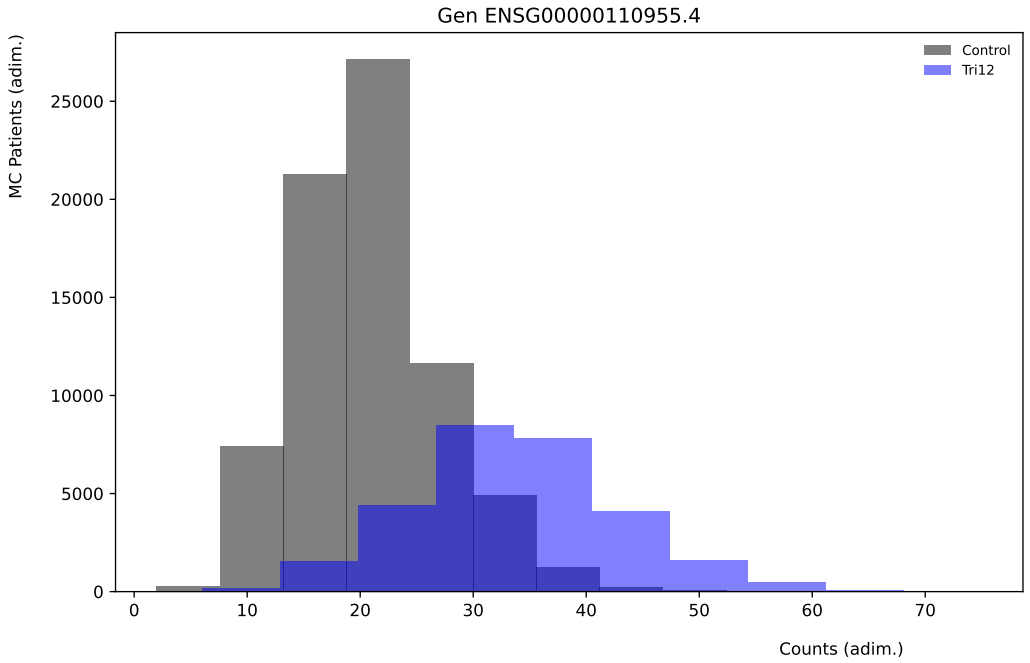
Las problemáticas no se consiguen eliminar con los dos principales añadidos a la metodología usada. La eliminación de casos extremos u *outliers*, que podría potencialmente justificar las variaciones entre semillas, no es satisfactoria, pues estas permanecen, con pequeñas diferencias al quitarlos (con el coste de reducción de estadística que además lleva aparejada). En cuanto a la preselección de genes, sí se mejoran notoriamente los modelos en cuanto a rendimiento, si bien seguimos manteniendo el sobreentrenamiento (a pesar de la reducción en dimensionalidad) y las problemáticas de variación en función a la semilla.

En todas las situaciones (más notoriamente cuando no se aplica la preselección), hay que reducir mucho la varianza preservada por el PCA para condensar la información y, a pesar del sobreentrenamiento, construir un modelo más potente. Esto puede deberse a la falta de datos para realizar correctamente el PCA, o apuntar a que los genes están muy descorrelacionados entre sí.

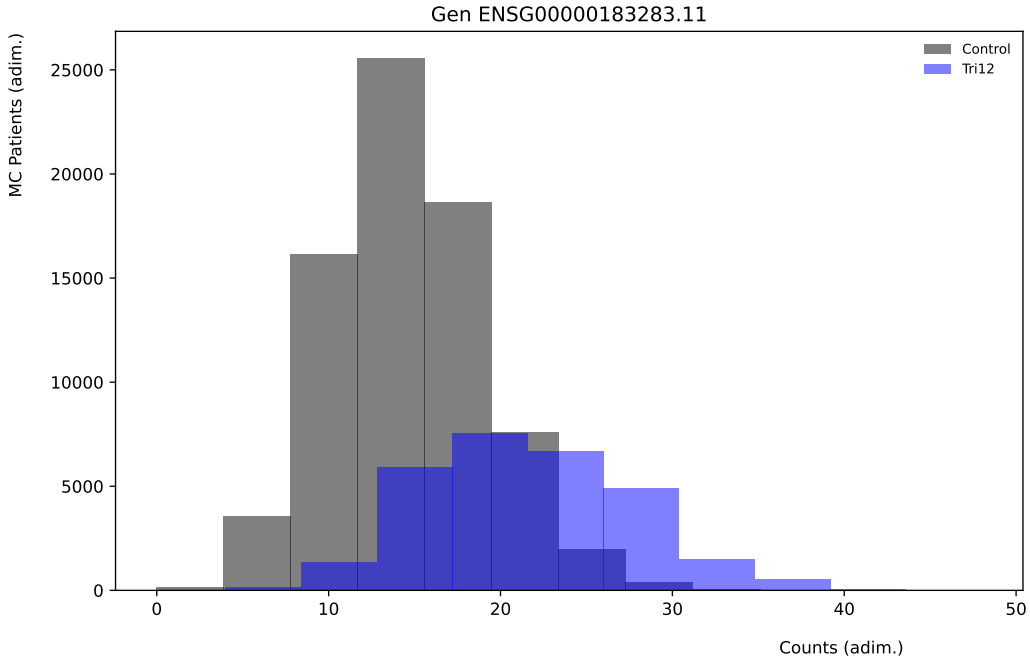
En el caso de los tests de hipótesis, tampoco obtenemos resultados ideales, observando sobreentrenamiento en todas las situaciones, si bien las variaciones en función a la semilla son bastante menos drásticas (excepto en el último caso de reparto 10-90 % prueba-entrenamiento, de la figura 3.19). La potencia de los modelos es elevada en el conjunto de prueba, comparable con los mejores obtenibles de aprendizaje automático, pero vemos cómo es necesario dotarnos de muchos datos con el fin de estimar correctamente el gran número de parámetros que necesitamos para trabajar con tests de hipótesis (medias muestrales y matriz de covarianza muestral).

Finalmente, los modelos entrenados con el conjunto de datos completo arrojan resultados parecidos, tanto para el de aprendizaje automático como para el de test de hipótesis, con AUC de $\sim$0.8 y exactitudes en torno al 75 %. Estos valores de esas figuras de mérito se asemejan a los que algunos modelos de aprendizaje automático anteriores obtuvieron, aunque como siempre, dependiendo de la semilla.

En todo caso, los resultados de las distintas aproximaciones al problema muestran inequívocamente que existen diferencias en la expresión génica, observables con un número de lecturas del orden de scRNA-seq., que permiten diferenciar las situaciones de mutación (tri12) de aquellas que no. Esto se puede comprobar en la figura 3.21 se muestran dos histogramas con cuentas normalizadas de los pseudodatos donde hay poder discriminatorio de forma directa en esos genes, a pesar de la reducción en lecturas.



(a)



(b)

Figura 3.21: Histogramas de los datos totales separados según tengan la mutación de la trisomía del cromosoma 12 o no para los genes ENSG00000110955.4 y ENSG00000183283.11 usando pseudosecuenciaciones.

4 Conclusiones

En este trabajo hemos estudiado un conjunto de datos provenientes de pacientes con leucemia linfática crónica (LLC) consistente en el número de cuentas para cada gen que se obtienen en secuenciaciones en bloque de ARN mensajero (*bulk RNA-seq*). A través de la creación de un número elevado de secuenciaciones ficticias de Montecarlo (basadas en números aleatorios), hemos sido capaces de simular un conjunto de datos parejo que se obtendría a través de secuenciación individual de ARN mensajero (*single cell RNA-seq*), con el fin de ver si es posible obtener un modelo de aprendizaje automático que identifique mutaciones con datos de ese tipo de secuenciaciones. Este estudio se hizo en primera instancia para discriminar una mutación muy reconocible: la trisomía del cromosoma 12.

Los resultados, detallados en la sección 3, muestran que en tales conjuntos de datos existen diferencias entre los genes que permiten discriminar entre las situaciones mutadas y las que no, lo que indica que potencialmente el objetivo del trabajo podría cumplirse. Sin embargo, los resultados obtenidos muestran que si bien es posible entrenar modelos de aprendizaje automático, el reducido número de datos reales acaba afectando al conjunto de pseudosecuenciaciones: no es posible tener con estos datos una representación fehaciente y estable de los conjuntos de entrenamiento y de prueba. Esto (i) impide evaluar correctamente el error de generalización de los modelos entrenados y (ii) pone en duda hasta cierto punto el rendimiento de un modelo obtenido, puesto que una redefinición de los conjuntos de prueba/entrenamiento (moviendo entradas de uno a otro), le afecta. Los resultados en cuanto a rendimiento de los modelos obtenidos tienen $\sim$0.8-0.9 de AUC y en torno al 75 % de exactitud, con cierto sobreentrenamiento dependiente de la definición de los conjuntos prueba/entrenamiento.

Con el fin de evitar este problema, se construyó un test de hipótesis que pudiera tener en cuenta la información de las correlaciones entre la distribución de los genes. Para ello se construyó un

test de hipótesis que sin embargo adolece de un sobreentrenamiento grave, puesto que es necesario precisar muchos más parámetros (medias muestrales y matriz de covarianza muestral) que en otras situaciones, lo que lo hace más dependiente de la cantidad de datos que se tenga, como se comprobó en la sección 3.2.

Habida cuenta de que los problemas de los que adolecemos tienden a arreglarse con un conjunto de datos mayor, se hizo un entrenamiento de un modelo de aprendizaje automático con el conjunto entero de datos, sin estimación de error de generalización, para comparar los resultados con los anteriores. Tal y como se explica en la sección 3.3, se obtienen resultados en figuras de mérito similares a los anteriores ($\sim$0.8 AUC y $\sim$75 % exactitud). Esto también se hizo con el test de hipótesis, con resultados similares. Por desgracia, no podemos evaluar el error de generalización en estos últimos modelos. De esta forma, si bien hemos confirmado que existen diferencias que pueden permitir diferenciar la mutación escogida (trisomía del cromosoma 12), se ha probado que no es posible con esta cantidad de datos construir un clasificador fiable.

El camino más directo para obtener un modelo así sería incrementar el conjunto de datos. Esto no es en absoluto sencillo, puesto que requiere hacer secuenciaciones de pacientes aquejados de LLC. Sin embargo, una mayor cantidad de estadística permitiría probablemente estabilizar los conjuntos de prueba y entrenamiento y otorgaría fiabilidad al clasificador. También permitiría estudiar otras alternativas más complejas para explotar mejor la información guardada de las secuenciaciones.

Además, con el fin de reducir el sobreentrenamiento, se podrían explorar opciones de modelos más sencillos. De hecho, las soluciones que existen para el estudio de la expresión diferencial de ARNm apuestan por modelos lineales generalizados (GLM) para representar el conjunto de datos, como los paquetes de R `edgeR` (Robinson y col., 2009) o `DESeq` (Anders & Huber, 2010). Estos modelos tienen la ventaja, para construir tests de hipótesis, de que se reducen mucho el número de parámetros a estimar, a través de ciertas hipótesis y simplificaciones, lo que permite mejorar los ajustes con menor cantidad de datos.

5 Bibliografía

Abrams, Z. B., Johnson, T. S., Huang, K., Payne, P. R. O. & Coombes, K. (2019). A protocol to evaluate RNA sequencing normalization methods. *BMC Bioinformatics*, *20*(24), 679. https://doi.org/10.1186/s12859-019-3247-x

Anders, S. & Huber, W. (2010). Differential expression analysis for sequence count data. *Nature Precedings*, 1-1.

Auer, P. L. & Doerge, R. W. (2011). A two-stage Poisson model for testing RNA-seq data. *Statistical applications in genetics and molecular biology*, *10*(1).

Bosch, F. & Dalla-Favera, R. (2019). Chronic lymphocytic leukaemia: from genetics to treatment. *Nature Reviews Clinical Oncology*, *16*(11), 684-701. https://doi.org/10.1038/s41571-019-0239-8

Breiman, L. (1998). Arcing classifier (with discussion and a rejoinder by the author). *The Annals of Statistics*, *26*(3), 801-849. https://doi.org/10.1214/aos/1024691079

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Carlberg, C. & Velleuer, E. (2021). *Cancer Biology: How Science Works* (First edition). Springer.

Gierliński, M., Cole, C., Schofield, P., Schurch, N. J., Sherstnev, A., Singh, V., Wrobel, N., Gharbi, K., Simpson, G., Owen-Hughes, T., Blaxter, M. & Barton, G. J. (2015). Statistical models for RNA-seq data derived from a two-condition 48-replicate experiment. *Bioinformatics*, *31*(22), 3625-3630. https://doi.org/10.1093/bioinformatics/btv425

Gilbert, W. & Maxam, A. (1973). The Nucleotide Sequence of the lacOperator. *Proceedings of the National Academy of Sciences*, *70*(12), 3581-3584. https://doi.org/10.1073/pnas.70.12.3581

Konishi, T. (2005). A thermodynamic model of transcriptome formation. *Nucleic Acids Research*, *33*(20), 6587-6592. https://doi.org/10.1093/nar/gki967

Li, X., Cooper, N. G. F., O'Toole, T. E. & Rouchka, E. C. (2020). Choice of library size normalization and statistical methods for differential gene expression analysis in balanced two-group comparisons for RNA-seq studies. *BMC Genomics*, *21*(1), 75. https://doi.org/10.1186/s12864-020-6502-7

Lodish, H. F. (2016). *Molecular cell biology* (Eighth edition.). W.H. Freeman.

McCarthy, D. J., Chen, Y. & Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, *40*(10), 4288-4297. https://doi.org/10.1093/nar/gks042

Nadeu, F., Clot, G., Delgado, J., Martín-García, D., Baumann, T., Salaverria, I., Beà, S., Pinyol, M., Jares, P., Navarro, A., Suárez-Cisneros, H., Aymerich, M., Rozman, M., Villamor, N., Colomer, D., González, M., Alcoceba, M., Terol, M. J., Navarro, B., ... Campo, E. (2018). Clinical impact of the subclonal architecture and mutational complexity in chronic lymphocytic leukemia. *Leukemia*, *32*(3), 645-653. https://doi.org/10.1038/leu.2017.291

Robinson, M. D., McCarthy, D. J. & Smyth, G. K. (2009). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139-140. https://doi.org/10.1093/bioinformatics/btp616

Robinson, M. D. & Smyth, G. K. (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, *23*(21), 2881-2887. https://doi.org/10.1093/bioinformatics/btm453

Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A. & Waterston, R. H. (2017). DNA sequencing at 40: past, present and future. *Nature*, *550*(7676), 345-353. https://doi.org/10.1038/nature24286

Svensson, V., Vento-Tormo, R. & Teichmann, S. A. (2018). Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols*, *13*(4), 599-604. https://doi.org/10.1038/nprot.2017.149

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K. & Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, *6*(5), 377-382. https://doi.org/10.1038/nmeth.1315

Khan Academy. (s.f.). Overview of transcription [[En línea; consulta el día 14-04-22]].

Wikipedia, the free encyclopedia. (s.f.). Single cell sequencing [[En línea; consulta los días 14-04-22]].

Zhang, H., Xu, J., Jiang, N., Hu, X. & Luo, Z. (2015). PLNseq: a multivariate Poisson lognormal distribution for high-throughput matched RNA-sequencing read count data. *Statistics in Medicine*, *34*(9), 1577-1589. https://doi.org/https://doi.org/10.1002/sim.6449