Visualización de la calidad de los datos de RNA-Seq relacionados con el sistema inmune y el cáncer visualizados mediante shiny.

Adrián Ferrer Torres
Máster en Bioinformática y Bioestadística
Análisis de datos ómicos con shiny

Nombre Consultor: Ricardo Gonzalo Sanz
Nombre Profesores responsables de la asignatura: Carles Ventura Royo y Jose Antonio Morán Moreno

2 de Enero de 2018

02/01/2018

Título del trabajo:	*Visualización de la calidad de los datos de RNA-Seq relacionados con el sistema inmune y el cáncer visualizados mediante shiny.*
Nombre del autor:	*Adrián Ferrer Torres*
Nombre del consultor/a:	*Ricardo Gonzalo Sanz*
Nombre del PRA:	*Jose Antonio Morán Moreno*
Fecha de entrega (mm/aaaa):	02/2018
Titulación::	*Máster en Bioinformática y Bioestadística*
Área del Trabajo Final:	*Análisis de datos ómicos*
Idioma del trabajo:	*castellano*
Palabras clave	*Shiny, RNA-Seq, inmunogenómica, expresión génica, control de calidad*

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

Debido al alza en la incidencia del cáncer en la población, ha llegado a ser la segunda causa de muerte en el mundo. Por ello, el estudio del sistema inmune relacionado con esta enfermedad es prioritario, investigar en qué falla y cómo se modifica la expresión de las células linfocíticas. Así se ganaría información acerca de la evasión inmune llegando a poder reconocer posibles checkpoints o dianas farmacológicas en las que actuar con diferentes tipos de fármacos, inmunoterapias u otras terapias personalizadas posibles.

Para poder analizar la expresión génica es preciso primero realizar un pre-filtrado, control y normalización de los resultados obtenidos por las muestras de un estudio RNA-seq. Esto es lo que se pretende en este estudio, mostrando además las diferentes etapas de los resultados obtenidos con gráficos y tablas en una aplicación hecha con el paquete shiny.

Las muestras se obtuvieron de la página web de gene expression omnibus. Se utilizó el paquete edgeR así como DT entre otros para la construcción de la aplicación.

Además se incluye en la memoria: el informe del presupuesto para un pequeño laboratorio y centro de investigación que secuencie el genoma de los clientes y que realice también otros estudios derivados de las ciencias ómicas de otros centros como hospitales, institutos y centros de investigación.

Abstract (in English, 250 words or less):

Due to the increasing incidence of cancer in the population, it has become the second cause of death worldwide. Due to this, the study of the relation of the immune system with cancer disease is critical to research in what fails and how the expression on lymphocytic cells is modified. Therefore information about immune evasion would be reached leading to possible checkpoints or pharmacological target in which we could treat with different types of drugs, immunotherapy, or other therapies as personalized therapy.

In order to conduct a study of genetic expression it is a must to first pre-filter, control and normalize the results obtained samples obtained from a RNA-Seq experiment. This is what is intended in this study, showing also the different stages of results obtained with graphics and tables in an application made with the shiny package.

The samples were obtained from the gene expression omnibus webpage. The edgeR and DT packages were used for the application construction.

It is as well included in the memory a report of the budget for a small lab and research center that sequences the client's genomes as well as other omic studies from other centers as hospitals, institutes and research centers.

Índice

Lista de figuras e ilustraciones

No se encuentran elementos de tabla de ilustraciones.

Lista de tablas

1. Introducción

1.1 Contexto y justificación del Trabajo

1.1.01 El Cáncer

El Cáncer es uno de los principales problemas de salud del mundo causando alta mortalidad y morbilidad. Además de la creciente incidencia del cáncer en la población[1], es la segunda causa de muerte en el mundo. Además se prevé que la incidencia del cáncer aumente un 70% en los próximos 20 años. En 2015 el cáncer fue responsable de 8,8 millones de muertes. Y en 2010 representó una pérdida económica de 1,16 trillones de dólares. (www.who.int).

Hipócrates fue uno de los primeros en describirlo sin embargo la primera vez que carcinoma fue mencionado en la historia data del año 3000 antes de Cristo en Egipto en el papiro de Edwin Smith. Desde la época romana se empezó a tratar el cáncer con cirugía (50 años antes de Cristo), Aulus Celsus, Romano, escribió *De Medicina*.[2]

Algunos de los descubrimientos sobre el cáncer son anteriores al año 1863 cuando Rudolf Virchow empezó a considerar el origen celular del cáncer. Algunos años más tarde, en 1889 Stephen Paget remarcó una relación y el concepto de metástasis a otros órganos llamando a la hipótesis como hipótesis de la semilla y tierra.[3]

Hoy en día se consideran los tumores malignos como una agrupación de células que han proliferado en tejidos localizados perdiendo su función en el tejido (anaplasia), crecientes y dividiéndose sin control. Cuando las células invaden y crecen en otros órganos o tejidos adyacentes esparciéndose de su localización inicial es cuando debemos usar el término metástasis.[4] Los tumores son tratados con quimioterapia y cirugía pero: ¿es suficiente?

1.1.02 Tratamientos contra el cáncer

La Cirugía existe desde la época de los romanos, en 1809 Ephraim McDowell extrajo un tumor de ovario sin anestesia. Más tarde el Dr.

Bernard Fisher mostró que remover el tumor no hacía nada sin el uso de además: quimioterapia o radioterapia. Con la llegada de los rayos X en 1895 por Roentgen permitió a Pierre y Marie Curie descubrir el radio con los que usados conjuntamente se combatieron cánceres de cabeza y cuello en 1928. También se usó teleterapia de cobalto como radioterapia en 1950.

El concepto de Vigilancia inmune fue incluido en 1909 por Paul Ehrlich quien sugirió que las nuevas células transformadas que aparecían en el cuerpo podrían ser eliminadas por el escaneo del sistema inmune antes del crecimiento del tumor [5].

Ehrlich también acuñó la expresión “balas mágicas” para referirse al concepto de usar fármacos para los receptores celulares. Los primeros fármacos prometedores usados contra el cáncer fueron el nitrógeno mostaza para lymphomas en 1943 y los antagonistas del ácido fólico contra la leucemia en 1948[3]. Thomas y Burnet propusieron la idea de antígenos específicos de tumores (STA en inglés)[5]. Tras el éxito de la quimioterapia en la década de 1960 para la leucemia en niños y enfermedad de Hodgking en adultos, los profesionales de la sanidad empezaron a utilizar químicos en adición a la radioterapia y a la cirugía. Durante esta década también se descubrió que la inmunidad celular a cánceres era mucho más importante que la inmunidad humoral, una creencia que sigue siendo cierta hoy en día[3].

Las estrategias para eliminar el cáncer por cirugía seguido de radioterapia o quimioterapia han mostrado éxito pero en algunos casos han llevado a una reaparición de cáncer o formación de metástasis. Además la toxicidad inespecífica de estos tratamientos a las células cancerígenas también afecta a las células sanas huésped.

Hoy en día hay ensayos clínicos con distintos tipos de inmunoterapia antitumoral así como para las alergias; y existen terapias para el asma con glucocorticoides y se ha visto que también pueden influir positivamente en el cáncer[6]. Para poder abordar verdaderamente el problema es necesario estudiar la expresión génica de las células afectadas por el microambiente tumoral así como por las interacciones de las células provenientes de los nódulos linfáticos.

¿Pero se necesitaría acaso un tratamiento más personal y con más datos en los que basarse en función de la expresión génica de cada uno? Esta cuestión indica la importancia de los estudios de NGS.

1.1.03 El sistema inmunitario

Es importante entender los mecanismos que intervienen en la regulación y defensa del cáncer en nuestro propio organismo como es el sistema inmunitario y la función que ejerce de inmunovigilancia, concepto acuñado en 1909 por Paul Ehrlich, quien sugería que las células transformadas podían ser eliminadas por el sistema inmune antes de que apareciese el crecimiento del tumor.

La expresión de los genes medida en “counts” en determinadas circunstancias asociadas al cáncer y otras enfermedades que comparten características similares, como es el caso del asma y la inflamación, pueden darnos pistas de los genes que cobran importancia en la regulación de procesos como la inflamación y del cáncer más en general. Es importante también conocer la inmunogenómica o genes que se expresan y son relativos al sistema inmune o que al expresarse lo afectan de alguna manera.

Un proceso como el de inmunovigilancia como muchos otros procesos de los que operan dentro del sistema inmunitario en condiciones normales podría verse afectado y desregulado en situaciones tumorales.

La inmunosupresión dentro del microambiente tumoral ocurre como resultado de mecanismos tumorales y productos tales como los Factores solubles derivados de tumor como IL-10, TGF- β, factor de crecimiento vascular endotelial (VEGF), producidas por células inmaduras mieloides (iMCs) y fosfatidilserina soluble (sPS) producidas por las células tumorales[7]. Es importante entender todos los factores que están implicados en la inmunosupresión del microambiente tumoral para poder tratar con por ejemplo inmunoterapia.

Así como de posibles tratamientos como la implicación de los glucocorticoides con efectos anti tumorales.

Si se encontrasen genes relacionados con el sistema inmunitario (o con la interferencia de alguno de los procesos) darían una pista de cómo se expresan en ciertas condiciones y si están regulados positiva o negativamente.

1.1.04 El término inmunovigilancia

La especificidad antitumoral inmune dada por las células linfocíticas CD8+ asesinas (killer), es una de las respuestas más importantes ante tumores. Las células linfáticas infiltradas en el tumor deberían matar las células tumorales vía los MHC de clase I[8]. De hecho el resultado clínico y prognosis muestran una mejora en ciertos cánceres donde hay presencia de linfocitos infiltrados en el tumor, de hecho los linfocitos CD8+ son los que mejor prognosis dan al paciente cuando se encuentran infiltrados si proliferan lo suficiente[9]. De hecho se ha visto que la transferencia de tumores a ratones RAG KO (gen knock out para RAG) que el crecimiento tumoral es mayor que en ratones normales (en sarcomas y carcinomas epiteliales). Como se conoce, las enzimas Rag son responsables de la recombinación y reordenación de los fragmentos VDJ del receptor de la célula T. Así, este experimento muestra que los linfocitos T pueden controlar el crecimiento tumoral[5]. En efecto, células T CD8+ activadas son las células (leucocitos) más específicas contra las células tumorales reconociendo antígenos específicos de tumor. Esto fue probado por inmunoterapia adoptiva de células T CD8+ [8][10].

1.1.05 Células T CD8+ citotóxicas y su función efectora en tumores

Linfocitos T citotóxicos CD8+ se caracterizan por una función efectora encargada de secretar perforina, granulisinas y granzima B, IFN-γ, y expresar Fas/FasL. Se vio si se compara un ratón al que le falta la subunidad IFN-γ, a un ratón al que no le falta la subunidad, que el IFN-γ disminuye el tamaño tumoral en fibrosarcoma. Es más el IFN-γ promueve la inmunogenicidad por reconocimiento y destrucción de células tumorales, mostrando la importancia de IFN-γ en la inhibición del crecimiento tumoral[5]. IFN-γ tiene diferentes efectos como citotoxicidad, anti-angiogénesis e inducir la apoptosis[5]. Granzima B, granulisina y perforina son contenidas en gránulos. Cuando la exocitosis ocurre, la perforina crea poros permitiendo la entrada de granzima B que produce la apoptosis en las células diana. La granulisina cuando es liberada durante la exocitosis también crea poros y produce apoptosis en las células diana. Descubrimientos recientes muestran que estas moléculas

son factores clave que juegan un rol principal en la lucha contra el cáncer [11]–[13]Células CD8+ expresan FasL. Cuando se activan, reconocen el receptor Fas de las células diana y se reclutan varios factores celulares como procaspasas para empezar la vía apoptótica de las caspasas. No en todos los cánceres se ha visto que tienen sensibilidad al FasL, en algunos experimentos in vitro e in vivo con RencaHA (línea celular tumoral) se ha descubierto que FasL ayudó a rechazar tumores[14]. De todas maneras, la función efectora de las células CD8+ se ha visto que se regula a la baja en el ambiente tumoral[15].

1.1.06 Inmunoedición del cáncer

Se compone de 3 etapas: eliminación, equilibrio y escape. El paso de la eliminación se compone de 4 fases: la angiogénesis la primera, seguido de la remodelación del estroma debido al crecimiento tumoral que contribuye a señales pro-inflamatorias como adenosina, la tercera fase siendo por la presencia de células de la immunidad innata como las células δγT, y la cuarta por las NKT y NK siguiendo estas señales proinflamatorias. La etapa de equilibrio es la fase más larga donde la presión adaptativa hace que las células tumorales mueran o muten en células menos inmunogénicas produciendo tumores inestables. Las células tumorales que han mutado su MHCI o su antígeno no serán capaces de presentar sus antígenos a las células T CD8+ [5].
Los linfocitos están en el origen de esta presión selectiva Darwiniana en las variantes de antígeno de células tumorales. Esto causa consecuentemente: supervivencia de células tumorales que no son capaces de presentar por MHCI o son insensibles al IFN-γ. Así también causa la muerte de células tumorales que pueden expresar esta complejo mayor de histocompatibilidad I[16].
Se ha visto que Linfocitos T citotóxicos en la etapa de escape inmune en immunoedición del cáncer en tumores, la cadena CD3-ζ del complejo receptor T (TCR) no transduce la señal. Las células son incapaces de proliferar, influenciadas por las citokinas immunosupresivas IL-10 y TGFβ[5]. Los linfocitos infiltrados en el tumor se vuelven apoptóticos debido a la ausencia de los factores antiapoptóticos Bcl-xl y Bcl-2[5].

1.1.07 Pérdida de función en Linfocitos T infiltrados en el tumor

Los linfocitos CD8+ son activados por tumores mediante presentación cruzada de antígeno los nódulos linfáticos drenantes del tumor. De todas maneras, una vez alcanzan el microambiente tumoral, los linfocitos pierden su función efectora. Pacientes ante estas circunstancias tienen una correlación con una pobre prognosis [5][8][17]. Se ha demostrado que en ciertos cánceres las células T tienen funciones dispares como la secreción de IFN-γ pero no de perforina en melanoma. Experimentos en terapia adoptiva de células T han mostrado que la expresión del receptor de IFN-γ de las células tumorales estromales y la secreción de IFN-γ de las células T transferidas son críticas para incrementar la presentación de los antígenos de tumor vía MHCI[18]. IFN-γ puede regular a la alza la presentación de MHCI por células tumorales si no incluyen mutación en la maquinaria productora de antígeno y proteínas presentadoras como TAP1, LMP, MHC, HLA, β2-microglobuline genes[19][20].

1.1.08 Inmunosupresión dentro del microambiente tumoral, un ejemplo de expresión

El ambiente inmunosupresivo dentro de los tumores sólidos ocurre como resultado de muchos mecanismos tumorales y factores producidos en el tumor llamados: factores solubles derivados de tumor como IL-10, TGF-β, vascular endothelial growth factor (VEGF), producidas por immature Myeloid cells (iMCs) y soluble phosphatidylserine (sPS) producida por las células tumorales[18]. IL-10 también conocida como la citokina factor inhibidor de síntesis, regula a la baja la síntesis de citokinas proinflamatorias como IL-2 disminuyendo la proliferación de células. TGF-β se ha visto regulando al alza las ectoenzimas responsables de convertir ATP en adenosina en TRegs: CD39 CD79 además de regular la diferenciación a Foxp3+. VEGF es responsable de la angiogénesis y vasculogénesis resultando en aportación de nutrientes al tumor. iMCs inducen la diferenciación de células T naive a células T reguladoras (Tregs). sPS induce la secreción de IL-10, TGF-β, y PGE2

(Prostaglandina E2, o dinoprostona) por parte de macrófagos activados del tumor resultando en la inhibición de células T.
La pérdida de expresión de la molécula MHCI de las células tumorales es uno de los mecanismos que hacen que el sistema de evasión inmune impida que las células T las reconozcan[19]. Esta pérdida de expresión de MHCI es debida a: una pérdida de heterocigosis en el locus de componentes para la clase I de MHC donde el gen se encuentra; una modificación en estos genes por mutación; una menor expresión de HLA, B y loci C; y una desregulación de la maquinaria procesadora de proteínas y antígenos que incluyen los genes LMP y TAP [19][20]. Además, células presentadoras de antígeno (APC) y células malignas *producen B7-H1 y B7-H1/PD-1 (B7 inhibitory family members)* que inmunosuprimen las células T efectoras volviéndose células T exhaustas perdiendo su función efectora y decreciendo la expresión de citokina o volviéndose anérgica , lo que impide que se active la célula T [21][22].

1.1.09 Tregs y adenosina: un ejemplo de epigenética en el microambiente tumoral

Células T reguladoras, también conocidas como células T supresoras, se sabe que inhiben el sistema inmune para esquivar la respuesta a auto-antígenos. Tregs son CD4+, CD25+ y Foxp3+. En tumores Tregs y adenosina desempeñan un rol importante en la pérdida de función de los linfocitos infiltrados en el tumor. Se ha visto que Treg pasa de Tconv a Tr1 en el ambiente tumoral[23]. La célula pro- oncogénica Tr1produce citokinas inmunosupresoras (TGFβ y IL-10) y factores inmunoinhibidores (adenosina y PGE2)[24]. En tumores el ATP se desfosforila en adenosina por dos ecto-nucleotidasas: CD39 y CD73. Nucleosido trifosfato difosfohydrolasa (NTPDasa) CD39 desfosforila ATP a AMP y la ecto-5'-nucleotidasa CD73 convierte AMP en adenosina [25][26].
Además se ha visto que TGFβ no solo inhibe las funciones activadoras y efectoras de las células T pero también puede convertir células T CD4+ CD25- en células T reguladoras, disminuyendo la proliferación de células T[27].
Dentro del microambiente tumoral suele haber hipoxia resultando en una menor producción de ATP y una acumulación intracelular de AMP. Este AMP se rompe por nucleotidasas citosólicas, así la adenosina puede

viajar del espacio intracelular a extracelular. También las ecto enzimas CD73 y CD39 son sintetizadas *de novo* cuando el oxígeno intracelular disminuye. Factores inducibles de hipoxia son estabilizados y unidos a la respuesta de hipoxia induciendo la producción de CD39 y CD73. Se ha visto recientemente que CD73 es también regulado positivamente por el TGFβ[26][28]. De esta manera, de acuerdo con un estudio de meta análisis la presencia de células T reguladoras está correlacionada con una pobre prognosis [9]. Es más, la hipoxia en el tumor vía CCL28 ha demostrado recientemente promover células T reguladoras CD4+ CD25+ Foxp3+ CCR10 + [29].

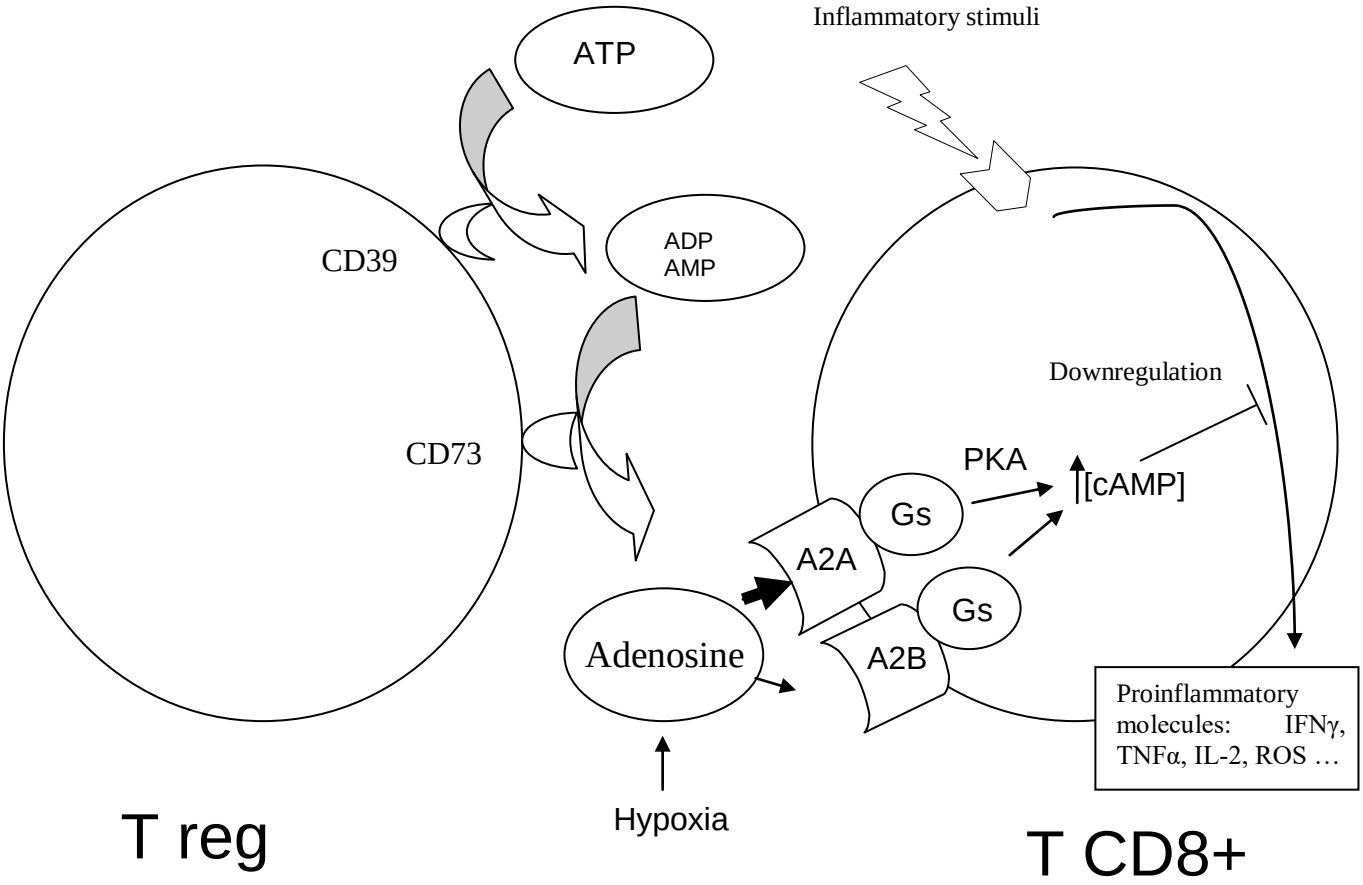


Figura 1: Regulación negativa de la respuesta de las células CD8+ por la unión de adenosina a los receptores A2A y A2B.

Las células Treg son responsables de la conversión de ATP a ADP vía las ecto-enzimas CD39 y CD73. La adenosina (metabolito) se une a los receptores A2A con más afinidad que con los A2B. Ambos adenosina Gs- receptores acoplados son capaces de incrementar AMPc en la

célula que regula negativamente la expresión genética que deriva en la función efectora y la proliferación en la célula CD8+.

1.1.10 Receptores de Adenosina A2A y A2B

Los receptores de adenosina son purinérgicos siendo parte de la clase P1 que contiene a la clase P2 unión al ATP. Existen 4 receptores de adenosina A1, A2a, A2b, y A3 y en los linfocitos están todos presentes.[30] Es sabido que los receptores de adenosina de las células T CD8+ están principalmente compuestos de A2A que controla la respuesta una vez activada [31]. A1 y A3 son Gi acoplados decrecientes de la producción de AMPc [32]. A2A y A2B receptores con proteínas Gs acopladas cambian la conformación cuando la adenosina se une, provocando la liberación de la subunidad Gα. La subunidad Gα estimulatoria activa la adenil ciclasa, así el ATP es hidrolizado a AMPc como segundo mensajero gracias a la enzima Adenilil ciclasa. AMPc es un fuerte inmunosupresor de la activación de las células T, la proliferación y la función efectora.[33] AMPc activa la proteína Kinasa A I (PKA I) localizada en los “lipid rafts” (zonas de la bicapa lipídica de la membrana celular que son ricas en ácidos grasos saturados), al lado del complejo TCR que incrementa la actividad de Csk por fosforilación. Csk también localizada en los lipid rafts, a su vez, inhibe la actividad Lck y Fyn por fosforilación. Consecuentemente al final de la vía Factor Nuclear de células T activadas (factor de transcripción NF-AT) y la respuesta de elemento uniendo a la proteína (CREB) se fosforilan. Una vez la serina se fosforila, NF-AT no puede migrar al núcleo, previniendo su unión en su sitio corriente arriba del gen IL-2, impidiendo la transcripción de IL-2. Además se ha visto que concentraciones altas de IL-2 ayudan a diferenciar células T CD8+ a linfocitos T citotóxicos con funciones efectoras.[34]

Se ha demostrado que con la administración de fármacos inmunosupresivos NF-AT no pueden ser desfosforilados por la calcineurina, impidiendo la migración al núcleo. La kinasa Src C-terminal está también activada por la proteína Kinasa A (PKA) inhibiendo la familia de Src tyr kinasas como Lck y Fyn dejando inefectiva la señalización del TCR, la primera señal de activación de las células linfocíticas T CD8+ .[33][35]

1.1.11 Rol de los receptores A2A y A2B en células T: posibles dianas para ensayos clínicos

La adenosina está presente en tejidos con inflamación aguda en una concentración elevada si se compara con tejidos no inflamados. La función efectora de las células T CD8+ puede dañar células huésped en el caso de excesiva activación, así debería ser controlado. Las células Tregs, como se ha visto anteriormente, pueden secretar factores inmunosupresivos (como TGFβ y IL-10) y factores inmunoinhibidores (como adenosina y PGE2). De todas maneras, la comprensión de los factores que involucran y producen la inhibición inmune en el microambiente tumoral es fundamental para el éxito de la inmunoterapia[36]. Además en experimentos anteriores se hizo un estudio de la proliferación en presencia de células T reguladoras CD4 CD25- y CD4 CD25+ con presencia o ausencia de adenosina y con un fármaco bloqueador de los receptores A2A y A2B: ZM-24138. Se observó una proliferación significativamente superior de células T CD8+ cuando se cultivaron con ZM-24138 que cuando se hizo sin él. Otros estudios han concluido que la inhibición de la ectoenzima CD39 disminuye la función inmunosupresora de las células tumorales[37] Por lo tanto en ambos casos los resultados parecen apuntar hacia la misma dirección.

Con estos datos podemos inferir que ésta es una de las posibles dianas farmacológicas de las muchas que podríamos encontrar con la ayuda de una mejor comprensión del transcriptoma celular de cada una de las células encontradas en el bulto tumoral o nódulo linfático y una mejor comprensión del metaboloma. Ya que se podrían entender mejor las interacciones celulares tanto dentro del nódulo linfático como en el microambiente tumoral. También se podrían descubrir nuevas dianas farmacológicas en función de qué genes se expresan en unas u otras condiciones. De ahí la importancia de los microarrays y NGS para estudiar la expresión génica o transcriptoma en diferentes situaciones o en una sola así como para estudiar posibles dianas terapéuticas. [38] Estudios de metabolómica permiten diseñar nuevos medicamentos[39]

Además RNA-seq permite clasificar diferentes subpoblaciones celulares dentro del tumor permitiendo clasificar mejor la prognosis del paciente[40].

1.1.12a Microarrays

Es una tecnología para realizar experimentos que permiten estudiar diversos genes, proteínas o metabolitos. Tiene una base sólida de cristal plástico o sílice, donde se estudian los anteriormente mencionados (genes o proteínas…) y están expuestos a moléculas cuya interacción se analiza midiendo la expresión (analógica o digital).

A pesar de que se pueden estudiar muchos genes a la vez quizás se encuentre más ruido de fondo que en los métodos hasta ahora utilizados.

Realiza comparaciones de clusters mediante modelos lineales implementados en el paquete limma de bioconductor. El descubrimiento de clases en microarrays es importante para estudiar la expresión diferencial por ejemplo o bien los SNP's en un estudio de asociación del genoma completo (GWAS). También existen otras tecnologías como los microRNAs y arrays que estudian la metilación además de la NGS.

Existen Microarrays de dos colores (spotted arrays) que se basan en la hibridación competitiva por dos fluorocromos distintos (como podrían ser Cy3 y Cy5) y microarrays de un color o arrays de oligonucleótidos.

Se basa en la hibridación de la muestra (target) y el microarray, y en la excitación de los fluorocromos con láser. Se mide la expresión relativa.

Los microarrays de un color se miden en expresión absoluta y solo utilizan un tipo de fluorocromo.

En el caso que nos ocuparía en ambos casos se utiliza ARN como muestra para medir la expresión génica.

1.1.12b Sanger, Microarrays y RNA-Seq

El análisis de los datos ómicos empieza con la secuenciación de Sanger, un método para secuenciar el DNA descubierto por Frederick Sanger y su equipo en 1977. Seguidamente se incorporaron los microarrays hacia el año 1990-1995[41] y más tarde el NGS.

Tanto los microarrays como RNA-Seq son capaces de perfilar el transcriptoma.[42] Hoy en día se utiliza NGS que tiene ventajas en cuanto a sanger como la amplificación clonal (más rápida) por la que

mediante la clonación se aumentan el numero de secuencias a detectar (ya que la técnica RNA-seq no es tan sensible). Ambos sistemas pueden realizar la comparación de grupos. Es decir pueden analizar la expresión diferencial en dos casos distintos.
RNA-Seq proporciona información más detallada en la expresión de patrones específicos que los microarrays en unas muestras de neuroblastoma[43]. Por tanto RNA-Seq supera a los microarrays determinando la transcriptómica del cáncer, sin embargo microarrays y RNA-seq determinan la prognosis del paciente con eficiencia[43]. En los microarrays tenemos como medida un ratio con lo cual es una medida relativa y en el NGS mide la expresión en counts, una medida discreta.

1.1.13 NGS Next Generation Sequencing, RNA-Seq, datos ómicos

La técnica RNA-Seq es novedosa, descrita en 2008 por primera vez.[44]
El RNA-Seq se utiliza para analizar el transcriptoma, es decir la expresión génica de una población celular en un momento dado. Se puede utilizar para comparar dos situaciones distintas como por ejemplo un cierto tratamiento a un control (como es el caso de la Differential Gene Expression) principalmente pero también para estudiar los SNPs (Single-nucleotide polymorphisms), la detección y cuantificación de RNA no génico, isoformas de splice, nuevos transcritos y sitios de interacción proteína RNA. Se ha visto que el RNA-Seq de Illumina por ejemplo es más sensible que los microarrays de Affymetrix[44]. RNA-Seq detecta niveles de expresión bajos en comparación a los microarrays, además la tecnología RNA-seq mide en counts por lo que no tiene máximo cuantificable como los microarrays[45]. Hay mayor especificidad y sensibilidad en RNA-seq que en los microarrays[45]. Debido al abaratamiento de las técnicas y maquinaria de ultrasecuenciación, con Illumina HiSeq, ABI 5500x, e ION PROTON, por ejemplo, podemos en poco tiempo y con relativamente poco dinero (unos pocos miles de dólares) analizar el transcriptoma de las muestras. El precio de cada una de los secuenciadores varía entre 89.000$ y 750.000$ (http://www.labx.com/product/illumina-sequencer). Por convención para estudiar la expresión génica diferencial en dos situaciones distintas se utiliza RNA-seq a pesar de que todavía no hay un protocolo o varios

protocolos estándar, homologados, o acordados con la comunidad científica ya que cada situación es diferente. Es tarea del bioinformático filtrar, normalizar, procesar los datos. Es cierto que existen aplicaciones y software que pueden tratar también datos de RNA-Seq como por ejemplo GALAXY, aunque con más limitaciones y menos opciones que R.

1.1.14 Illumina

Illumina es un proveedor de maquinarias de ultrasecuenciación. Sus diferentes sistemas y productos son capaces de secuenciar el ADN y ARN entre ellos se encuentran NextSeq, HiSeq, NovaSeq. Con secuenciadores de segunda o tercera generación como estos logramos una mayor información acerca de hasta todo un genoma completo y no solo unos genes como en los microarrays. Aunque en un microarray hubiese 100 genes a investigar en un RNA-seq se están investigando más de 10 veces esa cantidad de genes. Además la tecnología Illumina dispone de una mayor sensibilidad a la hora de detectar positivos que son realmente positivos sin alcanzar tantos falsos negativos o falsos positivos como en otras técnicas.

1.1.15 Preparación de las muestras para RNA-seq

Para el estudio de la expresión génica queremos obtener la información del RNA transcrito, secuenciar el RNA o el transcriptoma. Para ello hay que separarlo de su ambiente celular compuesto básicamente por proteínas y DNA. Los dos métodos más conocidos para obtener el RNA aislado son membranas de silica-gel y extracciones líquido-líquido con fenol-cloroformo acídico. En las membranas de silica-gel el RNA se une exclusivamente a la membrana de gel de silica, que requieren de etanol para la unión, mientras que el resto de componentes celulares son eliminados con el lavado. La cantidad de etanol influencia proporcional y positivamente la retención de RNA.[46]
Usando la extracción de fenol-cloroformo, los componentes celulares son disueltos en tres fases: la fase orgánica, la interfase, y la fase acuosa, en la que el RNA es retenido. A esta extracción le sigue una

precipitación con alcohol (etanol o isopropanol) para desalar y concentrar el RNA. El tipo de sal retiene solo un tipo de RNA.
Usar un RFP (RNA Fragmentation Plate) para que el mRNA se una a los "*beads*" para eliminar el rRNA residual. Seguidamente se sintetiza sscDNA (single strand copy) a partir de la transcriptasa inversa y primers.[47]

1.1.16 R

R fue el resultado de la evolución o inspiración de S un lenguaje para el análisis de datos.
Es un lenguaje de programación estadístico principalmente, usado también en investigación biomédica ya que con paquetes de Bioconductor (software libre para la Bioinformática) y sus herramientas podemos analizar y representar los resultados de datos genómicos de alto rendimiento o "high- throughput genomic data". Entre este tipo de datos genómicos de alto rendimiento encontramos: microarrays y RNA-seq (Next Generation Sequencing) o secuenciación de segunda generación.
Para el análisis de datos ómicos podemos utilizar R y sus paquetes tanto estadísticos, gráficos y de análisis de datos RNA-Seq con limma, EdgeR, glimma, Deseq2, bowtie… Existen múltiples opciones a la hora de realizar un análisis de datos de RNA-Seq y ningún protocolo estandarizado fijo que se considere el mejor.
R es un software libre tipo Unix, aunque su versión existe para diferentes tipos de sistema operativo incluyendo Windows, acepta varios editores de texto, entre ellos: RStudio, Eclipse, notepad++, así como varias interfaces gráficas como RStudio también y R Commander entre muchos otros. En el trabajo se utilizó RStudio como entorno de desarrollo integrado (IDE).

1.1.17 Shiny

Es una librería, paquete y herramienta de R para construir aplicaciones web y pretende favorecer el uso interactivo de los datos mediante

opciones que el usuario puede modificar. Por ejemplo: los no usuarios de R y los protocolos bioinformáticos para el análisis de datos ómicos pueden obtener la información visualmente e interactivamente con inputs modificables por el usuario de la aplicación. El objetivo de shiny en el ámbito médico sería representar los datos de manera personalizada para que un equipo no bioinformático, el equipo médico por ejemplo, pueda estudiar mejor los datos crudos o analizados de los experimentos, en este caso de NGS. Gracias a shiny esto sería posible, por el momento el análisis de datos ómicos más frecuente requiere de alguien para analizar los datos e ir modificando el código para representar los datos de manera diferente. Con shiny sin embargo podemos representarlos de la manera que sea más conveniente para cada caso.

1.1.18 UI Interfaz de Usuario

La interfaz de usuario o user interface (UI) es la parte de la aplicación con la que podemos interaccionar como usuarios y que visualizamos al abrir la aplicación. Ui.r. es el archivo que corresponde a esta parte de la aplicación. Cuando se inicie la aplicación será lo que veamos. El usuario de la aplicación puede seleccionar el input en el caso de la aplicación creada y tiene 4 opciones mediante selectInput(): seqdata1, seqdata2, seqdata3, y seqdata4.

El bioinformático tiene otras opciones como actionButton(), checkboxInput(), checkboxGroupInput(), dateInput(), dateRangeInput(), fileInput(), numericInput(), passwordInput(), radioButtons(), sliderInput() y textInput() para insertar inputs. Para poder crear una aplicación dinámica se ha de utilizar la función “reactive” (reactiva) para que los datos visualizados como outputs una vez procesados por el servidor puedan cambiar cada vez que el usuario seleccione un input diferente.

Si se comparase a una web creada con html php y mysql por ejemplo el UI correspondería a la misma función que realiza el código html.

1.1.19 Server

En el caso que nos ocupa utilizaremos el propio ordenador como servidor o server para ejecutar la aplicación. Se podría en efecto ejecutar

el server desde la web (webserver) para poder verlo desde la web. El archivo server.r es la parte de la aplicación que designa como se ejecuta, qué funciones se le da, qué y cómo es lo que va a representar en función de los inputs que seleccione el usuario. Llama y procesa outputs en función de los inputs de UI. La función render usada en server.R conjuntamente con la función de output de server.R permiten generar llamadas a la UI para que los resultados aparezcan donde está previsto por el UI.
En este caso si se comparase a una web creada con html php y mysql, el server.r realizaría la misma función que el código php.

1.1.20 MDS Plot

Dentro del paquete edgeR de R, obtenemos la función plotMDS() (gráfico de escala multidimensional). Medimos la similitud de las muestras en un gráfico con abscisas y ordenadas [48] Se escogió edgeR ya que al analizar la expresión diferencial se ha visto que PCA (principal component análisis) de DESeq tiende a sobreestimar la dispersión. Además para realizar un análisis de expresión diferencial sin sobreestimar falsos positivos sólo se tenían dos opciones o DESeq o edgeR.[49]
Aunque en el código de R se visualizó el plotMDS para ver el control de calidad de las muestras, la aplicación shiny no contiene dicho gráfico.

1.1.21 Importancia del control de calidad de las muestras para el consecutivo análisis

Importancia de detectar el efecto del lote o “Batch effect” y poder corregirlo a tiempo [50]. Se caracteriza por ser una fuente técnica de variación añadida en las muestras durante la manipulación si se hace de manera incorrecta. Si no es corregido puede dar lugar a confusión o mala interpretación de los resultados.

1.1.23 Firma de la expresión genética o signature genes

Es de vital importancia lograr el entendimiento de la expresión génica en melanoma. Para ello se pueden realizar diferentes técnicas como FACS y también de RNA-seq. El objetivo es agrupar en clusters los perfiles de expresión génica para poder determinar a qué células corresponden. Así mismo se pueden descubrir dianas farmacológicas ya que se podrá ver qué genes juegan un rol y qué proteínas se expresan. Se pueden conseguir nuevos tratamientos: entre ellos inmunoterapias o dianas farmacológicas para nuevos medicamentos por ejemplo[51].

1.1.24 Immunogenómica

Es importante entender el conjunto de los genes que juegan un papel esencial en el escape y evasión inmune en casos como el de melanoma u otros tipos de cáncer. La caracterización de las células infiltradas inmunes mostró que los genotipos tumorales determinaron los mecanismos de escape del tumor y los inmunofenotipos.[52] Por tanto es también de vital importancia conocer los linfocitos infiltrados en el tumor para poder predecir las posibles respuestas o no respuesta por inhibiciones ejercidas sobre del sistema inmune.[53] No es fácil tarea determinar los componentes del microambiente tumoral que interaccionan con los linfocitos provenientes de los nódulos linfáticos. Es más, la posible expresión de los linfocitos se puede modificar como se ha visto en el ejemplo de los linfocitos citotóxicos CD8+ con el efecto de la adenosina y los linfocitos Tregs. Y el microambiente tumoral además puede cambiar varias veces, o ser variado cambiando de un tumor a otro en pacientes que presentan mestástasis [54]. De ahí la importancia de determinar los diferentes tipos de expresión génica en cada uno de los tumores con la consiguiente clasificación de los linfocitos infiltrados.

1.2 Objetivos del Trabajo

El objetivo principal es realizar un control de calidad de los datos de expresión génica en RNA-Seq de nódulos linfáticos en humanos con melanoma para poderlo visualizar mediante una aplicación interactiva con gráficos y tablas. Para ello necesitamos de:

- Mejor entendimiento de la evolución del análisis de datos ómicos y del análisis de datos de NGS como RNA-Seq.
- Aprender a diseñar aplicaciones interactivas con la librería shiny de R para mostrar datos.
- Realizar varios análisis de datos de RNA-Seq
- Realizar aplicaciones prueba con shiny
- Entender mejor la evasión inmune e inmunosupresión en cáncer

Además se pretende presentar los gastos para un posible proyecto de investigación.

1.3 Enfoque y método seguido

La estrategia que se siguió constó de un aprendizaje constante en materia de cáncer, inmunología, RNA-Seq, y shiny. Para RNA-Seq y shiny se realizaron diversos tutoriales

Se quiere aprender a tener la capacidad de poder desarrollar aplicaciones a medida gracias a haber aprendido a realizar aplicaciones con shiny en R.

Primero se aprendió a realizar algunas aplicaciones sencillas con shiny para después poder utilizar los datos obtenidos en el análisis de expresión génica en una nueva aplicación que los integrase.

Las diferentes partes del código hacen referencia a paquetes de R ya existentes por lo tanto no es un producto nuevo, sin embargo la creación de la aplicación con el control de calidad de los datos sí que sería un producto nuevo.

Primero se realizó el principio del control de calidad de los datos y se intentó llegar hasta la creación de los heatmaps con los datos logaritmo en base 2 transformados pero aún sin normalizar. Después se empezó a desarrollar el código en shiny (server y ui).

A la hora de encarar el trabajo del análisis de expresión diferencial se usó el paquete edgeR en vez del paquete DESeq y no DESeq2 para no sobreestimar la dispersión. Se hubiese podido utilizar el paquete limma

1.4 Planificación del Trabajo

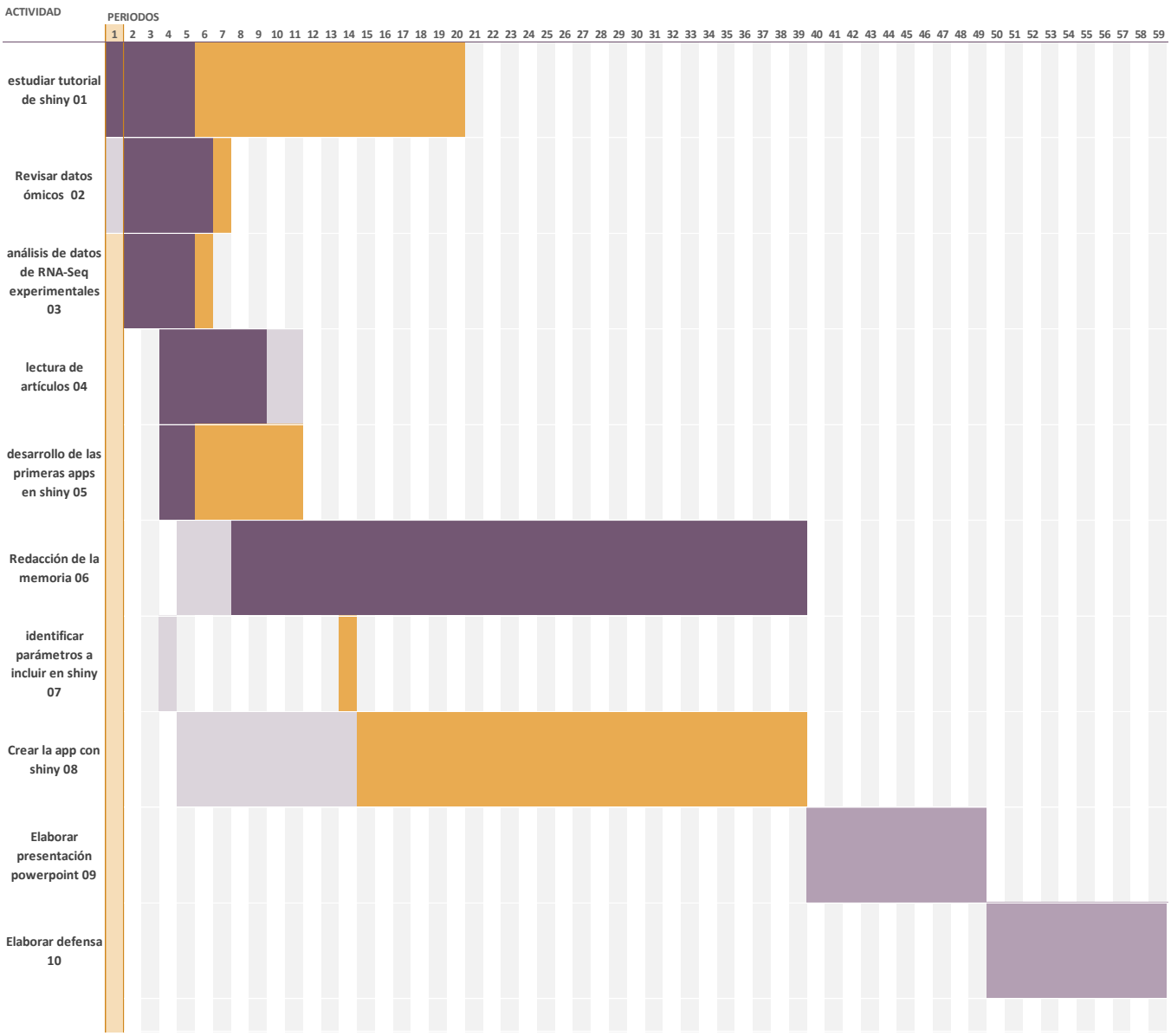


Figura 2: Diagrama de Gantt

ACTIVIDAD	INICIO DEL PLAN	DURACIÓN DEL PLAN	INICIO REAL	DURACIÓN REAL	PORCENTAJE COMPLETADO
estudiar tutorial de shiny 01	1	5	1	20	**100%**
Revisar datos ómicos 02	1	6	2	6	**100%**
análisis de datos de RNA-Seq experimentales 03	2	4	2	5	**100%**
lectura de artículos 04	4	8	4	6	**100%**
desarrollo de las primeras apps en shiny 05	4	2	4	8	**100%**
Redacción de la memoria 06	5	35	8	32	**100%**
identificar parámetros a incluir en shiny 07	4	1	14	1	**100%**
Crear la app con shiny 08	5	10	15	25	**100%**
Elaborar presentación powerpoint 09	40	10	40	10	**0%**
Elaborar defensa 10	50	10	50	10	**0%**

Tabla 1: Tabla asociada al diagrama de Gantt

1.5 Breve sumario de productos obtenidos

Memoria explicativa de por qué las ciencias ómicas son importantes y un ejemplo concreto en cómo se integran gracias al estudio de ejemplo de las CD8 y su relación con el ambiente inmunosupresivo tumoral. Además se obtuvo una aplicación shiny contenida en los archivos ui.r, server.r, y la carpeta data con los counts en formato texto descargados de https://www.ncbi.nlm.nih.gov/geo/ .

1.6 Breve descripción de los otros capítulos de la memoria

El resto de capítulos se subdividen como si de un artículo científico se tratase:

- Resultados: Exposición y explicación del código obtenido en R, dividiéndolo en los ficheros UI.r y server.r.
- Discusión: se discuten aspectos que se podrían haber estudiado, que faltan por estudiar, y los aspectos estudiados de los resultados.
- Materiales y métodos: donde se explica qué y cómo se utilizó para realizar los experimentos.

2. Resto de capítulos

2.1 Resultados

2.1.1 UI User Interface

Para que funcione el siguiente código hay que llamar runApp() con el directorio de trabajo de R (versión utilizada 3.4.3) en el que se encuentran las muestras (carpeta data), ui.r y server.r. Se utiliza setwd() para localizar el directorio de trabajo en esta carpeta de interés. Luego realizamos runApp() en la consola para ejecutar la aplicación.

```
#####################################################################
library(shiny)
library(DT)

# Define UI for application that draws a histogram
ui <- fluidPage(
# Insertamos un título a la página
  titlePanel("CPMs para normalizar"),

  mainPanel(
#llamamos al los input vsqd,( los counts en formato txt que lee el server.)
#Tenemos 4 muestras seqdata1, seqdata2, seqdata3, y seqdata4 con los
#identificadores respectivos 1, 2, 3, y 4 (usadas en el archivo server.r)
    selectInput("vsqd", "Sequencias a Escoger:",
                c("seqdata1" = 1,
                  "seqdata2" = 2,
                  "seqdata3" = 3,
                  "seqdata4" = 4)),

    # Se ordenan los outputs en diferentes tabs o lengüetas (3, 4 y 1) en 3
#paneles diferentes. Se llaman a las tablas con DT::dataTableOutput() del
#paquete DT, los gráficos con la función plotOutput () y la consola  que se ve
#con  verbatimTextOutput (). El orden de los paneles es consecutivo al orden
#en que se realizó el protocolo para visualizar los datos, control y distribución
#de las muestras, conteniendo los últimos paneles más información, o
#información más útil que los primeros
```

```
    tabsetPanel(type = "tabs",
                tabPanel("CPMs para nomalizar", DT::dataTableOutput("table")),
                tabPanel("Genoma Matriz logica", DT::dataTableOutput("tablelogic")),
                tabPanel("Media de Falsos", verbatimTextOutput("summary"))

    ),
    tabsetPanel(type = "tabs",
                tabPanel("Vista Y", verbatimTextOutput("vistaY")),
                tabPanel("Slots Guardados en Y", verbatimTextOutput("slotsY")),
                tabPanel("Distribución Muestras", plotOutput("plotxy")),
                tabPanel("Cabecera Genes", verbatimTextOutput("cabeceraGenes"))
    ),
    tabsetPanel(type = "tabs",
                tabPanel("Mas altos LCPMs", DT::dataTableOutput("tablelcpm"))
    )
  )
)
```

2.1.2 Server

```
# Instala el paquete edgeR desde el repositorio de www.bioconductor.org
source("http://bioconductor.org/biocLite.R")

# Descarga el paquete edgeR
biocLite("edgeR")

# Llama al paquete edgeR y shiny (ya instalado, sino habría que instalarlo)
library(edgeR)
library(shiny)

# Se leen los datos en R desde la carpeta data con los 4 archivos de texto que
#denominamos seqdata1, 2, 3,y 4.

seqdata1 <- read.delim("data/GSM2461003_LAU125.genes.results.txt",
stringsAsFactors = FALSE)
seqdata2 <- read.delim("data/GSM2461005_LAU355.genes.results.txt",
stringsAsFactors = FALSE)
seqdata3 <- read.delim("data/GSM2461007_LAU1255.genes.results.txt",
stringsAsFactors = FALSE)
seqdata4 <- read.delim("data/GSM2461009_LAU1314.genes.results.txt",
stringsAsFactors = FALSE)

# Se define el server para cargar los datos de entrada. Consideramos server
#una función con inputs de entrada y outputs de salida con varios objetos
#reactivos que permiten el cambio en UI de los resultados en función de la
#selección del input por parte del usuario de la aplicación.

server <- function(input, output) {

  selectorseqdata <- reactive({input$vsqd})

  # SEQDATA1
  # El comando head permite visualizar las seis primeras líneas
  head(seqdata1)
  # El comando dim devuelve las dimensiones de la matriz original
  dim(seqdata1)
```

```
  # Redimensiona la matriz en un arreglo con datos numéricos, quitamos las
#columnas 1 y 2 que contienen información acerca de la anotación ya que son
#prescindibles.
  countdata <- seqdata1[,-(1:2)]

  # SEQDATA2
  # Para ver las seis primeras líneas
  head(seqdata2)
  # Dimensión de la matriz original
  dim(seqdata2)
  # Redimensiona la matriz en un arreglo con datos numéricos sin las dos
#primeras 2 columnas.
  countdata2 <- seqdata2[,-(1:2)]

  # SEQDATA3
  # Para ver las seis primeras líneas
  head(seqdata3)
  # Dimensión de la matriz original
  dim(seqdata3)
  # Redimensiona la matriz en un arreglo con datos numéricos
  countdata3 <- seqdata3[,-(1:2)]

  # SEQDATA4
  # Para ver las seis primeras líneas
  head(seqdata4)
  # Dimensión de la matriz original
  dim(seqdata4)
  # Redimensiona la matriz en un arreglo con datos numéricos
  countdata4 <- seqdata4[,-(1:2)]


  # CPMS de SEQDATA1
  # Obtiene Counts Per Million o reads per kilobase per million son las lecturas
#por kilobase por millón. Estas CPMs son para normalizar en cuentas por
#millón para luego filtrar los genes (filas) no expresados.
  # Se normaliza con la función cpm()
  myCPM <- cpm(countdata)

#se realiza el mismo paso para las 3 muestras restantes.
```

```
# CMPS de SEQDATA2
# Obtiene CPMs para normalizar en cuentas por millón para luego filtrar.
# Se normaliza por defecto
myCPM2 <- cpm(countdata2)

# CMPS de SEQDATA3
# Obtiene CPMs para normalizar en cuentas por millón para luego filtrar.
# Se normaliza por defecto
myCPM3 <- cpm(countdata3)

# CMPS de SEQDATA4
# Obtiene CPMs para normalizar en cuentas por millón para luego filtrar.
# Se normaliza por defecto
myCPM4 <- cpm(countdata4)
```

#Ahora definiremos un threshold o valor mínimo umbral para el cual #consideraremos que los counts para un gen son TRUEs (verdaderos) si están #por encima del valor umbral. En este caso exigimos un mínimo de 0.5, siendo #entre 0.5 y 1 utilizados más corrientemente. En este caso quizás un valor de 1 #o 2 hubiese sido más adecuado.

```
######SEQDATA1################
thresh <- myCPM > 0.5
# Esto permite generar una matriz lógica con TRUEs y FALSEs.  tablalogic <-
head(thresh).
```

#Se repite el procedimiento para el RNA secuenciado en las muestras 2 3 y 4.

```
######SEQDATA2################
thresh2 <- myCPM2 > 0.5
# Esto permite generar una matriz lógica con TRUEs y FALSEs
tablalogic2 <- head(thresh2)
######SEQDATA3################
thresh3 <- myCPM3 > 0.5
# Esto permite generar una matriz lógica con TRUEs y FALSEs
tablalogic3 <- head(thresh3)
######SEQDATA4################
thresh4 <- myCPM4 > 0.5
# Esto permite generar una matriz lógica con TRUEs y FALSEs
tablalogic4 <- head(thresh4)
```

```
  ######SEQDATA1################
  head(myCPM)
  table(rowSums(thresh))
  # Se quieren mantener en la tabla los genes que tienen al menos dos TRUES
#en cada fila (para cada gen)
  keep <- rowSums(thresh) >= 2
  # Subconjunto de filas de countdata para mantener los genes más expresivos
#con el filtraje de mínimo dos trues por fila
  counts.keep <- countdata[keep,]
  summary(keep)
  dim(counts.keep)
#repetimos para seqdata  2, 3, y 4
  ######SEQDATA2################
  head(myCPM2)
  table(rowSums(thresh2))
  # Mantener genes que tienen al menos dos TRUES en cada fila
  keep2 <- rowSums(thresh2) >= 2
  # Subconjunto de filas de countdata para mantener los genes más expresivos
  counts.keep2 <- countdata[keep2,]
  summary(keep2)
  dim(counts.keep2)
  ######SEQDATA3################
  head(myCPM3)
  table(rowSums(thresh3))
  # Mantener genes que tienen al menos dos TRUES en cada fila
  keep3 <- rowSums(thresh3) >= 2
  # Subconjunto de filas de countdata para mantener los genes más expresivos
  counts.keep3 <- countdata[keep3,]
  summary(keep3)
  dim(counts.keep3)
  ######SEQDATA4################
  head(myCPM4)
  table(rowSums(thresh4))
  # Mantener genes que tienen al menos dos TRUES en cada fila
  keep4 <- rowSums(thresh4) >= 2
  # Subconjunto de filas de countdata para mantener los genes más expresivos
  counts.keep4 <- countdata[keep4,]
  summary(keep4)
  dim(counts.keep4)
```

```
  ######SEQDATA1################
  # Agrega una línea vertical opcional en 0.5 CPM
  # abline(v=0.5)
  #Objeto DGEList, y, utilizado para almacenar la información de la expresión
#génica en counts de los de los genes que expresan más de 2 counts por gen
#o fila o tránscrito
  y <- DGEList(counts.keep)
  ytabla <- DGEList(counts.keep)
  # La información del tamaño de la librería se guarda en slot de muestras y,
#conteniendo counts y samples.
  y$samples
  y$samples$lib.size
#Realizamos la misma operación para las 3 muestras restantes.
 ######SEQDATA2################
  # Agrega una línea vertical en 0.5 CPM
  # abline(v=0.5)
  y2 <- DGEList(counts.keep2)
  ytabla2 <- DGEList(counts.keep2)
  # Información del tamaño de Library con información acerca de cuantas
#lecturas o reads hay por secuencia, se guarda en slot de muestras
  y2$samples
  y2$samples$lib.size
  ######SEQDATA3################
  # Agrega una línea vertical en 0.5 CPM
  # abline(v=0.5)
  y3 <- DGEList(counts.keep3)
  ytabla3 <- DGEList(counts.keep3)
  # Información del tamaño de Library se guarda en slot de muestras
  y3$samples
  y3$samples$lib.size
  ######SEQDATA4################
  # Agrega una línea vertical en 0.5 CPM
  # abline(v=0.5)
  y4 <- DGEList(counts.keep4)
  ytabla4 <- DGEList(counts.keep4)
  # Información del tamaño de Library se guarda en slot de muestras
  y4$samples
  y4$samples$lib.size
```

```
  ######SEQDATA1################
  # Para determinar el control de calidad de las muestras es preciso realizar un
#gráfico de barras con cuantas lecturas o reads tiene cada muestra.
#Se marcan las etiquetas o labels en vertical. Se estudia la distribución de las
#muestras sin normalizar
  barplot(y$samples$lib.size,names=colnames(y),las=2)
  # Se agrega título al gráfico
  title("Barplot of library sizes")
  # Obtiene log2 de cuentas por millón con log=TRUE, aproximamos los datos a
#una distribución log2-normal para hacerlos comparables.
  logcounts <- cpm(y,log=TRUE)
  # Verifica distribuciones de muestras usando diagramas de caja
  boxplot(logcounts, xlab="", ylab="Log2 counts per million",las=2)
  # Agreguemos una línea azul horizontal que corresponde al logCPM mediano
  abline(h=median(logcounts),col="blue")
  title("Boxplots of logCPMs (unnormalised)")
  # Gráficos multidimensionales
  plotMDS(y)
  ######SEQDATA2################
  # El último argumento rota el nombre del eje X
  barplot(y2$samples$lib.size,names=colnames(y2),las=2)
  # Agrega título al gráfico
  title("Barplot of library sizes 2")
  # Obtiene log2 de cuentas por millón
  logcounts2 <- cpm(y2,log=TRUE)
  # Verifica distribuciones de muestras usando diagramas de caja
  boxplot(logcounts2, xlab="", ylab="Log2 counts per million",las=2)
  # Agreguemos una línea azul horizontal que corresponde al logCPM mediano
  abline(h=median(logcounts2),col="blue")
  title("Boxplots of logCPMs (unnormalised)")
  # Gráficos multidimensionales
  plotMDS(y2)
  ######SEQDATA3################
  # El Último argumento rota el nombre del eje X
  barplot(y3$samples$lib.size,names=colnames(y3),las=2)
  # Agrega título al gráfico
  title("Barplot of library sizes 2")
  # Obtiene log2 de cuentas por millón
```

```
logcounts3 <- cpm(y3,log=TRUE)
# Verifica distribuciones de muestras usando diagramas de caja
boxplot(logcounts3, xlab="", ylab="Log2 counts per million",las=2)
# Agreguemos una línea azul horizontal que corresponde al logCPM mediano
abline(h=median(logcounts3),col="blue")
title("Boxplots of logCPMs (unnormalised)")
# Gráficos multidimensionales
plotMDS(y3)
######SEQDATA4################
# El último argumento rota el nombre del eje X
barplot(y4$samples$lib.size,names=colnames(y4),las=2)
# Agrega título al gráfico
title("Barplot of library sizes 2")
# Obtiene log2 de cuentas por millón
logcounts4 <- cpm(y4,log=TRUE)
# Verifica distribuciones de muestras usando diagramas de caja
boxplot(logcounts4, xlab="", ylab="Log2 counts per million",las=2)
# Agreguemos una línea azul horizontal que corresponde al logCPM mediano
abline(h=median(logcounts4),col="blue")
title("Boxplots of logCPMs (unnormalised)")
# Gráficos multidimensionales
plotMDS(y4)
´


#Quedó pendiente integrar el plotMDS y heatmaps e integrarlo a la aplicación
#de shiny.
######SEQDATA1################
# Clustering jerárquico
# Se estima la varianza para cada fila en la matriz de logcounts
var_genes <- apply(logcounts, 1, var)
infogenes <- head(var_genes)
######SEQDATA2################
# Clustering jerárquico
# Se estima la varianza para cada fila en la matriz de logcounts
var_genes2 <- apply(logcounts2, 1, var)
infogenes2 <- head(var_genes2)
######SEQDATA3################
# Clustering jerárquico
# Se estima la varianza para cada fila en la matriz de logcounts
```

```
var_genes3 <- apply(logcounts3, 1, var)
infogenes3 <- head(var_genes3)
######SEQDATA4###############
# Clustering jerárquico
# Se estima la varianza para cada fila en la matriz de logcounts
var_genes4 <- apply(logcounts4, 1, var)
infogenes4 <- head(var_genes4)

######SEQDATA1###############
# Se obtienen los nombres de los genes para los 500 genes con la varianza
#mayor
select_var <- names(sort(var_genes, decreasing=TRUE))[1:500]
head(select_var)
######SEQDATA2###############
# Se obtienen los nombres de los genes para los 500 genes con la varianza
#mayor
select_var2 <- names(sort(var_genes2, decreasing=TRUE))[1:500]
head(select_var2)
######SEQDATA3###############
# Se obtienen los nombres de los genes para los 500 genes con la varianza
#mayor
select_var3 <- names(sort(var_genes3, decreasing=TRUE))[1:500]
head(select_var3)
######SEQDATA4###############
# Se obtienen los nombres de los genes para los 500 genes con la varianza
#mayor
select_var4 <- names(sort(var_genes4, decreasing=TRUE))[1:500]
head(select_var4)

######SEQDATA1###############
# Matriz de logcounts de subconjuntos seleccionados en el paso anterior
highly_variable_lcpm <- logcounts[select_var,]
dim(highly_variable_lcpm)
tablalcpm <- head(highly_variable_lcpm)
######SEQDATA2###############
# Matriz de logcounts de subconjuntos
highly_variable_lcpm2 <- logcounts[select_var2,]
dim(highly_variable_lcpm2)
tablalcpm2 <- head(highly_variable_lcpm2)
```

```
  ######SEQDATA3################
  # Matriz de logcounts de subconjuntos
  highly_variable_lcpm3 <- logcounts[select_var3,]
  dim(highly_variable_lcpm3)
  tablalcpm3 <- head(highly_variable_lcpm3)
  ######SEQDATA4################
  # Matriz de logcounts de subconjuntos
  highly_variable_lcpm4 <- logcounts[select_var4,]
  dim(highly_variable_lcpm4)
  tablalcpm4 <- head(highly_variable_lcpm4)

  # Expresión para generar la distribución deseada.
  # Se llama a esta función para cambiar los inputs. Las funciones output  se
#definen más abajo.
  d <- reactive({
    dist <- switch(input$dist,
                   norm = rnorm,
                   unif = runif,
                   lnorm = rlnorm,
                   exp = rexp,
                   rnorm)

    dist(input$n)
  })

  # Invocamos la función renderPrint para que imprima en este caso output
#resumen de los datos en función de qué muestra (seqdata) esté seleccionada
#para utilizarse. Genera los falsos y negativos en cuanto expresión génica se
#refiere según los criterios anteriores con un valor umbral, y los valores pre-
#filtrados.
  output$summary <- renderPrint({
    if(selectorseqdata() == 1){
      summary(keep)
    } else if(selectorseqdata() == 2){
      summary(keep2)
    } else if(selectorseqdata() == 3){
      summary(keep3)
    } else {
      summary(keep4)
```

```
  }
 })

  # Se genera una tabla HTML con las filas y columnas de los archivos de texto
#leídos con read.delim, sin contar con las columnas de las anotaciones que se
#han eliminado. El usuario selecciona el input en el UI que al ser reactivo hace
#que la se llame la función renderDataTable en este caso el server para que se
#recalculen los resultados y se vuelvan a enviar al UI para que puedan ser
#visualizados por el usuario. En este caso utilizamos operadores condicionales
#para seleccionar el distinto input en este caso llamado por la función
selectorseqdata()
 output$table <- DT::renderDataTable({
   if(selectorseqdata() == 1){
     countdata
   } else if(selectorseqdata() == 2){
     countdata2
   } else if(selectorseqdata() == 3){
     countdata3
   } else {
     countdata4
   }
 })

  # Se escoge la tabla con la matriz de valores lógicos (True o False) para la
#expresión génica en función de si los counts pasan el valor umbral mínimo
(threshold) impuesto previamente.
 output$tablelogic <- DT::renderDataTable({
   if(selectorseqdata() == 1){
     tablalogic
   } else if(selectorseqdata() == 2){
     tablalogic2
   } else if(selectorseqdata() == 3){
     tablalogic3
   } else {
     tablalogic4
   }
 })
```

```
  # Este objeto es la última tabla que corresponde a la tab de Mas altos LCPMs
#(log2 counts per million) representada en la UI de la aplicación con las
#varianzas más variables de los 500 primeros genes escogidos en orden
#decreciente
  output$tablelcpm <- DT::renderDataTable({
    if(selectorseqdata() == 1){
      tablalcpm
    } else if(selectorseqdata() == 2){
      tablalcpm2
    } else if(selectorseqdata() == 3){
      tablalcpm3
    } else {
      tablalcpm4
    }
  })

  # Contiene la descripción de counts y samples del objeto DGEList.
  output$vistaY <- renderPrint({
    if(selectorseqdata() == 1){
      ytabla
    } else if(selectorseqdata() == 2){
      ytabla2
    } else if(selectorseqdata() == 3){
      ytabla3
    } else {
      ytabla4
    }
  })

  # Se generan los logaritmos de las columnas length, effective_length,
expected_count, TPM y FPKM
  output$cabeceraGenes <- renderPrint({
    if(selectorseqdata() == 1){
      infogenes
    } else if(selectorseqdata() == 2){
      infogenes2
    } else if(selectorseqdata() == 3){
      infogenes3
    } else {
```

```
      infogenes4
    }
  })


  #muestra los Slots guardados en Y: counts y samples ----
  output$slotsY <- renderPrint({
    if(selectorseqdata() == 1){
      names(ytabla)
    } else if(selectorseqdata() == 2){
      names(ytabla2)
    } else if(selectorseqdata() == 3){
      names(ytabla3)
    } else {
      names(ytabla4)
    }
  })

# barplot o gráfico de barras mostrando la distribución para cada una de las
#muestras en función de la seleccionada.
  output$plotxy <- renderPlot({
    if(selectorseqdata() == 1){
      plot <- boxplot(logcounts, xlab="", ylab="Log2 counts per million",las=2)
    } else if(selectorseqdata() == 2){
      plot <- boxplot(logcounts2, xlab="", ylab="Log2 counts per million",las=2)
    } else if(selectorseqdata() == 3){
      plot <- boxplot(logcounts3, xlab="", ylab="Log2 counts per million",las=2)
    } else {
      plot <- boxplot(logcounts4, xlab="", ylab="Log2 counts per million",las=2)
    }
  })
}
```

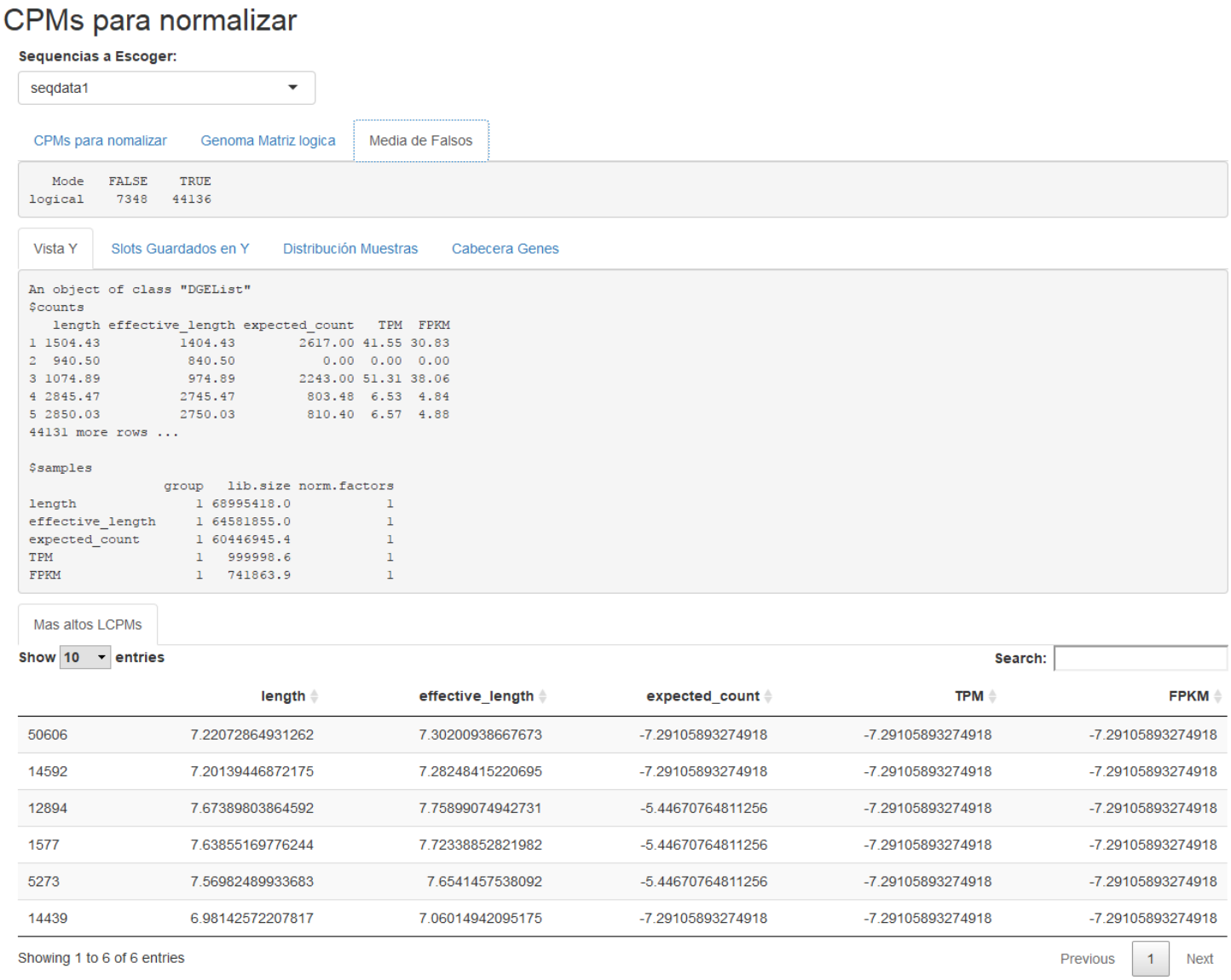
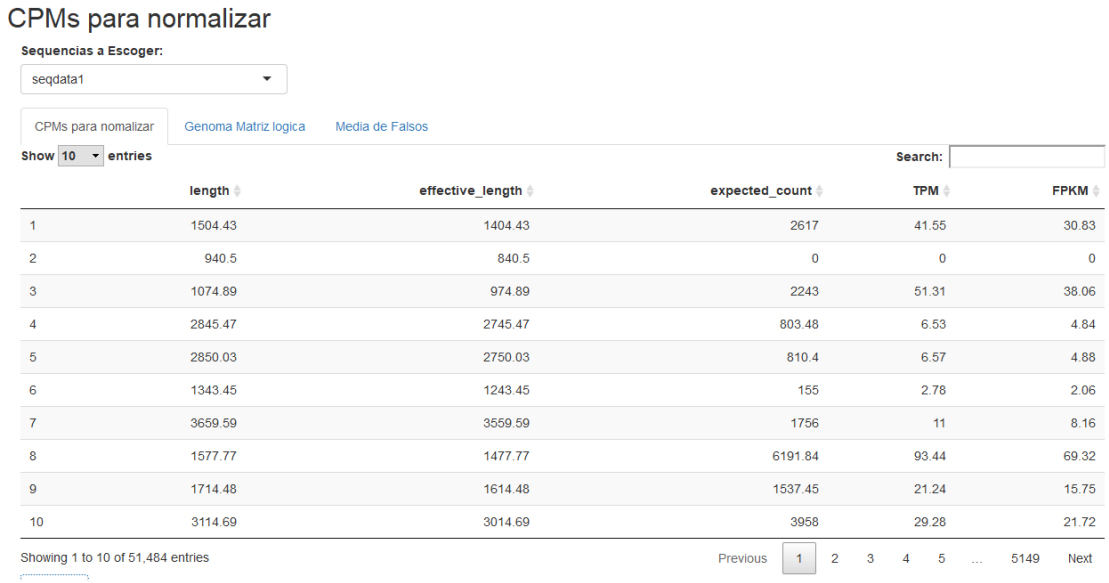


Figura 3: Vista general de la aplicación

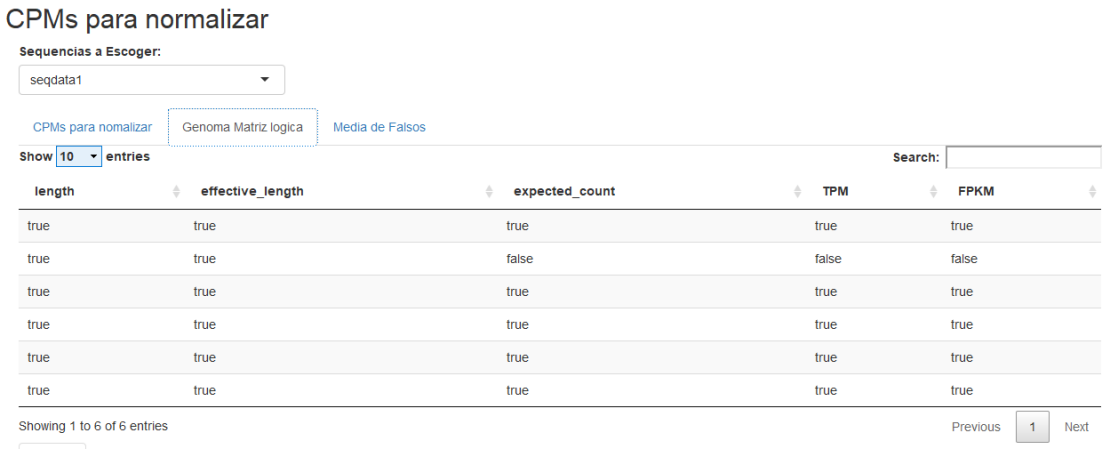
Tal como definimos en el UI hay 3 paneles definidos con tabsetPanel distribuidos verticalmente y 3 tabs en el primero, 4 tabs en el segundo, y 1 tab en el tercero. Cada tab se definió dentro de la función tabsetPanel y la subfunción tabPanel.



	length	effective_length	expected_count	TPM	FPKM
1	1504.43	1404.43	2617	41.55	30.83
2	940.5	840.5	0	0	0
3	1074.89	974.89	2243	51.31	38.06
4	2845.47	2745.47	803.48	6.53	4.84
5	2850.03	2750.03	810.4	6.57	4.88
6	1343.45	1243.45	155	2.78	2.06
7	3659.59	3559.59	1756	11	8.16
8	1577.77	1477.77	6191.84	93.44	69.32
9	1714.48	1614.48	1537.45	21.24	15.75
10	3114.69	3014.69	3958	29.28	21.72

Figura 4: CPMS para normalizar

Tabla con length , effective length, counts, TPM y FPKM en la que se han eliminado las dos primeras columnas de las anotaciones.



length	effective_length	expected_count	TPM	FPKM
true	true	true	true	true
true	true	false	false	false
true	true	true	true	true
true	true	true	true	true
true	true	true	true	true
true	true	true	true	true

Figura 5: Matriz lógica

Matriz lógica con valor True aquellos valores que superen los valores umbrales mínimos marcados durante el filtraje de los datos.

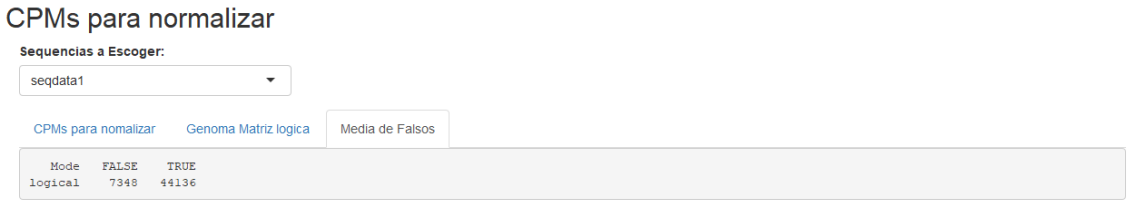


Figura 6:Total de falsos, donde no hay expresión y verdaderos donde sí hay expresión

Se marcan como trues los genes cuyos counts (expresión génica) se hayan considerado superiores al valor umbral y por lo tanto expresado.

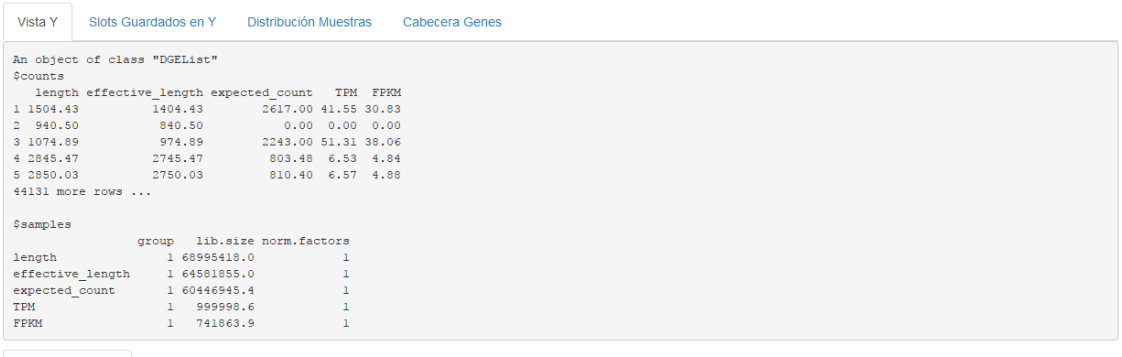


Figura 7: Descripción del objeto DGE

Descripción de los counts y de las muestras almacenados en el objeto “y” DGEList.

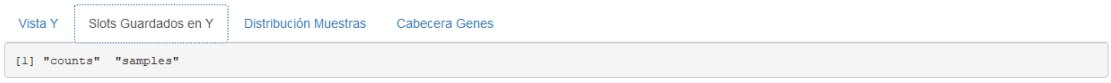


Figura 8:Objetos guardados en Y: counts y samples

Counts y samples guardados en el objeto Y DGEList.

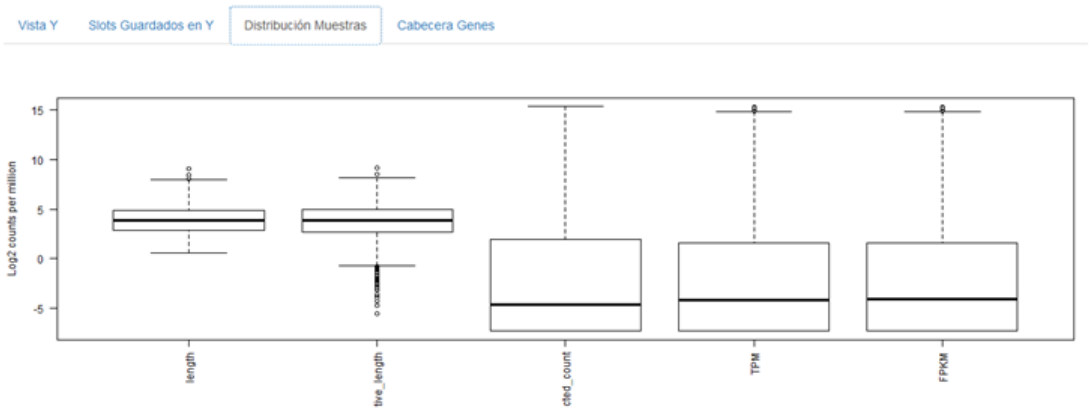


Figura 9: Diagramas de caja demostrando la distribución de la muestra

Se observa que para las 4 muestras los counts son bastante variables con una media negativa de -5log2 counts.

Figura 10: Cabecera genes

Medidas obtenidas de la varianza para los seis primeros genes en el objeto infogenes (utilizando la función head).

Mas altos LCPMs

Show 10 entries Search:

	length	effective_length	expected_count	TPM	FPKM
50606	7.22072864931262	7.30200938667673	-7.29105893274918	-7.29105893274918	-7.29105893274918
14592	7.20139446872175	7.28248415220695	-7.29105893274918	-7.29105893274918	-7.29105893274918
12894	7.67389803864592	7.75899074942731	-5.44670764811256	-7.29105893274918	-7.29105893274918
1577	7.63855169776244	7.72338852821982	-5.44670764811256	-7.29105893274918	-7.29105893274918
5273	7.56982489933683	7.6541457538092	-5.44670764811256	-7.29105893274918	-7.29105893274918
14439	6.98142572207817	7.06014942095175	-7.29105893274918	-7.29105893274918	-7.29105893274918

Showing 1 to 6 of 6 entries Previous 1 Next

Figura 11: cpm toptable

Tabla resumen con los genes más variables.

2.1.3 Beneficios económicos de un posible proyecto

Quedaría pendiente realizar un estudio de si daría beneficios crear un estudio o laboratorio en el cual se secuencie el DNA y RNA del cliente con dispositivos tan baratos como MinION (https://nanoporetech.com/products/minion) o otros más convencionales como IonProton. Para poder analizar los resultados gracias a un bioinformático, con aplicaciones mediante shiny por ejemplo, y que fuesen estudiados por personal médico y/o centros de investigación asociados. Las muestras podrían enviarse por correo o bien tomarse en el estudio / laboratorio (https://www.cnio.es/es/programas/secuenciacion/pedidos.asp).
En el caso de secuenciación de DNA, debería purificarse previamente por lo cual se necesitaría de un técnico de laboratorio experimentado o alguien cualificado desde MSC hasta PhD con experiencia en NGS y secuenciación del DNA.
Se necesitaría de una pequeña estructura computacional para el NGS:
Dos máquinas de:4TB de disco duro, 12 núcleos y 48Gb de ram cada una.
Se podría considerar un pequeño instituto de investigación y consultoría biomédica privado que realizase colaboraciones con otros institutos de investigación así como con instituciones académicas.

	1er año	2º año	3er año	
MinION kit o Ion Proton	2000-115.000 €	0 €	0 €	
Técnico de laboratorio /bioinformático	18.000 €	18.000 €	18.000 €	
phD o experto en secuenciación NGS y datos ómicos	22.000 €	22.000 €	22.000 €	
Médico interpretador de resultados	33.000 €	33.000 €	33.000 €	
Diseñador web	20.000 €	20.000 €	20.000 €	
recepcionista	18.000 €	18.000 €	18.000 €	
Alquiler del estudio lab	20.000 €	20.000 €	20.000 €	
Campana de extracción	100.000 €	0 €	0 €	
Equipo de laboratorio	20.000 €	20.000 €	20.000 €	
Infraestructura de computadores pequeña	10.000 €	0 €	0 €	
TOTAL	376.000 €	151.000 €	151.000 €	**TOTAL (3 años)** Plan A 678.000€ Plan B 632000€

Tabla 2: Gastos en tres años

La rentabilidad del centro/instituto de investigación y consultoría privada dependería no sólo de los contratos establecidos con otros centros o institutos de investigación y hospitales sino también de los genomas secuenciados a los clientes, además de las becas que proporciona tanto el gobierno, ayuntamiento y Generalitat, como la unión europea.

La viabilidad del proyecto es discutible puesto que ya existen centros que secuencian el genoma por un precio aproximado de mil euros el genoma. Suponiendo que se hiciese por el mismo precio que los otros centros, inicialmente, se necesitarían 678 clientes para que la empresa no haya supuesto ninguna pérdida. Además en el presupuesto no se han tenido en cuenta posibles inconvenientes, imprevistos o gestiones o gastos no planificados, como seguros tanto del equipo material como personal, el pago a la seguridad social por los empleados, el mantenimiento del equipamiento, la posible necesidad de aumentar el número de computadoras y/o de consultores bioinformáticos.

Para aumentar la viabilidad de un proyecto sería posible aumentar la inversión en computadoras y eliminar el ION Proton (y utilizar un MinION kit, si se requiere), eliminar la campana ya que supone un gasto inicial excesivo. Con ello se reduciría en 215000€ los gastos y se podría aprovechar mejor el presupuesto para una infraestructura de computadores mediana de 100000€. Se necesitarían alrededor de 50TB para el sistema de archivos distribuido, un clúster de 10 nodos de hasta 120 núcleos. A lo que habría que añadirle el salario de un informático para mantenerlo que rondaría los 23000€ (69000€ en 3 años). Por lo tanto el cambio supondría un abaratamiento de costes en 46000€ suponiendo un gasto total en 3 años de 632000€ (planB) En vez de técnico de laboratorio, el técnico podría ser bioinformático. Así se podrían realizar hasta estudios metagenómicos medios, montaje de genomas, y secuenciar más de unos cuantos exomas.

2.2 Discusión

2.2.1 Acerca de este experimento

El código R para la aplicación de este experimento podría ser optimizado, quizás en vez de cuadriplicar el código para cada una de las muestras se podría haber realizado un loop en el experimento. Ya que es interesante comparar muchas muestras a la vez como es el caso en la investigación en general.

A la hora del filtrado, se podría haber utilizado como umbral un CPM mínimo mayor de 0,5 counts ya que solo se estudiaron 4 muestras. Es sabido que cuanto mayor es la librería menor el umbral de CPM para considerarse como expresado. 0,5 es un valor umbral un poco bajo en este caso. Un valor más acertado podría ser de 1 por réplica (en este caso 1 o 4 depende de cómo se cuente).

Se hubiese habido de utilizar la anotación de los genes para poder introducirlo en las tablas de la aplicación ya que hubiese sido más intuitivo e informativo a primera vista.

Se utilizó el paquete edgeR ya que en la presencia de gran cantidad de counts solo DESeq y edgeRR son capaces de mantener una ratio razonable de falsos positivos sin pérdida de poder estadístico.[49] Se utilizó edgeR ya que en un principio se quería realizar un análisis de expresión diferencial y no estimar los niveles absolutos de expresión.

Se podría haber obtenido el control de calidad de las muestras con el gráfico MDS integrado en la aplicación shiny. Además de realizar un clustering jerárquico con heatmaps e integrarlo a la aplicación de shiny.

En la tabla los más altos CPMs habría que inspeccionar qué genes son los que dan alta variabilidad y gran varianza y averiguar por qué.

Quedaría por demostrar el perfil génico mediante otro método como FACS por ejemplo de las células con la expresión de los genes obtenida en el RNA-Seq.

No hubo tiempo de realizar la normalización de los datos. Pero tampoco hubo muestras de nódulos linfáticos sanos para comparar en el análisis de expresión diferencial que tenía pensado hacerse.

No se observaron batch effects o efectos de lote después del análisis de las muestras con RNA-Seq.

No se hizo ningún estudio de expresión diferencial en este caso. Se podría seguir el estudio, más allá del filtrado, pre-procesado y normalización de las

muestras en individuos con melanoma, con muestras de nódulos linfáticos sanos para realizar un análisis de expresión diferencial.
Quizás algunas lengüetas (tabs) del código de shiny como Slots guardados en Y podrían haber sido substituido por un plotMDS o si se hubiese podido realizar un heatmap.

2.2.2 Aspectos a investigar

Sería interesante realizar estudios proteómicos y metabolómicos paralelamente a los estudios transcriptómicos y genómicos[55]. Los estudios proteómicos se encargan de estudiar las funciones de las proteínas mediante la comprensión de su estructura y actividad. La metabolómica se caracteriza por estudiar los metabolitos. Se podrían determinar nuevos biomarcadores como metabolitos o proteínas en este tipo de estudios además de obtener una lista de posibles dianas farmacológicas. En el caso de realizar análisis proteómico, se podría realizar en diferentes situaciones temporales para determinar la evolución o los cambios proteómicos durante ese período ya que las proteínas expresadas varían en función del tiempo, en estadíos, tipos de cáncer y de las condiciones microambientales diferentes observadas. (http://www.ciber-bbn.es). Siempre comparando muestras de donantes sanos con muestras provenientes de pacientes presentando algún tipo de cáncer.
Además el Proyecto del Proteoma Humano estará a punto de finalizar y los subsecuentes proyectos como proyecto de proteoma en cáncer y humano (CANCER–HPP). Por lo que se podrá comparar el proteoma de una persona sana con las muestras directamente del paciente (o bien desde alguna base de datos como TCGA o gene expression omnibus) presentando algún tipo de enfermedad proinflamatoria, cáncer y alergias. Por lo que se tendrá más información acerca de las proteínas presentes en este tipo de procesos y se podrá saber con más precisión nuevas posibles dianas farmacológicas o dianas para otro tipo de terapias como la inmunoterapia. De hecho se podrá realizar un meta-análisis estadístico integrando los diferentes descubrimientos hasta ahora de los distintos tipos de ciencias ómicas que se pretenden interpretar.

2.3 Materiales y Métodos

Dataset: GSE93722
de https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2461009 web de gene expression omnibus de NCBI (https://www.ncbi.nlm.nih.gov/geo/).
Los datasets que corresponden a 4 pacientes son los siguientes:

LAU125
Melanoma metastásico, muestra del nódulo linfático Ilíaco.

LAU355
Nódulo linfático ilíaco-obturador

LAU1255
Melanoma metastásico, muestra del nódulo linfático axilar.

LAU1314
Melanoma metastásico, muestra del nódulo linfático Ilíaco-obturador.

Las muestras fueron analizadas con la siguiente maquinaria de ultrasecuenciación: Illumina HiSeq2500.

Pacientes de melanoma en estadío III donaron tejidos metastásicos, las muestras se obtuvieron de la disección de nódulos linfáticos de la axila derecha y de la región ilíaca y iliaca-obturadora[56]. Las suspensiones unicelulares se obtuvieron por disrupción mecánica y fueron criopreservadas en medio RPMI 1640 con 40% FCS y 10% DMSO. Las suspensiones unicelulares de los nódulos linfáticos fueron descongeladas y utilizadas para extracción de RNA para su secuenciación. Se usó un kit para eliminar las células muertas después de congelación (Miltenyi Biotech). El RNA se extrajo usando el kit RNAeasy Plus mini kit (Qiagen) siguiendo el protocolo. Se empezó con 200.000 células. RNA fue cuantificado y se analizó usando un analizador de fragmentos (Advanced Analytical). El RNA total de las muestras usado para su secuenciación obtuvo un RNA Quality Number (RQN) superior o igual a 7, siendo 1 el mínimo y 10 el máximo.
El análisis de expresión diferencial pretendía estudiarse enteramente con edgeR[57]

```
> sessionInfo()
R version 3.4.3 (2017-11-30)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows >= 8 x64 (build 9200)

Matrix products: default

locale:
[1] LC_COLLATE=Spanish_Spain.1252  LC_CTYPE=Spanish_Spain.1252
LC_MONETARY=Spanish_Spain.1252
[4] LC_NUMERIC=C                   LC_TIME=Spanish_Spain.1252

attached base packages:
[1] parallel  stats4    stats     graphics  grDevices utils
datasets  methods   base

other attached packages:
 [1] BiocInstaller_1.28.0 DT_0.2               shiny_1.0.5
knitr_1.17           edgeR_3.20.2
 [6] limma_3.34.0         GenomeInfoDb_1.14.0  IRanges_2.10.5
S4Vectors_0.14.7     BiocGenerics_0.24.0

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.14           compiler_3.4.3         bitops_1.0-6
tools_3.4.3
 [5] digest_0.6.13          bit_1.1-12             jsonlite_1.5
tibble_1.3.4
 [9] lattice_0.20-35        rlang_0.1.4            DBI_0.7
yaml_2.1.16
[13] GenomeInfoDbData_1.0.0 htmlwidgets_0.9        locfit_1.5-9.1
bit64_0.9-7
[17] grid_3.4.3             Biobase_2.38.0         R6_2.2.2
XML_3.98-1.9
[21] magrittr_1.5           blob_1.1.0             htmltools_0.3.6
rsconnect_0.8.5
[25] mime_0.5               xtable_1.8-2           httpuv_1.3.5
RCurl_1.95-4.8
```

3. Conclusiones

Con este trabajo se ha aprendidoa usar shiny para la creación de una aplicación en la que se visualizasen resultados de datos ómicos, se aprendió acerca del control de calidad para este tipo de datos transcriptómicos de RNA-seq. Se logró estudiar la gestión del proyecto y el tiempo.

Reflexión crítica:

En un principio se quería realizar todo el protocolo de análisis de expresión diferencial en un ejemplo de análisis de transcriptómica con microarrays y otro con RNA-seq en dos especies distintas.
Entonces se supuso que era preferible realizar la comparación de los resultados entre dos ejemplos de la misma especie. Además se recomendó utilizar una única y misma tecnología para el análisis transcriptómico: Affymetrix o Illumina, para poder realizar las comparaciones pertinentes de los resultados obtenidos en la misma unidad o ratio. Ya que se tenía pensado desde un principio comparar resultados de ambas tecnologías. Tras algún intento de escoger datasets adecuados encontramos uno pero por la falta de tiempo para la entrega del desarrollo del trabajo se realizó un trabajo menos extenso que el que inicialmente se había planteado. En vez de todo el análisis de expresión diferencial se propone realizar una aplicación para visualizar el control de calidad de los datos.

Análisis crítico:

Se organizó la planificación metodológica y escrupulosamente. Sin embargo aunque en un principio se pudo seguir, la planificación fue retrasándose para la creación de la aplicación en sí. Aunque se empezó a escribir la memoria a tiempo y antes de lo planeado, el hecho de no tener encarrilada la aplicación desde el principio hizo que se retrasasen otras partes. En parte se siguió más de la cuenta realizando aplicaciones de prueba en shiny en vez de centrarse en la aplicación del trabajo en sí.
Quizás se podría haber seguido otra metodología ya que se ha tenido que rehacer partes del trabajo o volver a empezar debido a que se empezó a realizar parte del trabajo cuando aún no se había confirmado con el tutor si se iba por el buen camino en cuanto a la selección de muestras por ejemplo. Se

pretendía escoger un código para el análisis de RNA seq y conseguir adaptarlo a una aplicación shiny. Viendo que esto no era posible se tuvieron que escoger otras muestras para poder analizar. Debido a estos últimos ajustes del trabajo se tuvo que acortar el objetivo en cuanto a cantidad de código para poder llegar a la fecha clave con un script funcional. Por ello se realizó un control de calidad de los datos (counts) y no se realizó ningún análisis de expresión diferencial. Porque además al mismo tiempo se debía de ir redactando la memoria, por ello se decidió acortar el estudio ya que si no, no hubiese habido tiempo de escribir un informe tan largo en tan poco tiempo. A pesar de ello se pudo desarrollar una aplicación shiny funcional.
Hubiese sido interesante poder utilizar Linux para el procesado de los datos y la ejecución completa del análisis de expresión diferencial ya que su potencia como sistema operativo agiliza el trato de datos en gran cantidad. Debido a la falta de otro ordenador con el sistema operativo Linux y dada la insuficiente capacidad del disco duro, en éste, para que una máquina virtual funcionase realizando una partición en el mismo, se decidió utilizar Windows y R (sistema operativo ya instalado).

Líneas de trabajo:

En este trabajo se tienen en cuenta unos datos de expresión génica provenientes del nódulo linfático. Se podría estudiar el transcriptoma del microambiente tumoral también.
Sería interesante investigar la relación del cáncer u otros tipos de cáncer con otras enfermedades proinflamatorias y alergias y comparar la expresión génica para buscar coincidencias de grupos. Durante este trabajo sólo se hizo el control de calidad que es el primero de los pasos para poder realizar seguidamente un análisis de expresión diferencial. Y en las muestras no hay un caso control por lo que para realizar un análisis de expresión diferencial se necesitarían de 4 muestras de nódulos linfáticos en personas sanas para poder compararlas a las 4 muestras de los pacientes con melanoma.
Sería interesante poder secuenciar el genoma de los pacientes (cada paciente) para poder inferir qué genes son los que están más o menos regulados en el conjunto de pacientes para poder determinar qué vías y qué dianas farmacológicas se podrían intervenir con nuevos fármacos. O si fuese posible con inmunoterapia.

4. Glosario

Definición de los términos y acrónimos más relevantes utilizados dentro de la Memoria.

AMP	Adenosina monofosfato
ATP	Adenosina trifosfato
APC	Células Presentadoras de Antígeno
STA	Specific Tumour Antigen
IL-10	Interleucina inmunosupresiva
TGF-β	Factor Transformante Beta
VEGF	Factor de crecimiento vascular endotelial
iMCs	Células mieloides inmaduras
sPS	Fosfatidilserina soluble
MHC	Complejo Mayor de Histocompatibilidad
RAG	Gen de Activación de Recombinación
IFN-γ,	Interferón Gamma
RencaHA	Cepa de células tumorales
NKT	Células T Asesinas naturales
NK	Natural Killer
TCR	Complejo Receptor T
Bcl-xl y Bcl-2[	Factores antiapoptóticos
PGE2	Prostaglandina E2
HLA	Antígenos Leucocitarios Humanos
Treg	Linfocito T regulador
Tconv	Linfocito T convenvional
Tr1	Linfocito T reguladora de tipo 1
NTPDasa	Nucleosido trifosfato difosfohydrolasa
AMPc	Adenosina monofosfato cíclico
PKA	Proteína Kinasa A
lipid rafts	Localización de membrana con alto contenido en colesterol y esfingolípidos (ácidos grasos

	saturados)
Lck	Tirosina quinasa no específica
Csk	Tirosina quinasa
Fyn	Proto-oncogen proteína tirosin quinasa
NF-AT	Factor Nuclear de transcripción de linfocitos T activados
CREB	Elemento de Respuesta y unión al AMPc (factor de transcripción)
UI	Interfaz de Usuario
SNPs	Polimorfismo de de nucleótido único
FPKM	Fragmentos por Kilobase por Millón
RPKM	Lecturas por Kilobase por Millón
TPM	Tránscritos por Kilobase por Millón
SNP	Polimorfismo de nucleótido simple
GWAS	Genome-Wide Assoaciation Study
IDE	Entorno de Desarrollo Integrado
GB	Gigabyte (10^9 bytes)
TB	Terabyte (10^{12} bytes)
CPM	Counts per million
MDS	Multi Dimensional Scaling
FACS	Fluorescence-Activated Cell Sorting
TCGA	The Cancer Genome Atlas
html	HiperText Markup Language
php	HypertextPreprocessor
mysql	Sistema de gestión de bases de datos relacional
RBP	RNA Bead Plate
RFP	RNA Fragmentation Plate

5. Bibliografía

[1] M. Arnold *et al.*, “Recent trends in incidence of five common cancers in 26 European countries since 1988: Analysis of the European Cancer Observatory,” *Eur. J. Cancer*, vol. 51, no. 9, pp. 1164–1187, Jun. 2015.

[2] S. I. Hajdu, “A note from history: Landmarks in history of cancer, part 1,” *Cancer*, vol. 117, no. 5, pp. 1097–1102, Mar. 2011.

[3] V. T. DeVita and S. A. Rosenberg, “Two Hundred Years of Cancer Research,” *N. Engl. J. Med.*, vol. 366, no. 23, pp. 2207–2214, Jun. 2012.

[4] D. Tarin, “Erratum to: Clinical and Biological Implications of the Tumor Microenvironment,” *Cancer Microenviron.*, vol. 5, no. 2, pp. 113–113, Aug. 2012.

[5] R. Kim, M. Emi, and K. Tanabe, “Cancer immunoediting from immune surveillance to immune escape,” *Immunology*, vol. 121, no. 1, pp. 1–14, May 2007.

[6] B. E. Himes *et al.*, “RNA-Seq transcriptome profiling identifies CRISPLD2 as a glucocorticoid responsive gene that modulates cytokine function in airway smooth muscle cells.,” *PLoS One*, vol. 9, no. 6, p. e99625, 2014.

[7] R. Kim, M. Emi, and K. Tanabe, “Cancer immunosuppression and autoimmune disease: beyond immunosuppressive networks for tumour immunity.,” *Immunology*, vol. 119, no. 2, pp. 254–64, Oct. 2006.

[8] C. N. Janicki, S. R. Jenkinson, N. A. Williams, and D. J. Morgan, “Loss of CTL Function among High-Avidity Tumor-Specific CD8+ T Cells following Tumor Infiltration,” *Cancer Res.*, vol. 68, no. 8, pp. 2993–3000, Apr. 2008.

[9] W. H. Fridman, F. Pagès, C. Sautès-Fridman, and J. Galon, “The immune contexture in human tumours: impact on clinical outcome,” *Nat. Rev. Cancer*, vol. 12, no. 4, pp. 298–306, Mar. 2012.

[10] H. L. Hanson *et al.*, “Eradication of established tumors by CD8+ T cell adoptive immunotherapy.,” *Immunity*, vol. 13, no. 2, pp. 265–76, Aug. 2000.

[11] S. Okada and T. Morishita, “The Role of Granulysin in Cancer Immunology,” *ISRN Immunol.*, vol. 2012, pp. 1–5, Jan. 2012.

[12] A. J. Brennan, J. Chia, J. A. Trapani, and I. Voskoboinik, “Perforin deficiency and susceptibility to cancer,” *Cell Death Differ.*, vol. 17, no. 4, pp. 607–615, Apr. 2010.

[13] I. Rousalova and E. Krepela, “Granzyme B-induced apoptosis in cancer cells and its regulation (review).,” *Int. J. Oncol.*, vol. 37, no. 6, pp. 1361–78, Dec. 2010.

[14] A. Shanker *et al.*, “Antigen Presented by Tumors In vivo Determines the Nature of CD8+ T-Cell Cytotoxicity,” *Cancer Res.*, vol. 69, no. 16, pp. 6615–6623, Aug. 2009.

[15] L. Gattinoni *et al.*, “Acquisition of full effector function in vitro paradoxically impairs the in vivo antitumor efficacy of adoptively transferred CD8+ T

cells," *J. Clin. Invest.*, vol. 115, no. 6, pp. 1616–1626, Jun. 2005.

[16] M. T. Spiotto, D. A. Rowley, and H. Schreiber, "Bystander elimination of antigen loss variants in established tumors," *Nat. Med.*, vol. 10, no. 3, pp. 294–298, Mar. 2004.

[17] O. J. Finn, "Immuno-oncology: understanding the function and dysfunction of the immune system in cancer," *Ann. Oncol.*, vol. 23, no. suppl 8, p. viii6-viii9, Sep. 2012.

[18] J. J. Listopad *et al.*, "Fas expression by tumor stroma is required for cancer eradication.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, no. 6, pp. 2276–81, Feb. 2013.

[19] F. Garrido, I. Algarra, and A. M. García-Lora, "The escape of cancer from T lymphocytes: immunoselection of MHC class I loss variants harboring structural-irreversible 'hard' lesions," *Cancer Immunol. Immunother.*, vol. 59, no. 10, pp. 1601–1606, Oct. 2010.

[20] B. Seliger, "Novel insights into the molecular mechanisms of HLA class I abnormalities," *Cancer Immunol. Immunother.*, vol. 61, no. 2, pp. 249–254, Feb. 2012.

[21] S. Burugu, A. R. Dancsok, and T. O. Nielsen, "Emerging targets in cancer immunotherapy," *Semin. Cancer Biol.*, Oct. 2017.

[22] J. Crespo, H. Sun, T. H. Welling, Z. Tian, and W. Zou, "T cell anergy, exhaustion, senescence, and stemness in the tumor microenvironment," *Curr. Opin. Immunol.*, vol. 25, no. 2, pp. 214–221, Apr. 2013.

[23] A. M. White and D. C. Wraith, "Tr1-Like T Cells - An Enigmatic Regulatory T Cell Lineage.," *Front. Immunol.*, vol. 7, p. 355, 2016.

[24] T. L. Whiteside, "What are regulatory T cells (Treg) regulating in cancer and why?," *Semin. Cancer Biol.*, vol. 22, no. 4, pp. 327–334, Aug. 2012.

[25] "Tumor metabolism as modulator of immune response and tumor progression," *Semin. Cancer Biol.*, vol. 22, no. 4, pp. 335–341, Aug. 2012.

[26] F. S. Regateiro, S. P. Cobbold, and H. Waldmann, "CD73 and adenosine generation in the creation of regulatory microenvironments," *Clin. Exp. Immunol.*, vol. 171, no. 1, pp. 1–7, Jan. 2013.

[27] V. C. Liu *et al.*, "Tumor evasion of the immune system by converting CD4+CD25- T cells into CD4+CD25+ T regulatory cells: role of tumor-derived TGF-beta.," *J. Immunol.*, vol. 178, no. 5, pp. 2883–92, Mar. 2007.

[28] M. V. Sitkovsky *et al.*, "P hysiological C ontrol of I mmune R esponse and I nflammatory T issue D amage by H ypoxia -I nducible F actors and A denosine A_{2A} R eceptors," *Annu. Rev. Immunol.*, vol. 22, no. 1, pp. 657–682, Apr. 2004.

[29] A. Facciabene *et al.*, "Tumour hypoxia promotes tolerance and angiogenesis via CCL28 and Treg cells," *Nature*, vol. 475, no. 7355, pp. 226–230, Jul. 2011.

[30] B. B. Fredholm, A. P. IJzerman, K. A. Jacobson, J. Linden, and C. E. Muller, "International Union of Basic and Clinical Pharmacology. LXXXI. Nomenclature and Classification of Adenosine Receptors--An Update," *Pharmacol. Rev.*, vol. 63, no. 1, pp. 1–34, Mar. 2011.

[31] S. Huang, S. Apasov, M. Koshiba, and M. Sitkovsky, “Role of A2a extracellular adenosine receptor-mediated signaling in adenosine-mediated inhibition of T-cell activation and expansion.,” *Blood*, vol. 90, no. 4, pp. 1600–10, Aug. 1997.

[32] D. Lukashev, M. Sitkovsky, and A. Ohta, “From "Hellstrom Paradox" to anti-adenosinergic cancer immunotherapy.,” *Purinergic Signal.*, vol. 3, no. 1–2, pp. 129–34, Mar. 2007.

[33] R. Mosenden and K. Taskén, “Cyclic AMP-mediated immune regulation — Overview of mechanisms of action in T cells,” *Cell. Signal.*, vol. 23, no. 6, pp. 1009–1016, Jun. 2011.

[34] N. Zhang and M. J. Bevan, “CD8+ T Cells: Foot Soldiers of the Immune System,” *Immunity*, vol. 35, no. 2, pp. 161–168, Aug. 2011.

[35] J. Linden and C. Cekic, “Regulation of lymphocyte function by adenosine.,” *Arterioscler. Thromb. Vasc. Biol.*, vol. 32, no. 9, pp. 2097–103, Sep. 2012.

[36] T. Seremet *et al.*, “Illustrative cases for monitoring by quantitative analysis of BRAF/NRAS ctDNA mutations in liquid biopsies of metastatic melanoma patients who gained clinical benefits from anti-PD1 antibody therapy,” *Melanoma Res.*, p. 1, Dec. 2017.

[37] J. Bastid *et al.*, “Inhibition of CD39 Enzymatic Function at the Surface of Tumor Cells Alleviates Their Immunosuppressive Activity,” *Cancer Immunol. Res.*, vol. 3, no. 3, 2015.

[38] S. Shrestha *et al.*, “Integrated MicroRNA–mRNA Analysis Reveals miR-204 Inhibits Cell Proliferation in Gastric Cancer by Targeting CKS1B, CXCL1 and GPRC5A,” *Int. J. Mol. Sci.*, vol. 19, no. 1, p. 87, Dec. 2017.

[39] V. Raškevičius *et al.*, “Genome scale metabolic models as tools for drug design and personalized medicine,” *PLoS One*, vol. 13, no. 1, p. e0190636, Jan. 2018.

[40] A. M. Horning *et al.*, “Single-cell RNA-seq reveals a subpopulation of prostate cancer cells with enhanced cell cycle-related transcription and attenuated androgen response,” *Cancer Res.*, p. canres.1924.2017, Dec. 2017.

[41] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, “Quantitative monitoring of gene expression patterns with a complementary DNA microarray.,” *Science*, vol. 270, no. 5235, pp. 467–70, Oct. 1995.

[42] L. Chen *et al.*, “Correlation between RNA-Seq and microarrays results using TCGA data,” *Gene*, vol. 628, pp. 200–204, Sep. 2017.

[43] W. Zhang *et al.*, “Comparison of RNA-seq and microarray-based models for clinical endpoint prediction,” *Genome Biol.*, vol. 16, no. 1, p. 133, Dec. 2015.

[44] Z. Wang, M. Gerstein, and M. Snyder, “RNA-Seq: a revolutionary tool for transcriptomics.,” *Nat. Rev. Genet.*, vol. 10, no. 1, pp. 57–63, Jan. 2009.

[45] Illumina, “Transitioning from Microarrays to mRNA-Seq,” 2000.

[46] M. Sultan *et al.*, “Influence of RNA extraction methods and library selection schemes on RNA-seq data,” *BMC Genomics*, vol. 15, no. 1, p.

675, Aug. 2014.

[47] Illumina, “FOR RESEARCH USE ONLY TruSeq ® RNA Sample Preparation v2 Guide,” 2014.

[48] S. Ruddy, “edgeR Tutorial: Differential Expression in RNA-Seq Data,” 2011.

[49] M. A. Dillies *et al.*, “A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis,” *Brief. Bioinform.*, vol. 14, no. 6, pp. 671–683, Nov. 2013.

[50] G. Nyamundanda, P. Poudel, Y. Patil, and A. Sadanandam, “A Novel Statistical Method to Diagnose, Quantify and Correct Batch Effects in Genomic Studies,” *Sci. Rep.*, vol. 7, no. 1, p. 10849, Dec. 2017.

[51] Y. Xie, R. Wang, and J. Zhu, “Construction of breast cancer gene regulatory networks and drug target optimization,” *Arch. Gynecol. Obstet.*, vol. 290, no. 4, pp. 749–755, Oct. 2014.

[52] P. Charoentong *et al.*, “Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade,” *Cell Rep.*, vol. 18, no. 1, pp. 248–262, Jan. 2017.

[53] C.-Y. Ock *et al.*, “Pan-Cancer Immunogenomic Perspective on the Tumor Microenvironment Based on PD-L1 and CD8 T-Cell Infiltration,” *Clin. Cancer Res.*, vol. 22, no. 9, pp. 2261–2270, May 2016.

[54] A. Jiménez-Sánchez *et al.*, “Heterogeneous Tumor-Immune Microenvironments among Differentially Growing Metastases in an Ovarian Cancer Patient,” *Cell*, vol. 170, no. 5, p. 927–938.e20, Aug. 2017.

[55] C. Sandhu, A. Qureshi, and A. Emili, “Panomics for Precision Medicine,” *Trends Mol. Med.*, Dec. 2017.

[56] J. Racle, K. de Jonge, P. Baumgaertner, D. E. Speiser, and D. Gfeller, “Simultaneous Enumeration Of Cancer And Immune Cell Types From Bulk Tumor Gene Expression Data,” *bioRxiv*, p. 117788, Mar. 2017.

[57] A. T. L. Lun, Y. Chen, G. K. Smyth, and A. 2015, “It’s DE-licious: a recipe for differential expression analyses of RNA-seq experiments using quasi-likelihood methods in edgeR.”

shiny.rstudio.com/ 01/10/2017

bioconductor.org/ 30/10/2017

www.who.int 01/11/2017

www.cancer.gov 06/11/2017

chagall.med.cornell.edu/RNASEQcourse/Intro2RNAseq.pdf 08/11/2017

www.rna-seqblog.com/ 08/11/2017

www.bioconductor.org/help/workflows/RNAseq123/ 08/11/2017

combine-australia.github.io/RNAseq-R/ 08/11/2017

bioinformatics-core-shared-training.github.io/RNAseq-R/ 08/11/2017

biocluster.ucr.edu/~rkaundal/workshops/R_mar2016/RNAseq.html 08/11/2017

bioconductor.org/packages/release/bioc/html/TCGAbiolinks.html 08/11/2017

www.biostars.org/p/219024/ 08/11/2017

cran.r-project.org/package=TCGAretriever 08/11/2017

www.researchgate.net/post/How_to_get_TCGA_data 08/11/2017

www.ciber-bbn.es 27/12/2017

www.labx.com/product/illumina-sequencer 08/01/2018

6. Anexos

- Dynamic Documents with R and knitr (Chapman & Hall/CRC The R Series)1st Edition
  Yihui Xie

- Learning Shiny, 2015
  Hernan G. Resnizky

- Web Application Development with R Using Shiny
  Harness the graphical and statistical power of R and
  rapidly develop interactive user interfaces using the
  superb Shiny package
  Chris Beeley

- The R Book
  Michael J. Crawley

- Hands-On Programming with R
  Garrett Grolemund

- Advanced R
  Hadley Wickham

- edgeR: differential expression analysis of digital gene expression data User's Guide
  Yunshun Chen, Davis McCarthy,
  Matthew Ritchie, Mark Robinson, Gordon K. Smyth
  First edition 17 September 2008
  Last revised 30 June 2016

- Hospital Universitari Vall d'Hebron
  Institut de Recerca - VHIR
  Institut d'Investigació Sanitària de l'Instituto de Salud Carlos III (ISCIII) Bioinformatics for Biomedical Research http://ueb.vhir.org/201BBR Alex Sánchez 11/11/2015 Next Generation Sequencing Technologies, Application, Hardware, Software

- Hospital Universitari Vall d'Hebron
  Institut de Recerca - VHIR Institut d'Investigació Sanitària de l'Instituto de Salud Carlos III (ISCIII)
  Bioinformatics for Biomedical Research
  http://ueb.vhir.org/201BBR
  Alex Sánchez
  11/11/2015 From raw images to preprocessed data

- RNA-seq Data Analysis: A Practical Approach (Chapman & Hall/CRC Mathematical and Computational Biology)
  De Eija Korpelainen,Jarno Tuimala, Panu Somervuo, Mikael Huss, Garry Wong

- R Graphics Cookbook
  Winston Chang