UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
DOCTORADO EN CIENCIAS BIOMÉDICAS
LABORATORIO INTERNACIONAL DE INVESTIGACIÓN SOBRE EL GENOMA HUMANO

IDENTIFICACIÓN DE LOS PRINCIPALES GENES INVOLUCRADOS EN PATOLOGÍAS PULMONARES POR MEDIO DE ANÁLISIS INTEGRATIVO

EXAMEN DE TITULACIÓN
QUE PARA OPTAR POR EL GRADO DE
DOCTORA EN CIENCIAS BIOMÉDICAS

PRESENTA:
ANA BEATRIZ VILLASEÑOR ALTAMIRANO

TUTOR PRINCIPAL: DRA. ALEJANDRA MEDINA RIVERA
LABORATORIO INTERNACIONAL DE INVESTIGACIÓN SOBRE EL GENOMA HUMANO, UNAM

CO-TUTOR: DR. JULIO COLLADO VIDES
CENTRO DE CIENCIAS GENÓMICAS, UNAM

CO-TUTOR: DR. MOISÉS SELMAN LAMA
INSTITUTO NACIONAL DE ENFERMEDADES RESPIRATORIAS

JURIQUILLA, QUERÉTARO, MÉXICO. MAYO 2021

Agradecimientos

Al posgrado de ciencias biomedicas de la UNAM. A la UNAM, al campus Juriquilla, al LIIGH y al LAVIS.

Agradezco a mi tutora y mentora, Alejandra Medina Rivera sin ella esto no habría sido posible.

A mis co-tutores Julio Collado y Moisés Selman por todo el apoyo, consejo y tiempo que me otorgaron. En especial a Yalbi Balderas por ser parte no oficial pero imprescindible en mi formación.

A mis compañeros de laboratorio que estuvieron y están en el RegGenLab en especial a Karen Núñez.

A todos mis compañeros del LIIGH con los que me tocó convivir durante mi doctorado en especial a Miriam Bravo.

A todos los profesores del LIIGH, en especial a Daniela Robles, Diego Ortega, María Ávila y Lucía Morales, por formar parte en mi educación.

A todos los técnicos Luis, Alejandra, Karina y en especial a Jair por todos sus consejos y apoyo durante está tésis desde el inicio. A los profesores que me dieron clase y los que me ayudaron en está investigación.

A CONACYT por el apoyo económico.

- CONACYT, Infraestructura 269449.
- Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica – Universidad Nacional Autónoma de México (PAPIIT-UNAM) [IA206517, IA201119, IA203021].
- CONACYT FORDECYT-PRONACES 11311.
- CONACYT “Fronteras de la Ciencia” 5.
- Estímulos a Investigaciones Médicas “Miguel Alemán Valdés”

A la Fundazione Edmud Mach por recibirme en mi estancia de investigación, a Kristof Engelen, Paolo Sonego y en especial a Marco Moretto por todo su apoyo, grazie mille.

A Dana-Farber Institute por recibirme en mi estancia de investigación, en especial a Rafael Irrizary, Patrick Kilmes y Alejandro Reyes, thank you so much.

A todas las instituciones como Bioconductor, ATS, Rladies, CDSB, RMB y personas que a lo largo del doctorado me dieron apoyo moral, guía en mi educación y ayuda económica para becas, congresos y cursos para mi mejoramiento académico.

A las secretarias de la UNAM, en especial a Denny y a Carmelita que sin su ayuda no me hubiera podido graduar por la burocracia.

.... Familia y amigos....

Agradezco a mi hermosa madre y padre por cada momento a su lado, gracias a ustedes soy lo que soy. A mi hermano Carlos por todas las platicas con tono informático a altas horas de la noche, por su amistad y apoyo incondicional. A mi hermanita Leticia por ser mi mejor amiga, por mantenerme sana emocionalmente y ser mi apoyo en todo momento. Los amo mucho familia.

A mi familia materna, los Altamirano, a mi familia paterna, los Villaseñor. Que orgullo ser Villaseñor-Altamirano.

A todos mis amigos, a los de siempre, a los que eran y a los que se convirtieron. Todos formarom parte importante en mi vida. Gracias por su ayuda en momentos difíciles, en la salud, en la enfermedad, en el desamor, en la felicidad, en la angustia, en la emoción y en su apoyo para desarrollarme como científica mexicana. Gracias de todo corazón.

Abstract

The most common platforms to evaluate gene expression on a high-throughput scale are microarrays and RNA sequencing which measure levels of RNA from almost all genes in an organism. Transcriptomics has allowed us to understand molecular pathways and differences in a variety of conditions and human disorders such as lung diseases. However, integrating different experiments from different laboratories and platforms in order to have a robust profile is challenging due to batch effects and confounders.

Here, we used lung diseases (*e.i.* Chronic Obstructive Pulmonary Disease; COPD and Idiopathic Pulmonary Fibrosis; IPF) to generate a repository of lung experiments called PulmonDB that are downloaded, stored in a relational database, manually curated (with a consistent gene annotation), and checked for quality control. We used PulmonDB to analyze robustness across experiments, differences and similarities between COPD and IPF, and to analyze the regulation of Hedgehog Interacting Protein (HHIP) signal in COPD samples.

Resumen

Las plataformas más comunes para evaluar la expresión génica en una escala masiva son los microarreglos y la secuenciación de RNA donde se mide los niveles de RNA en prácticamente todos los genes de un organismo. La transcriptómica nos ha permitido comprender las vías moleculares y las diferencias en una variedad de condiciones y trastornos humanos, como las enfermedades pulmonares. Sin embargo, la integración de diferentes experimentos de diferentes laboratorios y plataformas para tener un perfil robusto es un desafío debido a los efectos de lote y factores confusos.

En este estudio, usamos enfermedades pulmonares (*e.i.* Enfermedad Pulmonar Obstructiva Crónica; EPOC y Fibrosis Pulmonar Idiopática; FPI) para generar un repositorio de experimentos pulmonares nombrado PulmonDB que se descargan, almacenan en una base de datos relacional, se curan manualmente (con una anotación genética consistente) y se verifican que pasen un control de calidad. Se usó PulmonDB para analizar la robustez de los datos entre experimentos, las diferencias y similitudes entre la EPOC y FPI, a demás de analizar la regulación de la proteína Hedgehog Interacting Protein (HHIP) en muestras de EPOC.

Índice general

Índice de tablas **13**

Índice de figuras **15**

Información extra **19**

1 Introdución **21**
1.1 Transcriptómica 21
1.2 Enfermedades pulmonares 22
1.3 Antecedentes 35

2 Objetivos **39**
2.1 Hipótesis 39
2.2 Objetivo General 39
2.3 Objetivos específicos 39

3 Base de datos: PulmonDB **41**
3.1 Artículo: COMMAND 41
3.2 Artículo: PulmonDB 53

4 Robustez en PulmonDB **65**
4.1 Metodología 66
4.2 Resultados 68
4.3 Conclusión 70

5 Heterogeneidad en EPOC y FPI **73**
5.1 Introducción 73
5.2 Metodología 76
5.3 Resultados 78
5.4 Composición celular de las muestras en IPF 78
5.5 Composición celular de las muestras en EPOC 84
5.6 Conclusiones 86

6 Regulación de HHIP **89**
6.1 Introducción 89

6.2 Metodología 90
6.3 Resultados 92
6.4 Conclusiones 96
6.5 Poster en ISMB 97

7 Conclusiones **99**
7.1 Perspectivas 100

Apéndice: Técnicas de secuenciación y expresión génica a nivel ómico **103**
Contribución personal 103
Publicación 1: *Review of 'omics-level gene sequencing and gene expression techniques* 104

Bibliografía **141**

Índice de tablas

4.1 Información de los experimentos usados. EPOC: Enfermedad Pulmonar Obstructiva Crónica; AAD: Deficiencia de alfa-1-antitripsina; ILD: Enfermedades intersticiales 67

5.1 Datos de bulk RNA-seq 77

Índice de figuras

1.1 Características de la elasticidad pulmonar. La distensibilidad pulmonar y el retroceso elástico pulmonar son parte del mecanismo de ventilación pulmonar. La distensibilidad es la cualidad de poder inflar y expandir los pulmones (izquierda-rojo). Mientras que el retroceso elástico pulmonar es el poder retraerse a su tamaño original después de inhalar (derecha-verde). Figura creada usando BioRender.com y generada por Mauricio Guzmán. 23

1.2 Gráfica de presión transpulmonar (cm H_2O) contra volumen pulmonar (L). Esta gráfica representa la distensibilidad pulmonar, la cual es inversa a la elasticidad (Pierce et al., 2005b). La presión transpulmonar está dada por la diferencia entre la presión intrapulmonar y la presión en la pleura, representando la fuerza ejercida para la dilatación pulmonar (Grieco et al., 2017). La gráfica presenta las curvas representativas de pacientes con fibrosis, enfisema y la curva normal en diferentes colores. El punto máximo indica la capacidad total del pulmón (*TLC*), la cual varia dependiendo de la patología. Figura modificada de (Hammer et al., 2015). 25

1.3 Figura con valores de *FEV1* y *FVC* de enfermedades obstructivas (enfisema) y restrictivas (fibrosis). El eje x representa el tiempo al momento de la expiración, en el eje y se gráfica el volumen en litros (L). En la gráfica se muestra el tiempo transcurrido en un segundo (|1sec|) y las medidas del volumen espiratorio forzado en un segundo (FEV) y la capacidad vital forzada (FVC). Se puede observar el cambio en las gráficas con respecto a los valores normales (A) en las enfermedades obstructivas (B), en las restrictivas (C). Tomada de (Pierce et al., 2005b). 26

1.4 Diagrama de toma de decisiones basado en la relación *FEV1/FVC*. Figura modificada de (Paraskeva et al., 2011). Volumen Espiratorio forzado el primer segundo (FEV1). Capacidad Vital Forzada (FVC). Capacidad de difusión de transferencia del monóxido de carbono (DLCO). Capacidad pulmonar total (TLC). Límite inferior del rango normal (LLN) en una población específica. . 27

1.5 Patologías de la EPOC (Tomada de (Barnes et al., 2015)). EPOC se caracteriza principalmente por tener enfisema (imagen superior), donde la estructura alveolar se rompe. Asimismo, presenta bronquitis crónica, donde los bronquios se hiper inflaman (imagen inferior), existe secreción de moco, se cierran las vías aéreas, lo que ocasiona de manera conjunta un aumento en el retroceso elástico pulmonar. 30

1.6 Imagen con los factores de riesgo asociados a EPOC. En la izquierda se encuentran algunos factores ambientales (representando el humo de tabaco, contaminación, polvos y químicos) y en la derecha los factores genéticos. AAT es alfa-1 antitripsina, SNPs es *Single Nucleotide Polymorfism*. 32
1.7 Imagen representando la FPI (Tomada de (Martinez et al., 2017)). La FPI es una enfermedad que se caracteriza por dilatación en los bronquios, remodelación de la matriz extra-celular y presencia de fibrosis en el intersticio pulmonar, lo que provoca un aumento en el retroceso elástico pulmonar y mayor dificultad para el recambio de O_2. 34
1.8 Figura de experimentos públicos asociados a EPOC. En el eje de las x se encuentran los años hasta la fecha. En el eje y la cantidad de experimentos medido en número de identificadores GSEs que estaban disponibles para ese año. 37
1.9 Figura de experimentos públicos asociados a FPI En el eje de las x se encuentran los años hasta Enero 15, 2021. En el eje y la cantidad de experimentos medido en número identficadores de GSEs que estaban disponibles para ese año. 38

4.1 Definiciones de reproducibilidad, replicabilidad, robustez y generalizado. Figura, tomada de la sección *Definitions* el libro de *Turing Way* (Arnold et al., 2019). 66
4.2 Diagrama de la metodología usada para explorar la robustez de PulmonDB. Como primer paso fue la selección de experimentos que contrastan muestras de EPOC contra muestras controles en tejido pulmonar. Después la obtención de datos normalizados, para seguir 1) el flujo de trabajo clásico se usaron los datos crudos y se normalizaron utilizando RMA, 2) los datos normalizados de PulmonDB se obtuvieron usando su paquetería de R. Luego se realizó la expresión diferencial por experimento y después se compararon los datos obtenidos de RMA con los que se obtuvieron de PulmonDB. 68
4.3 Diagrama de dispersión entre los datos normalizados utilizando el flujo de trabajo típico y los contrastes de muestras normalizados en PulmonDB para la muestra GSE27597. En el eje X se observan los resultados de *logFC* de PulmonDB, en el eje Y los resultados de *logFC* usanndo el flujo de trabajo típico. En color azul se encuentra la regresión linear y el coeficiente de correlación se indica en la esquina superior izquierda. 69
4.4 Diagrama de dispersión para todos los experimentos selecionados. El eje X tiene los resultados de *logFC* para PulmonDB, el eje Y los resultados de *logFC* obtenidos usando el flujo de trabajo clásico. La línea azul representa la regresión lineal. 70

5.1 Tipos celulares en el pulmón, modificada de (Zepp and Morrisey, 2019). a) representación del pulmón que se encuentra dividido en b) Tráquea y vías aéreas largas (Bronquios), c) Vías aéreas medias y bajas (Bronquiolos) y d) alvéolos, cada una de las regiones con diferente composición celular. 74

5.2 Figura de secuenciación *bulk* contra *single-cell* obtenida de (Sheila-10x, 2017). Se representa un tejido compuesto de diversas células, las cuales son analizadas con *single-cell* (arriba) o *bulk* (abajo) y como la expresión genética obtenida de ambas tecnologías representa la expresion de un tipo celular o un promedio de todo. 75
5.3 Figura general de MuSiC modificada de (Wang et al., 2019). El primer paso es calcular similitudes usando agrupamiento jerárquico (clustering jerárquico). Después se obtienen las proporciones priorizando los genes con mayor peso. Esto se hace iterativamente hasta obtener la proporción de todos los tipos celulares. 76
5.4 Árbol con agrupamiento jerárquico. Se calculó similitud entre tipos celulares para la matriz completa (*Cluster log (Design Matrix)*) o para la matriz del promedio entre individuos (*Cluster log(Mean of RA)*).Los 7 clusters o grupos (línea punteada roja) están indicados por colores: C1: *Lymphatic* (azul), C2: *Fibroblast* (morado), C3: *NK cell* (verde), C4: *Endothelium* (rojo), C5: *Ciliated* (aqua), C6: *T cell* y *Mast cell* (rosa), C7: *Type 1*, *Secretory*, *Transformed epithelium*, *Type 2*, *Macrophages* y *B cell* (amarillo-verde). 79
5.5 Proporción de tipos celulares en FPI obtenidos por deconvolución. Porcentajes de tipos célulares para el experimento **GSE52463**, el cual tiene muestras control (covariable en ocre) y de FPI (covariable en verde). El porcentaje está indicado en azul, entre más obscuro mayor proporción y más claro indica menor proporción. Tanto las muestras como los tipos celulares están agrupados jerárquicamente. 81
5.6 Comparaciones entre las proporción de tipos celulares en FPI obtenidos por deconvolución. El p-valor de la prueba Wilcoxon se muestra en cada gráfica. Verde es el grupo de pacientes con FPI y ocre los controles. 82
5.7 Varianza de las proporciones celulares para los datos de la FPI. La varianza entre pacientes calculada por tipo celular y mostrada por grupo. El p-valor de la prueba Wilcoxon se muestra en la gráfica. Verde es el grupo de pacientes con FPI y ocre los controles. 83
5.8 Proporción de tipos celulares en EPOC obtenidos por deconvolución. Porcentajes de tipos celulares para el experimento **GSE57148**, el cual tiene muestras control (covariable en ocre) y de la EPOC (covariable en rojo). El porcentaje está indicado en azul, entre más obscuro mayor proporción y más claro indica menor proporción. Tanto las muestras como los tipos celulares están agrupados jerárquicamente. 84
5.9 Comparaciones entre las proporción de tipos celulares en COPD obtenidos por deconvolución. El p-valor de la prueba Wilcoxon se muestra en cada gráfica. Verde es el grupo de pacientes con FPI y ocre los controles. 85
5.10 Varianza de las proporciones celulares para los datos de la EPOC. La varianza entre pacientes calculada por tipo celular y mostrada por grupo. El p-valor de la prueba Wilcoxon se muestra en la gráfica. Verde es el grupo de pacientes con FPI y ocre los controles. 87

6.1 Descripción de HHIP y variantes asociadas a la EPOC. 92

6.2 Expresión de HHIP en datos de PulmonDB. Muestras de tejido pulmonar de los 8 estudios selecionados se usaron para graficar la expresión de HHIP. Blanco representa a las muestras control, rojo a las muestras de EPOC y verde a muestras con FPI. 93
6.3 Regulones de *ATF3* (MA0018) y *HLF* (MA0043). Los regulones de *ATF3* y *HLF* también cuentan con variantes que modifican el motivo transcripcional. *ATF3* y *HLF* son factores transcripcionales, usando pySCENIC se calculó los módulos de regulación de dichos genes. Los genes mostrados en la figura son los modulos de regulación resultantes de pySCENIC respectivamente. 94
6.4 Motivo de *HLF* (MA0043) y la variante **rs1032295**. Se muestra la ubicación en la región río arriba de *HHIP* (chr 4, 144513432); los resutltados de *Variation-scan* para cada variante; y su interpretación en el contexto de pacientes con la EPOC (o *COPD* por sus siglas en inglés). 95
6.5 Motivo de *ATF3* (MA0018) y la variante **rs12509311**. Se muestra la ubicación en la región río arriba de *HHIP* (chr 4, 144557510); los resutltados de *Variation-scan* para cada variante; y su interpretación en el contexto de pacientes con la EPOC (o *COPD* por sus siglas en inglés). 96

Información extra

Está tésis fue escrita en **Markdown** usando la paquetería de **blookdown**. El código para generar la tésis se encuentra en Github privado (si se desea, favor de solicitar acceso):

```
git clone https://github.com/AnaBVA/PhdThesis.git
```

Cuenta con:

-PDF

-Word

-Html

Los cuales se pueden descargar de: Drive

Capítulo 1

Introdución

1.1 Transcriptómica

El ácido ribonucleico (RNA) son macromoléculas generadas a partir de platillas de ácido desoxirribonucleico (DNA) por un proceso llamado transcripción (Liang, 2013, Cramer (2019)). La transcripción es un proceso complejo que involucra varias moléculas pero una vez que la maquinaria necesaria se encuentra accesible, se reconoce la región promotora del gen en regiones disponibles del DNA y se transcriben las moléculas de RNA (Cramer, 2019, Ong and Corces (2014), Bell et al. (2011)).

El concepto de gene ha cambiado, en un principio se definió como la unidad de herencia genética pero su significado fue evolucionando con el paso del tiempo (Gerstein et al., 2007). Gerstein, Mark B, et. al definieron el gen como *"una unión de secuencias genómicas que codifican un conjunto coherente de productos funcionales potencialmente superpuestos"* (Gerstein et al., 2007). Otros autores han discutido sobre la definición llegando a conclusiones similares (Portin and Wilkins, 2017, Pearson (2006)).

Estudiar la transcripción de los genes se le conoce como transcriptómica, la cual se centra en entender el conjunto completo de transcritos para una célula, tejido u órgano en particular bajo condiciones específicas para caracterizar el transcriptoma (Yadav et al., 2018, McGettigan (2013)). El transcriptoma incluye todos los tipos de RNA sintetizados,

incluyendo RNA mensajero (mRNA), RNAs largos no codificantes, *micro*RNA, RNAs pequeños (*e.i.* siRNAs, piRNAs, snoRNAs, snRNAs), poliadenilados, isoformas alternativas, RNA-editado, etc. (Milward et al., 2016, Liang (2013), Lowe et al. (2017), Ganser et al. (2019)).

Durante el desarrollo de nuevas tecnologías como microarreglos y secuenciación de RNA, la medición del transcriptoma ha evolucionado y mejorado (Lowe et al., 2017). Actualmente, con el progreso de las tecnologías masivas se puede caracterizar la expresión en diferentes condiciones (Yadav et al., 2018). Estas tecnologías tienen la capacidad de medir la expresión genética de miles de genes (Wang et al., 2009), lo cual ha requerido el desarrollo de programas computacionales que ayuden al análisis de los datos (Conesa et al., 2016). Con esto, se ha podido identificar el perfil transcriptómico en diversas condiciones y enfermedades, identificando genes que se encuentran activados o reprimidos (Wang et al., 2009, Yadav et al. (2018)).

Dichas tecnologías son los microarreglos y la secuenciación masiva de RNA. Ambas tecnologías tienen ventajas y desventajas y ambas se siguen utilizando (Wang et al., 2009, Lowe et al. (2017)). Históricamente, los microarreglos surgieron primero, pero la secuenciación de RNA es actualmente la más usada y cuenta con mayor resolución, lo que permite a los investigadores poder tener un perfil transcriptómico con más precisión (Lowe et al., 2017). Ver Apéndice: Técnicas de secuenciación y expresión génica a nivel ómico.

1.2 Enfermedades pulmonares

El sistema respiratorio tiene diversas funciones. Una de las más importantes es realizar el intercambio gaseoso de CO_2 <-> O_2 para que los eritrocitos distribuyan el O_2 en todo el cuerpo (Krogh, 1915, Hughes and Bates (2003), Jensen (2004)). Dentro de los componentes del sistema respiratorio se encuentran la nariz, la faringe, la tráquea, el tórax, etc. Sin embargo, el órgano que se encarga de hacer el recambio gaseoso es el pulmón, el cual conduce el aire a través de la inspiración y la expiración por los bronquios y bronquiolos hasta los alveolos en donde se lleva a cabo la difusión de los gases (Krogh, 1915, Boron and Boulpaep

(2016), Hughes and Bates (2003)).

El pulmón requiere elasticidad para poder extenderse y contraerse durante la inspiración y la expiración, lo que permite el proceso mecánico de la respiración (Boron and Boulpaep, 2016). Existen dos conceptos importantes que describen la elasticidad del pulmón, uno es la **distensibilidad** o **compliancia pulmonar** (*compliance* en inglés) y otro es el **retroceso elástico pulmonar** (*elastic recoil* en inglés). La **distensibilidad pulmonar** es la capacidad del pulmón y la caja torácica de expandirse, también se conoce como la medida de que tan fácil es inflar el pulmón. En contraste, el **retroceso elástico pulmonar** es la capacidad del pulmón a regresar a su forma inicial después de la inhalación, esta propiedad reacciona al cambio de presión en los pulmones y produce la disminución en el volumen pulmonar (Boron and Boulpaep, 2016; Barrett, 2013; Chaitow et al., 2002). Véase 1.1.

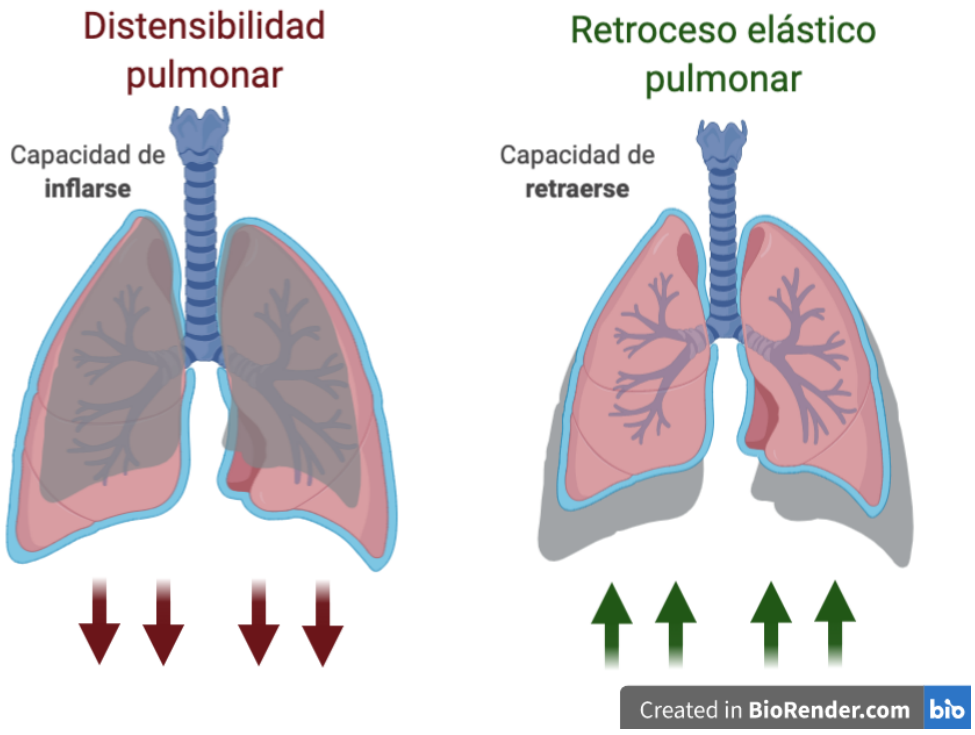


Figura 1.1: Características de la elasticidad pulmonar. La distensibilidad pulmonar y el retroceso elástico pulmonar son parte del mecanismo de ventilación pulmonar. La distensibilidad es la cualidad de poder inflar y expandir los pulmones (izquierda-rojo). Mientras que el retroceso elástico pulmonar es el poder retraerse a su tamaño original después de inhalar (derecha-verde). Figura creada usando BioRender.com y generada por Mauricio Guzmán.

Alteraciones en la elasticidad del pulmón ocasionan estados patológicos que complican la respiración. A partir de estas características elásticas, las enfermedades pulmonares se dividen de manera canónica en obstructivas, restrictivas o mixtas. El aumento en la **distensibilidad pulmonar** y la reducción en el **retroceso elástico pulmonar** se asocia a enfermedades obstructivas, las cuales se caracterizan por presentar limitación en el flujo aéreo durante la expiración y se distinguen por presentar defectos en las vías aéreas (Pierce et al., 2005a, Chaitow et al. (2002),Paraskeva et al. (2011)). Mientras que las enfermedades restrictivas se caracterizan por una disminución en la **distensibilidad pulmonar** y un aumento en el **retroceso elástico pulmonar**. Estás enfermedades se les asocia generalmente al parénquima pulmonar (Scano et al., 2010, Paraskeva et al. (2011),Hammer et al. (2015)). Véase 1.2.

1.2.1 Espirometría

La espirometría es un estudio que se realiza para medir la capacidad pulmonar y la velocidad con la que se vacían y llenan los pulmones (Petty, 2006). El estudio no produce dolor y se utiliza un aparato llamado espirómetro, puede hacerse en presencia de broncodilatadores o no (Pierce et al., 2005a). Existen términos importantes para describir la capacidad pulmonar, por ejemplo, la cantidad de aire que hay en los pulmones después de una inspiración máxima es la capacidad pulmonar total (*Total lung capacity: TLC*), el volumen residual (*Recidual volume: RV*), la capacidad vital forzada (*Forced vital capacity: FVC*), el volumen espiratorio forzado en el primer segundo (*Forced expiratory volume in 1 second: FEV1*), el volumen inspiratorio forzado en un segundo (*Forced inspiratory volume in 1 second: FIV1*), entre otros términos (Pierce et al., 2005a, Paraskeva et al. (2011)). Algunas recomendaciones y estandarizaciones importantes para tomar estas mediciones se encuentran en (Graham et al., 2019), donde se expone la importancia de aumentar la precisión, exactitud y calidad de las mismas utilizando buenas prácticas.

Los resultados de la espirometría ayudan a determinar patologías pulmonares. Para las enfermedades **obstructivas** existe una reducción en la relación *FEV1/FVC*, mientras que las **restrictivas** pueden incrementar o mantenerse igual (Pierce et al., 2005b, Paraskeva et al. (2011)) como se observa en la Figura 1.3.

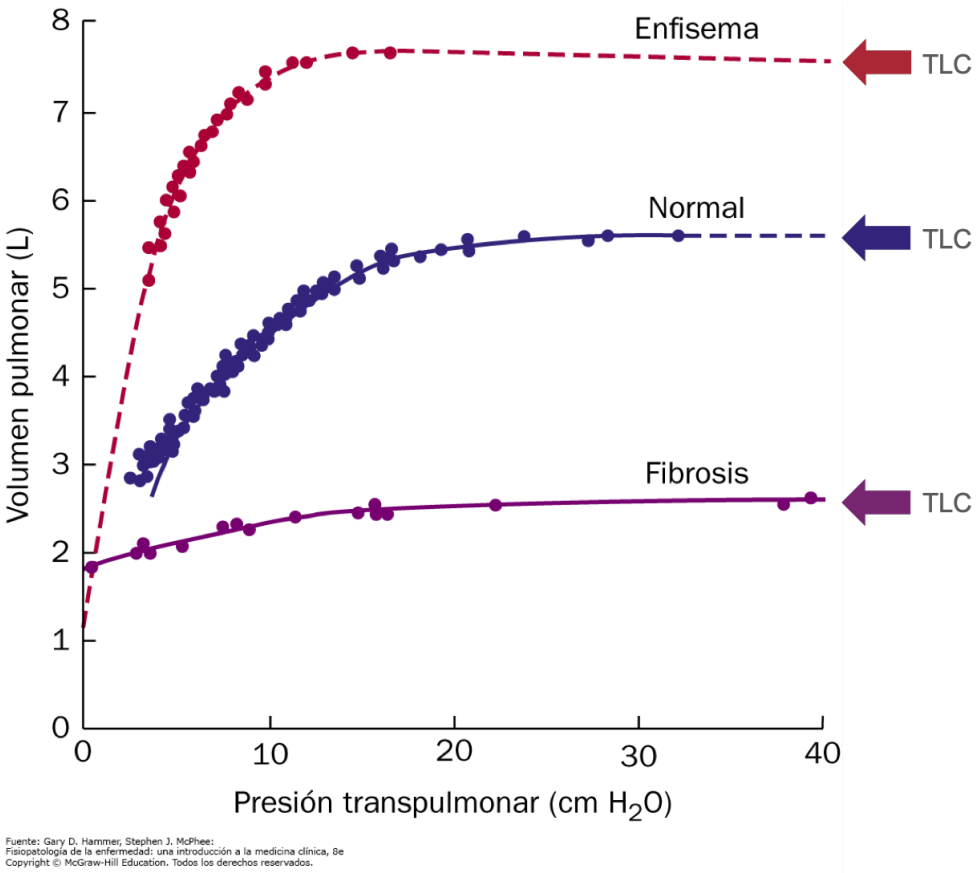


Figura 1.2: Gráfica de presión transpulmonar (cm H_2O) contra volumen pulmonar (L). Esta gráfica representa la distensibilidad pulmonar, la cual es inversa a la elasticidad (Pierce et al., 2005b). La presión transpulmonar está dada por la diferencia entre la presión intrapulmonar y la presión en la pleura, representando la fuerza ejercida para la dilatación pulmonar (Grieco et al., 2017). La gráfica presenta las curvas representativas de pacientes con fibrosis, enfisema y la curva normal en diferentes colores. El punto máximo indica la capacidad total del pulmón (TLC), la cual varia dependiendo de la patología. Figura modificada de (Hammer et al., 2015).

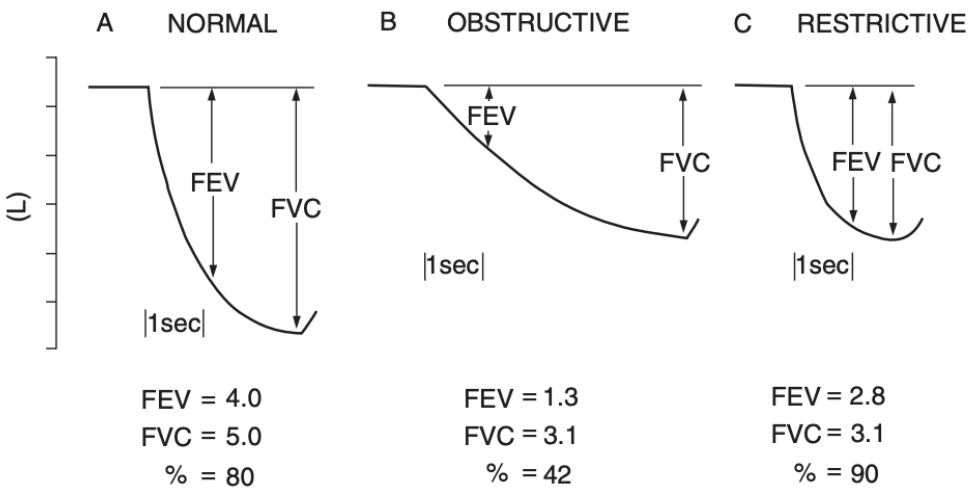


Figura 1.3: Figura con valores de *FEV1* y *FVC* de enfermedades obstructivas (enfisema) y restrictivas (fibrosis). El eje x representa el tiempo al momento de la expiración, en el eje y se gráfica el volumen en litros (L). En la gráfica se muestra el tiempo transcurrido en un segundo (|1sec|) y las medidas del volumen espiratorio forzado en un segundo (FEV) y la capacidad vital forzada (FVC). Se puede observar el cambio en las gráficas con respecto a los valores normales (A) en las enfermedades obstructivas (B), en las restrictivas (C). Tomada de (Pierce et al., 2005b).

La espirometría ayuda a caracterizar la capacidad pulmonar de forma indolora y relativamente sencilla, además de que la relación *FEV1/FVC* permite tomar decisiones acerca del tipo de enfermedad (1.3 y 1.4) (Paraskeva et al., 2011). Sin embargo, es importante considerar el historial clínico así como otros estudios para diagnosticar alguna enfermedad pulmonar. Por ejemplo, en pacientes con enfermedades obstructivas se recomienda hacer la espirometría con y sin la inhalación de un broncodilatador (Graham et al., 2019). Por otro lado, las enfermedades restrictivas no pueden diagnosticarse sólo por espirometría ya que en algunas ocasiones la relación *FEV1/FVC* no cambia (Paraskeva et al., 2011).

1.2.2 Fibrosis pulmonar idiopática y la enfermedad pulmonar obstructiva crónica

Un ejemplo característico de las enfermedades pulmonares obstructivas es la enfermedad pulmonar obstructiva crónica o EPOC por sus siglas (en inglés *Chronic Obstroctive Pulmonary Disease, COPD*). Como su nombre lo sugiere y como se describió anteriormente, EPOC al ser una enfermedad obstructiva presenta disminución en el retroceso elástico

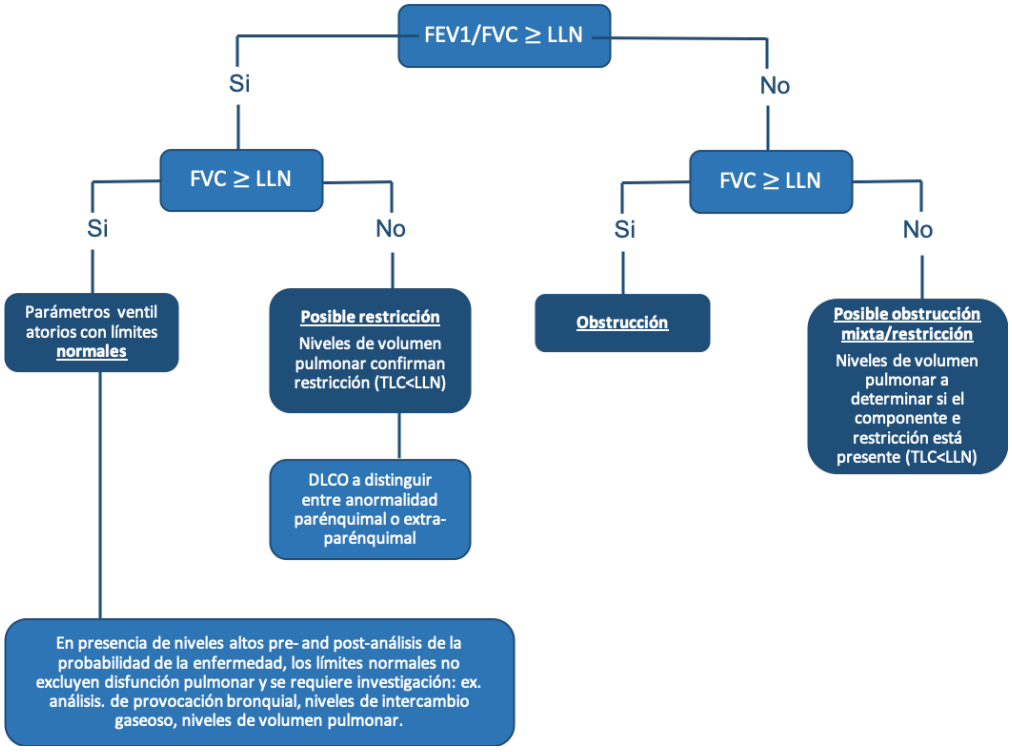


Figura 1.4: Diagrama de toma de decisiones basado en la relación *FEV1/FVC*. Figura modificada de (Paraskeva et al., 2011). Volumen Espiratorio forzado el primer segundo (FEV1). Capacidad Vital Forzada (FVC). Capacidad de difusión de transferencia del monóxido de carbono (DLCO). Capacidad pulmonar total (TLC). Límite inferior del rango normal (LLN) en una población específica.

pulmonar e hiperinflación dada por aumento en la distensibilidad pulmonar (Scano et al., 2010).

En contraste, las enfermedades pulmonares intersticiales difusas (EPID) son un ejemplo de enfermedades restrictivas, caracterizadas por un menor volumen pulmonar debido a un aumento en el retroceso elástico pulmonar (Scano et al., 2010). Las EPID, son patologías que afectan el intersticio pulmonar aumentando su grosor, generando inflamación y provocando fibrosis. Una enfermedad en particular que pertenece a las EPID es la fibrosis pulmonar idiopática (FPI) (Martinez et al., 2017).

EPOC y FPI son dos enfermedades que se encuentran en lados opuestos del espectro, tienen orígenes anatómicos distintos y fenotipos patológicos antagónicos. Sin embargo, ambas enfermedades son crónicas, progresivas, irreversibles, inflamatorias, presentan envejecimiento acelerado y están asociadas al humo de tabaco (Selman et al., 2019).

La prevalencia de ambas enfermedades está correlacionada con la edad, a mayor edad aumentan los casos de EPOC y FPI (Ito and Barnes, 2009, Mora et al. (2017)). El envejecimiento es un proceso biológico intrínseco que genera un deterioro progresivo en los tejidos, los cuales presentan mayor vulnerabilidad ante los cambios ambientales (Ito and Barnes, 2009). Durante el envejecimiento, la acumulación de daños, exposición acumulada de diferentes factores ambientales (como el humo de tabaco), alta concentración de especies reactivas de oxígeno, acortamiento aberrante de los telómeros, cambios epigenéticos y predisposición genética pueden afectar la reparación del tejido resultando en enfermedades crónicas como EPOC y FPI (Meiners et al., 2015; Selman et al., 2019, Ito and Barnes (2009), Mora et al. (2017)).

Por lo que en este trabajo decidimos enfocarnos a estudiar EPOC y FPI a nivel transcriptómico con el objetivo de poder estudiar los mecanismos similares y las diferencias moleculares que tienen estas dos enfermedades canónicamente antagonistas, pero que comparten grandes similitudes entre sí.

1.2.3 La Enfermedad Pulmonar Obstructiva Crónica

1.2.3.1 ¿Qué es la enfermedad pulmonar obstructiva crónica?

La enfermedad pulmonar obstructiva crónica (EPOC) es una enfermedad progresiva irreversible e incapacitante que provoca insuficiencia respiratoria por destrucción de la arquitectura pulmonar (Rennard and Drummond, 2015). Su definición a lo largo de la historia ha cambiado (Petty, 2006), pero actualmente la comunidad de *Global Initiative for Chronic Obstructive Lung Disease (GOLD)* ha definido la EPOC en el 2020 como:

"La Enfermedad Pulmonar Obstructiva Crónica (EPOC) es una enfermedad común, prevenible y tratable que se caracteriza por síntomas respiratorios persistentes y limitación del flujo de aire debido a anomalías de las vías respiratorias y / o alveolares generalmente causadas por una exposición significativa a partículas o gases nocivos."

– Traducido de (2020 Global Initiative for Chronic Obstructive Lung Disease, 2020)

Se caracteriza principalmente por presentar obstrucción bronquial y enfisema pulmonar (Rennard and Drummond, 2015). La obstrucción bronquial ocurre por la inflamación de los bronquios y la sobreproducción e hipersecreción de moco por las células caliciformes (Ramos et al., 2014). En el enfisema las células endoteliales y epiteliales presentan apoptosis (Horowitz et al., 2009), los fibroblastos pierden la capacidad de reparar estos daños y la matriz extracelular se ve afectada (Togo et al., 2008), lo que se refleja en la destrucción en la pared de los alveolos pulmonares (Rennard and Drummond, 2015). En la bronquitis crónica los bronquios se inflaman, existe secreción de moco y reduce la cantidad de aire que puede pasar por las vías aéreas (Barnes et al., 2015) (Figura 1.5).

1.2.3.2 ¿Cómo se diagnostica?

Existen controversias en debate sobre la forma más adecuada para diagnosticar EPOC, en la cual diversos grupos de investigación participan sin poder llegar a un acuerdo global, pero coincidiendo que la clasificación por espirometría requiere ser evaluada en contexto clínico y que existe gran cantidad de diagnósticos incorrectos cuando no se evalúa de manera adecuada

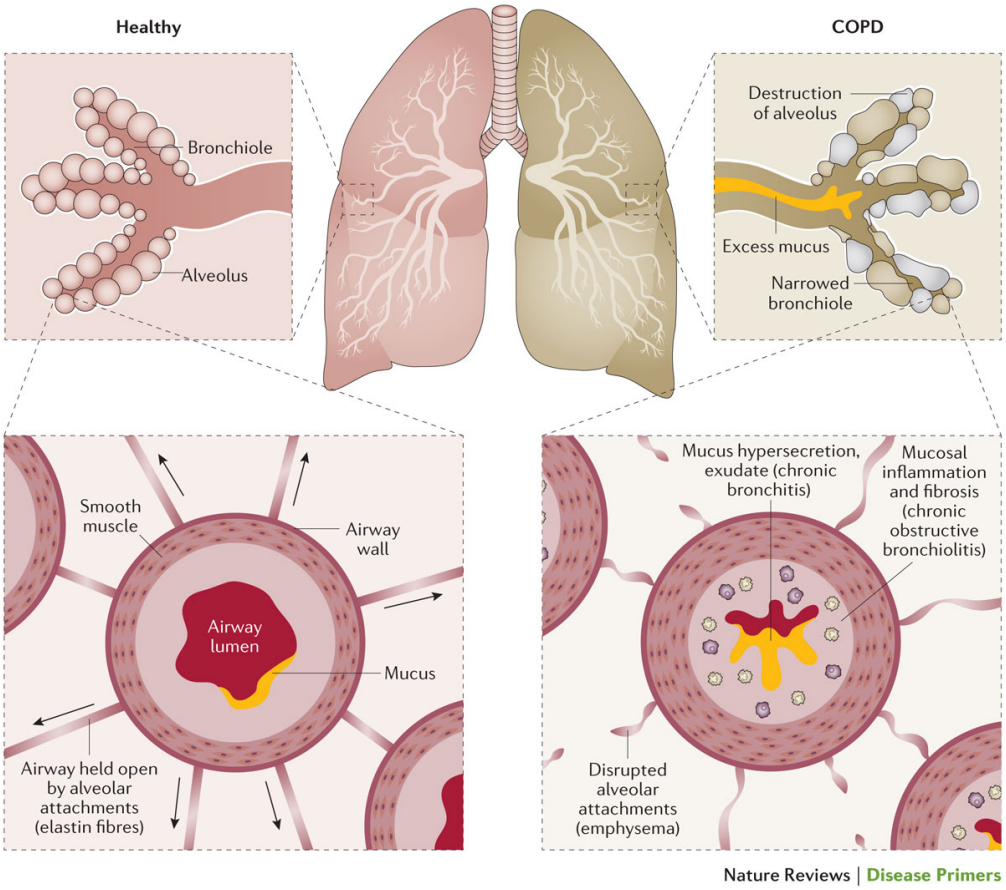


Figura 1.5: Patologías de la EPOC (Tomada de (Barnes et al., 2015)). EPOC se caracteriza principalmente por tener enfisema (imagen superior), donde la estructura alveolar se rompe. Asimismo, presenta bronquitis crónica, donde los bronquios se hiper inflaman (imagen inferior), existe secreción de moco, se cierran las vías aéreas, lo que ocasiona de manera conjunta un aumento en el retroceso elástico pulmonar.

(MacNee, 2019).

El diagnostico de la EPOC, según la guía *Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Lung Disease* (2020 Global Initiative for Chronic Obstructive Lung Disease, 2020) se hace considerando los factores de riesgo (*e.i.* disnea, tos crónica, esputo, infecciones recurrentes, humo de tabaco, humo de leña, historial familiar de EPOC, etc.), espirometría con y sin broncodilatador y con un valor $FEV1/FVC < 0.70$ (2020 Global Initiative for Chronic Obstructive Lung Disease, 2020).

La inflamación de los pulmones a causa de la inhalación de humo de tabaco o partículas como humo de leña es clave en el desarrollo de la EPOC (Orozco-Levi et al., 2006). Los pacientes muestran una respuesta de inflamación crónica por estos irritantes y la inflamación puede persistir aún después de no tener contacto (*e.i.* dejar de fumar) (Rutgers et al., 2000). Muchos de estos mecanismos son aún desconocidos y siguen en investigación, sin embargo se han encontrado auto-anticuerpos (Wen et al., 2018; Feghali-Bostwick et al., 2008) y alteraciones en la microbiota de los pulmones (2020 Global Initiative for Chronic Obstructive Lung Disease, 2020).

A pesar de desconocer con precisión la patogénesis de la EPOC, se sabe que el estrés oxidativo amplifica la inflamación, el cual se encuentra incrementado en los pacientes con exacerbaciones (Kirkham and Barnes, 2013). Existe también un desbalance en la relación proteasas-antiproteasas que ocasiona la destrucción de elastina debido al incremento de proteasas, dando lugar a enfisema pulmonar (Barnes et al., 2015). El aumento de células del sistema inmune en los pacientes con la EPOC está caracterizado por aumento de macrófagos, neutrófilos, linfocitos T y en algunos casos eosinófilos (Rutgers et al., 2000). Después de la inflamación, se ha observado fibrosis en las vías aéreas y en el intersticio contribuyendo a la limitación del flujo aéreo (Barnes et al., 2015, 2020 Global Initiative for Chronic Obstructive Lung Disease (2020)).

1.2.3.3 Factores de riesgo

El principal factor de riesgo es el humo de tabaco, sin embargo, no es el único factor ambiental (Churg et al., 2008, Halbert et al. (2006), Wright and Churg (2002)). Se sabe

que menos del 50% de los fumadores desarrollan esta enfermedad (Lundbäck et al., 2003), así que otros factores también están participando en la EPOC (Halbert et al., 2006). Por ejemplo, la exposición a humo de leña, humo de pipas, cigarros eléctricos, mariguana, humo de combustión, polvos, vapores, gases y otros químicos contribuye al desarrollo de esta enfermedad (2020 Global Initiative for Chronic Obstructive Lung Disease, 2020)(Figura 1.6). También la edad es un factor de riesgo importante y aunque anteriormente se pensó que los hombres tenían más riesgo, actualmente algunos estudios han demostrado mayor susceptibilidad en las mujeres (SILVERMAN et al., 2000, Lundbäck et al. (2003)).

El factor genético más estudiado es la deficiencia de la Alfa-1 Antitripsina (AAT), una proteína inhibidora de proteasas que evita que la elastasa degrade a la elastina y, a su vez, causa la destrucción pulmonar, pero solo representa el 1-2% de los casos (DeMeo and Silverman, 2004). En estudios de asociación genética se ha visto varios polimorfismos implicados tanto en la EPOC, como en la disminución de los valores de *FEV1*, estas mutaciones se han encontrado cercanas a genes como *HHIP, CHRNA5, FAM13A, HTR4, RIN3, MMP12, SFTPD*, entre muchos otros (Agustí and Hogg, 2019; Silverman, 2020).

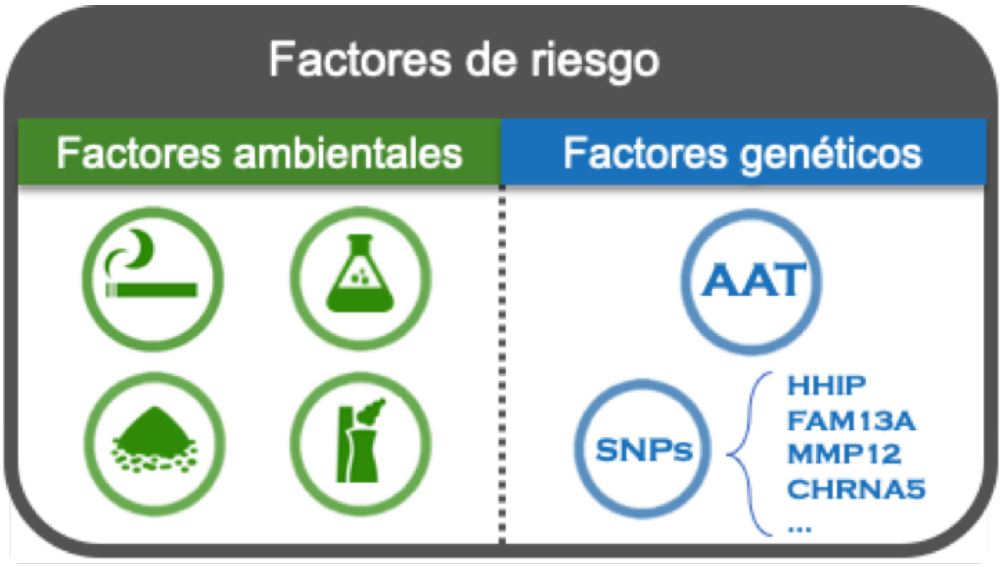


Figura 1.6: Imagen con los factores de riesgo asociados a EPOC. En la izquierda se encuentran algunos factores ambientales (representando el humo de tabaco, contaminación, polvos y químicos) y en la derecha los factores genéticos. AAT es alfa-1 antitripsina, SNPs es *Single Nucleotide Polymorfism*.

1.2.4 Fibrosis Pulmonar Idiopática

1.2.4.1 ¿Qué es la fibrosis pulmonar idiopática?

La fibrosis pulmonar idiopática (FPI) pertenece a las EPID, un grupo de enfermedades que tienen cambios histopatológicos en el intersticio pulmonar caracterizándose por engrosamiento en la pared pulmonar (Figura 1.7 (Selman and Pardo, 2014). La FPI es una enfermedad crónica que afecta a los pulmones, está asociada al envejecimiento y es considerada de etiología desconocida (Selman and Pardo, 2014). Sin embargo, se sabe que las células epiteliales alveolares (principalmente las células tipo II) producen mediadores profibróticos estimulando el crecimiento de fibroblastos y su diferenciación generando la sobre producción de matriz extra-celular (ECM, por su siglas en inglés) (Selman and Pardo, 2014; Selman et al., 2019).

Algunos de estos factores secretados que contribuyen al remodelación de la matriz extracelular (MEC) son *TFG-beta*, *TNF*, algunas metaloproteinasas como *MMP1*, *MMP7*, *MMP19*, o *CXCL12*, entre otros los cuales generan una expansión en la población de fibroblastos y promueven la diferenciación a miofibroblastos (Martinez et al., 2017; Pardo et al., 2016). Algunos de los procesos moleculares que caracterizan a la FPI que es una enfermedad ligada al envejecimiento son acortamiento de los telómeros, inestabilidad genómica, senescencia celular y por otro lado la pérdida del balance proteasas-antiproteasas con acumulación de proteínas aberrantes (Selman and Pardo, 2014; Meiners et al., 2015).

1.2.4.2 ¿Cómo se diagnostica?

El diagnóstico de FPI se lleva a cabo a través de tomografía, en donde pacientes con FPI tienen una lesión tipo *honeycombing* o panal de abeja, se deben descartar otras EPIDs y se toma una biopsia para analizarse con histopatología (Raghu et al., 2018; Martinez et al., 2017). Recomendaciones para el correcto diagnóstico de FPI se encuentra en (Raghu et al., 2018).

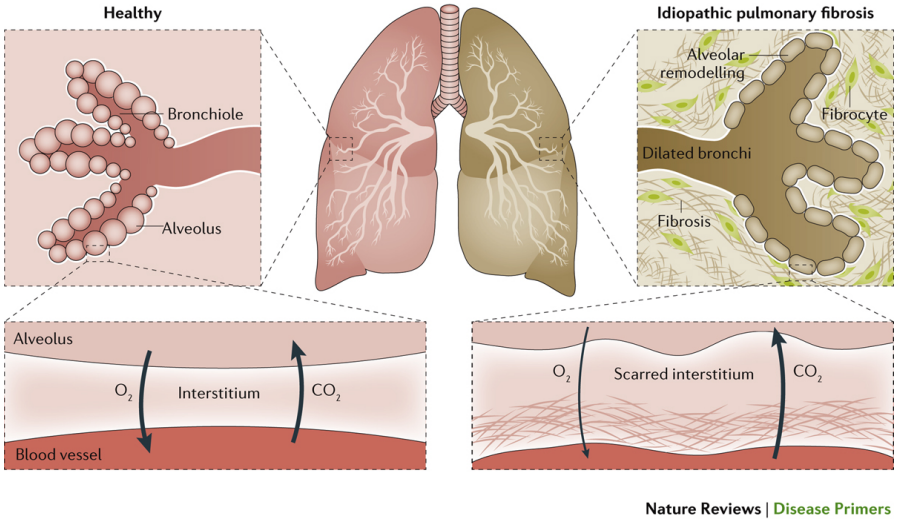


Figura 1.7: Imagen representando la FPI (Tomada de (Martinez et al., 2017)). La FPI es una enfermedad que se caracteriza por dilatación en los bronquios, remodelación de la matriz extra-celular y presencia de fibrosis en el intersticio pulmonar, lo que provoca un aumento en el retroceso elástico pulmonar y mayor dificultad para el recambio de O_2.

1.2.4.3 Factores de riesgo

Varios factores de riesgo se han asociado a FPI, como el humo de cigarro, infecciones virales (*e.i.* virus de *Epstein-Barr*) y bacterianas, exposición ocupacional como a polvo de madera, metal, sílice, etc. y algunos componentes genéticos (Meiners et al., 2015). Se calcula que hasta un tercio del riesgo de la enfermedad podría podría ser explicado por variantes genéticas. Se han podido detectar variantes raras con efectos grandes como mutaciones en *TERT, TERC, SFTPC*, entre otros y mutaciones comunes con efectos pequeños pero asociadas a la FPI como mutaciones cercanas a *MUC5B, FAM13A, MAPT*, por mencionar algunos (Mathai et al., 2016).

La FPI está altamente asociada al envejecimiento y principalmente al acortamiento de los telómeros y senescencia celular. Sin embargo, EPOC es la enfermedad característica de un envejecimiento acelerado en el pulmón, por lo que la relación y los mecanismos puntuales que definen porque un paciente en edad avanzada desarrolla EPOC o FPI permanecen en discusión (Selman and Pardo, 2014; Meiners et al., 2015; Selman et al., 2019).

1.3 Antecedentes

Tecnologías como *northern blot*, reación en cadena de la polimerasa (*PCR*) o reacción en cadena de la polimerasa con transcriptasa inversa (*RT-PCR*) permitieron la cuantificación de los transcritos particulares en enfermedades y diversas condiciones biológicas (Heid et al., 1996, Lowe et al. (2017)). Por lo que el desarrollo de tecnologías como microarreglos y secuenciación masiva para cuantificar múltiples transcritos simultáneamente, atrajo la atención de diversas áreas incluida la investigación biomédica (Lowe et al., 2017). Estás tecnologías se han aplicado desde la investigación básica para caracterizar condiciones biológicas hasta el diagnostico de enfermedades (Gordon et al., 2002, Costa et al. (2013)). Esto ha permitido generar nuevo conocimiento sobre el desarrollo de las enfermedades, biomarcadores, y blancos terapéuticos (Kori and Yalcin Arga, 2018).

1.3.1 Estudios transcriptómicos de EPOC

EPOC se ha estudiado desde la perspectiva de la transcriptómica desde que se utilizaron los microarreglos para analizar la expresión de genes. En el 2003, se publicaron los primeros análisis de datos masivos en EPOC datos reportados por Eric P Hoffman (GSE475). Los autores no tienen publicación asociada a este experimento. Posteriormente se desarrollaron 3 nuevos estudios transcriptómicos en pulmón/sangre de pacientes con EPOC (Spira et al., 2004a, Golpon et al. (2004), Spira et al. (2004b)). El aumento en experimentos se encuentra asociado al desarrollo biotecnológico de técnicas de secuenciación masiva (Ver Capítulo **??**).

Se ha descrito que la EPOC no solo es una enfermedad compleja con diversos fenotipos clínicos, sino también heterogénea indicando que dichas características pueden o no estar presentes en los pacientes en un momento particular y que varían en proporción entre individuos (Agusti, 2014); Esto ha sido corroborado en los perfiles transcripcionales de la EPOC, que han mostrado heterogeneidad en las muestras (Ham et al., 2019, Wedzicha (2000)).

En un caso particular se mostró como la composición celular afecta el perfil transcripcional en las muestras con EPOC. Se usaron muestras de RNA-seq de pacientes con EPOC

(experimento de (Kim et al., 2015)) y se mostró que contienen diferentes composiciones celulares ocasionando heterogeneidad en la expresión de genes. Por lo cual, los autores proponen incluir como covariable al porcentaje de tipos celulares que se obtiene por deconvolución al análisis de expresión diferencial (Ham et al., 2019).

Adicionalmente, se sabe que el momento en el que se toma la muestra tiene gran relevancia, es decir, si el paciente presenta un cuadro exacerbado durante la toma de muestra o si se encuentra estable (Wedzicha, 2000).

La importancia de conocer una firma robusta para la EPOC está en los diversos comportamientos que presentan los pacientes ante los tratamientos (MacNee, 2019; Dransfield et al., 2019). En este trabajo se usarán experimentos previamente publicados en *Gene Expression Omnibus (GEO)* (Figura 1.8). Para mayor información sobre GEO por favor véase el capítulo **??**.

1.3.2 Estudios transcriptómicos de FPI

Existen menos experimentos realizados de FPI que se encuentran registrados en GEO a comparación de EPOC. Esto se puede observar en la Figura 1.9, teniendo 29 experimentos a la fecha (Enero 15, 2021).

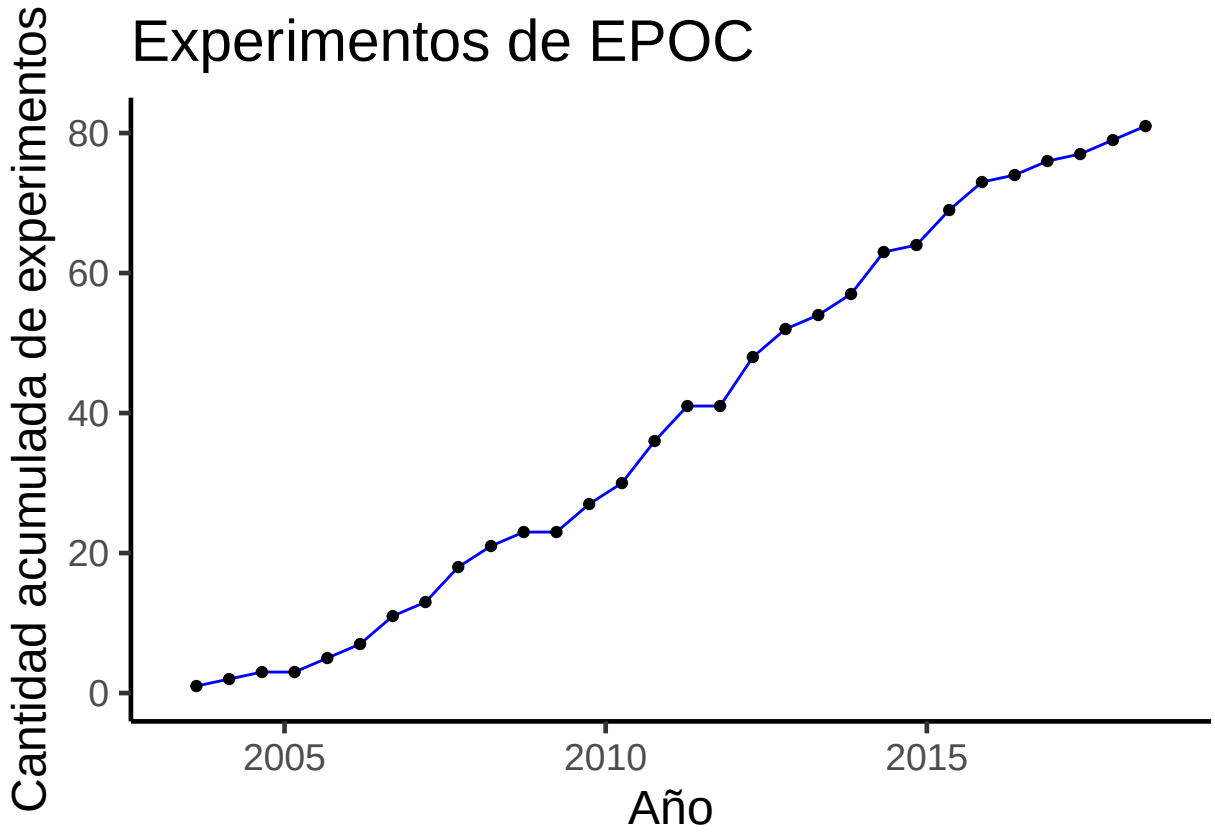


Figura 1.8: Figura de experimentos públicos asociados a EPOC. En el eje de las x se encuentran los años hasta la fecha. En el eje y la cantidad de experimentos medido en número de identificadores GSEs que estaban disponibles para ese año.

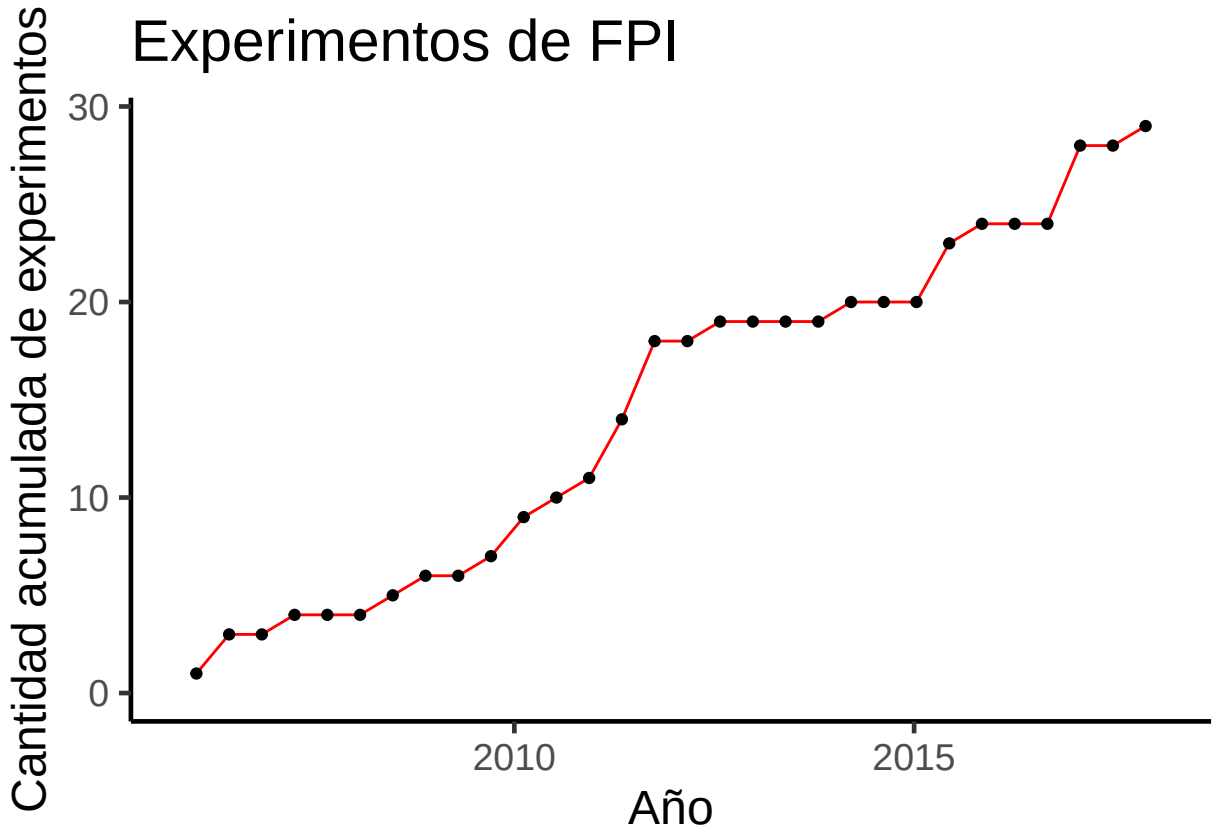


Figura 1.9: Figura de experimentos públicos asociados a FPI En el eje de las x se encuentran los años hasta Enero 15, 2021. En el eje y la cantidad de experimentos medido en número identficadores de GSEs que estaban disponibles para ese año.

Capítulo 2

Objetivos

2.1 Hipótesis

La EPOC y la FPI son enfermedades pulmonares con similitudes y diferencias etiológicas, biológicas y clínicas, el uso de un compendio con datos transcriptómicos públicos de la EPOC y FPI permitirá identificar los genes involucrados que no se han identificado.

2.2 Objetivo General

Realizar un compendio con datos públicos de expresión genética para determinar las diferencias transcripcionales entre individuos sanos y pacientes con EPOC y FPI.

2.3 Objetivos específicos

2.3.1 Objetivo específico 1

Creación del compendio de expresión genética con datos de EPOC y FPI, a través de la i) búsqueda, ii) selección de los experimentos, iii)descarga de los datos, iv)curación manual y la iv) homogeneización de los datos.

2.3.2 Objetivo específico 2

Analizar los datos para encontrar firmas transcripcionales en pacientes con EPOC y FPI, caracterizando los genes diferencialmente expresados y las vías enriquecidas.

Capítulo 3

Base de datos: PulmonDB

3.1 Artículo: COMMAND

Desde que se empezaron a usar los microarreglos y la secuenciación masiva para entender la expresión de los genes en las diferentes enfermedades, bases de datos públicas como lo es GEO y ArrayExpress, han acumulado experimentos de diversos laboratorios. Tal es el caso de EPOC, donde desde el 2005 encontramos experimentos que analizan la expresión diferencial en esta enfermedad.

Sin embargo, estos experimentos se realizaron utilizando diversas plataformas en distintos laboratorios y procesados de distinta manera. Por lo cual, integrar esta información requiere utilizar los datos crudos y pre procesarlos de manera uniforme. Y adicionalmente, se requiere integrar los datos para que puedan ser comparables entre sí.

Existen varias formas de integrar diferentes experimentos, la forma canónica para hacerlo sería a través de un meta-análisis, otra alternativa es integrar los datos corrigiendo por *batch effect* cuando los experimentos han sido realizados utilizando la misma tecnología (*e.i. Affymetrix, Agilent*).

En el 2010, Kristoff y colaboradores integraron los experimentos transcriptómicos publicados para *E. coli* que se encontraban hasta el momento para generar un compendio con estos datos y poder analizarlos de una manera integral. Este compendio se llamó COLOMBOS

(Engelen et al., 2011), el cual se construyó utilizando una herramienta llamada COMMAND. Después de unos años el compendio se expandió a más bacterias permitiendo integrar datos transcriptómicos de diferentes organismos (Moretto et al., 2016a, Meysman et al. (2014)).

Más tarde en el 2015 se publicó un nuevo compendio ahora enfocado a estudiar la vid (*Vitis spp.*), llamado VESPUCCI (Moretto et al., 2016b). Los resultados de este compendio han permitido encontrar factores transcripcionales que se co-expresan y actúan en conjunto ayudando a expandir el conocimiento de la regulación en este organismo (Pilati et al., 2017, Malacarne et al. (2018)).

Como parte del aprendizaje para ocupar y aprender a manejar COLOMBOS, se realizó un manual de usuario, el cual fue escrito por el grupo de trabajo para manejar dicha herramienta. COLOMBOS te permite 1) descargar los datos desde repositorios públicos, 2) guardarlos en una base de datos MySQL, 3) re anotar las sondas de microarreglos en una versión nueva del genoma (en nuestro caso ocupamos el genoma 37 de humano), 4) Curar la información de cada muestra (*e.i.* tipo de muestra, enfermedad, sexo, edad) 5) Crear *sample contrasts*, de los cuales hablaremos más adelante, y 6) normalizar las muestras de manera homogénea.

Dichos "sample contrasts" son contrastes que se hacen entre una muestra que se toma como referencia y otra muestra, la cual puede ser del grupo control o de otro grupo (e.i. COPD, IPF), con el afán de emular los contrastes que se tenían en los microarreglos de dos canales. Esta idea se generó para que los contrastes puedan ser comparables entre sí, ya que se usa una muestra control como referencia para obtener valores relativos. Los contrastes se calculan como la resta de los valores de expresión, que se encuentran en logaritmo base 2.

$$sample\ contrast = log_2(test) - log_2(ref)$$

COMMAND fue actualizado y se publicó la nueva versión de la aplicación (Moretto et al., 2019). Está actualización no realiza todos los pasos que se describieron anteriormente pero se enfoca en la creación de la base de datos. Está actualización permite 1) descargar los datos desde repositorios públicos, 2) guardarlos en una base de datos PostgreSQL y 3) re anotar las sondas. Durante mi doctorado participé con los autores para probar la aplicación,

reportar errores y hacer pruebas para la publicación, así como apoyo en la redacción del artículo el cual se encuentra adjunto a continuación. Está publicación permitió conocer la plataforma de COMMAND. La cual se usó para generar un compendio de expresión genética con experimentos de la EPOC y FPI que nos permitiera estudiar las diferencias y similitud entre enfermedades.

3.1.1 Contribución personal

Para la publicación de este artículo participé en evaluar, probar, y revisar la aplicación de COMMAND, creación de la figura 4, participación en el desarrollo de la figura 3 y 5, así como en la revisión del manuscrito final.

3.1.2 Publicación 2: *First step toward gene expression data integration: transcriptomic data acquisition with COMMAND>_*

Moretto et al. BMC Bioinformatics (2019) 20:54
https://doi.org/10.1186/s12859-019-2643-6

BMC Bioinformatics

SOFTWARE Open Access

First step toward gene expression data integration: transcriptomic data acquisition with COMMAND>_

Marco Moretto[1*], Paolo Sonego[1], Ana B. Villaseñor-Altamirano[2,3] and Kristof Engelen[1*]

Abstract

Background: Exploring cellular responses to stimuli using extensive gene expression profiles has become a routine procedure performed on a daily basis. Raw and processed data from these studies are available on public databases but the opportunity to fully exploit such rich datasets is limited due to the large heterogeneity of data formats. In recent years, several approaches have been proposed to effectively integrate gene expression data for analysis and exploration at a broader level. Despite the different goals and approaches towards gene expression data integration, the first step is common to any proposed method: data acquisition. Although it is seemingly straightforward to extract valuable information from a set of downloaded files, things can rapidly get complicated, especially as the number of experiments grows. Transcriptomic datasets are deposited in public databases with little regard to data format and thus retrieving raw data might become a challenging task. While for RNA-seq experiments such problem is partially mitigated by the fact that raw reads are generally available on databases such as the NCBI SRA, for microarray experiments standards are not equally well established, or enforced during submission, and thus a multitude of data formats has emerged.

Results: COMMAND>_ is a specialized tool meant to simplify gene expression data acquisition. It is a flexible multi-user web-application that allows users to search and download gene expression experiments, extract only the relevant information from experiment files, re-annotate microarray platforms, and present data in a simple and coherent data model for subsequent analysis.

Conclusions: COMMAND>_ facilitates the creation of local datasets of gene expression data coming from both microarray and RNA-seq experiments and may be a more efficient tool to build integrated gene expression compendia. COMMAND>_ is free and open-source software, including publicly available tutorials and documentation.

Keywords: Transcriptomic, Gene expression, Microarray, Rna-seq, Compendia, Data integration

Background

Transcriptomic studies started over 20 years ago with the first spotted microarray [1] while the first RNA-seq experiments appeared about a decade ago [2–4]. Since then the number of transcriptomic experiments performed has constantly grown, favoured, among other things, by the increase of technical quality and the decreasing prices [5]. Nowadays large studies profiling expression of genes and their association with several experimental conditions are commonplace, and the wealth of public information is a huge help for scientific investigation. Nevertheless, most of the true potential for reuse and integration remains untapped because of the vast heterogeneity of such datasets and the difficulties in combining them. With the advent of systems biology, data integration emerged as a prevailing aspect to take full advantage of such rich sources of information [6]. Several approaches have been proposed to fulfill the need to effectively integrate gene expression data and they can generally be categorized as being either direct integration or meta-analysis. The former directly consider the sample-level measurements within each study, and merge these into a single data set [7]. Meta-analysis, on

* Correspondence: marco.moretto@fmach.it; engelen.kristof@gmail.com
[1]Unit of Computational Biology, Research and Innovation Centre, Fondazione Edmund Mach, via E. Mach 1, 38010 San Michele all'Adige, Italy
Full list of author information is available at the end of the article

the other hand, integrates gene expression analysis combining information from primary statistics (such as *p*-values) or secondary statistics (such as lists of differentially expressed genes) resulting from single studies. Those studies combine the information from several data sources defining confidence levels subjectively for each individual study without a general scheme. Meta-analysis is a common method to integrate conclusions from different studies [8].

Both approaches have been widely adopted and many tools have been developed to exploit or further analyse such datasets [9–13]. Regardless of the strategy used to combine and analyse a large amount of gene expression experiments, the first step in common with all these approaches is the acquisition of raw data. COMMAND>_ (COMpendia MANagement Desktop) is a web application developed in order to facilitate the creation and maintenance of local collection of gene expression data and have been successfully used to build gene expression compendia such as COLOMBOS [14] and VESPUCCI [15]. It has been designed with flexibility in mind in order to deal with the disparate ways in which gene expression data are published, and to be easily extended to deal with new technologies.

Implementation

COMMAND>_is a multi-user web application developed in Python 3 using the Django 1.11 framework for the backend; the web interface has been developed using ExtJS 6.2 with a look and feel typical of desktop applications (Fig. 1). Despite being developed as a single page application, it allows users to navigate using browser buttons. By default it relies on PostgreSQL as Database Management System (DBMS), but the Django Object Relational Mapping (ORM) allows it be used with other DBMSs as well. COMMAND>_ uses both AJAX and WebSocket (via Django Channel) for client-server communications. WebSocket ensures a two-way communication between the web interface and a Python backend, easing the problem of continuously polling the server

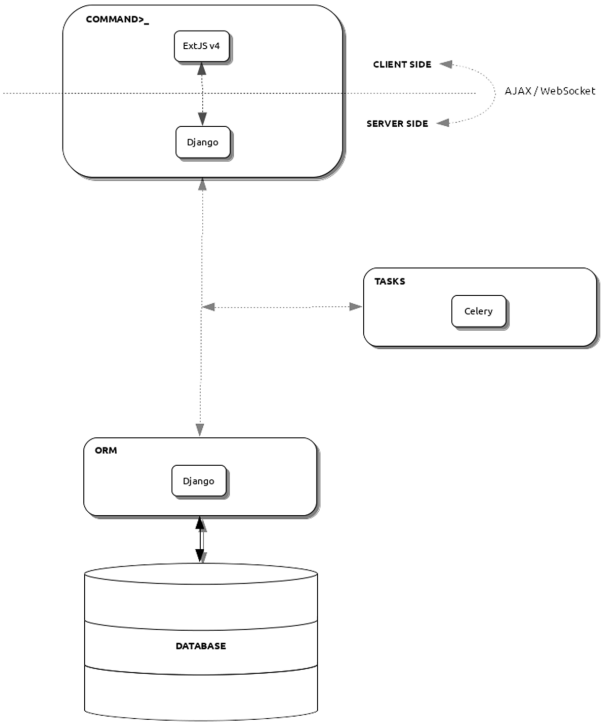


Fig. 1 COMMAND>_ infrastructure. The Graphical User Interface (GUI), on the client-side, is developed using the ExtJS Javascript framework that communicates using both AJAX calls and WebSocket to the server-side part of the application developed using the Django framework. All the business-logic has been developed in Python within the Django framework that is in charge of managing the database connection with the Object Relational Mapping (ORM) layer. Celery is the task queue distributed system used to run and manage tasks

for updates on time-consuming tasks. Intensive tasks such as downloading and parsing files are managed asynchronously by the Celery task queue system so that many processes can run simultaneously (8 by default). COMMAND>_ is a complex application with several layers that work together. To ease the deployment process we provide a Docker Compose file, thus having a working instance is just a matter of running one configuration file. Since COMMAND>_ relies on several third-party software, performance depends in part on the specific software requirements. The default Python scripts are designed to keep the memory footprint as low as possible and scale linearly with respect to the input size, because many of them might run concurrently. The complete requirements list is available at the documentation page. COMMAND>_ has been designed to be adapted to different gene expression platforms and currently handles platforms of two kinds, microarray and RNA-seq, but can be extended to allow for more platforms to be managed. Gene expression data itself are modeled as one possible type of data that can be collected. By extending specific classes, as reported in the online documentation, COMMAND>_ can be adapted to potentially handle any kind of quantitative data.

Data model

The basic concept behind the data model and how it is implemented in the database (Fig. 2) revolves around the idea that a set of measurements for several biological features (such as genes in case of gene expression data) are collected across different samples. The collected values might be direct or indirect measurements of such biological features and depends on the type of platform used in the experiment. In case of microarrays for example, each measurement refers to a single probe (a *reporter* in the data model) and thus it is an indirect measurement of gene activity. Samples can then be thought as a set of *reporter* measurements taken with a *platform* that is therefore a set of reporters. Biological features (as genes) and reporters (as probes) might have different properties (fields) such as name and sequence that can be used to couple the two entities. The three entities *experiment*, *platform*, and *sample* as well as *biological features* and *reporters* also hold meta-data, such as original ids, names and descriptions.

Results and discussion

Workflow

COMMAND>_ is a multi-user application. From the web interface it is possible to create users and groups and grant privileges. Admin users have unlimited access, while normal users might be limited to work only on specific compendia and/or with a subset of functionalities. The typical workflow can be divided into three steps: i) search and download experiment data, ii) parsing downloaded files, iii) preview and import experiment data into the local database (Fig. 3).

A mandatory prerequisite for being able to perform these steps is to first establish the genomic background

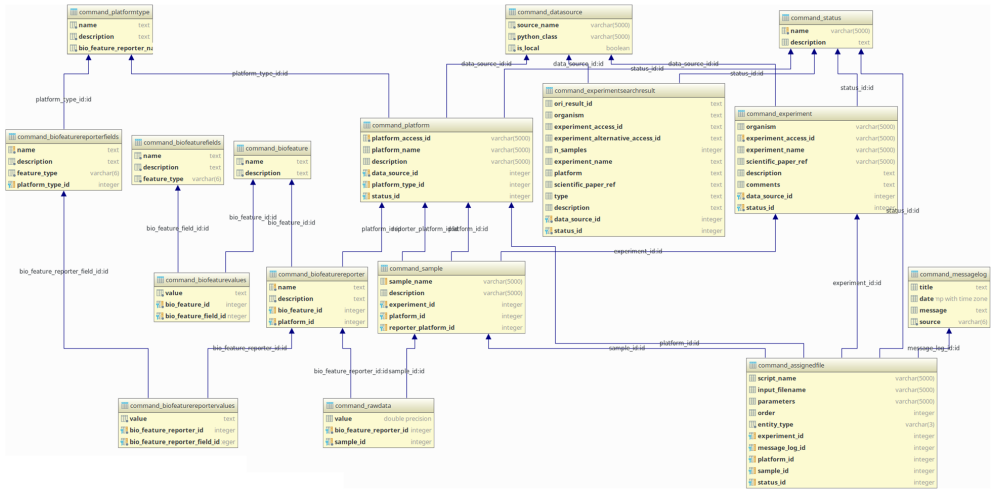


Fig. 2 The database schema that represents the data model. The core part is represented by the *sample* table that holds the information of the *experiment* it belongs to, the *platform* used to measure it as well as the link to the *raw_data* measurements

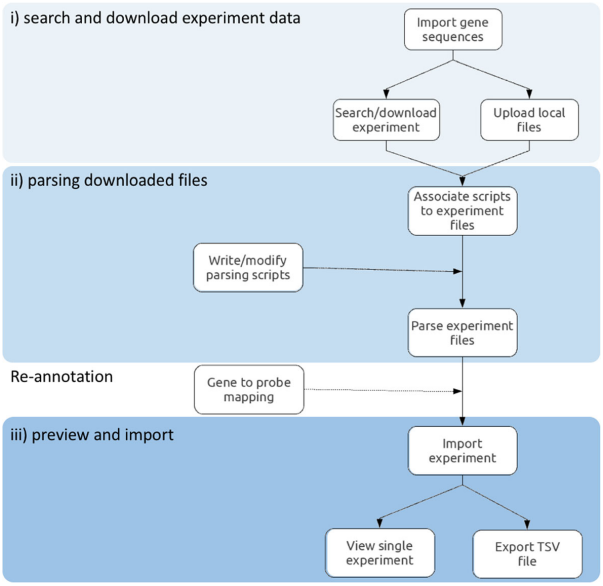


Fig. 3 The flowchart of the typical workflow. Users start by searching and downloading experiments from public databases or uploading files for local experiments. Experiment files are then associated with parsing scripts and the parsing phase runs in background. Once experiment is parsed can be imported into the database and, if necessary, probes can be mapped to genes

for the expression data, by uploading a FASTA file with gene sequences. Users can then import experiments starting by searching and downloading them from public databases or uploading local files (Fig. 4a). The supported databases are (at the moment) NCBI GEO [16], SRA [17] and EBI ArrayExpress [18]. Once the search has been performed, users can select one or more experiments and start the download process. Compressed files will be automatically extracted in a temporary folder.

The pivotal point is the assignment of downloaded files together with parsing scripts to entities (*experiment*, *platforms* and *samples*) to mine only the relevant information (Fig. 4b). The scripts can be created or modified directly within the interface and are responsible for parsing input files and populating each part of the data model, i.e. measurement data and meta-data for *experiment*, *platforms* and *samples*. Once scripts are assigned to downloaded files, they can run independently and the results can be inspected using the preview interface. If the experiment appears to be complete, it can be imported into the database. Any possible error that might occur during parsing or importing of the experiment will be reported in the system log.

When a new microarray platform gets imported, it would be necessary to map its probes to genes. The probe to gene mapping is a fundamental process carried out performing a BLAST+ [19] alignment and a two-step filtering (Fig. 4c). The alignment might take a while for platforms with a lot of probes, especially when using the short-blastn option, and the result cannot be used as-is for the probe to gene mapping. Bad alignments need to be filtered out in order to retain only the most plausible ones, i.e. the alignments that most likely represent the "true" mRNA-to-probe hybridization process. The filtering step is usually fast and can be performed several times on the same alignment result to test different threshold choices. The two-steps filtering tries to mitigate the side effect of a simpler filtering (Fig. 5) and it is performed to guarantee that probes map to genes with high similarity (restrictive alignment threshold), while also mapping unambiguously to a unique position avoiding cross-hybridization issues in the measurements (less restrictive alignment threshold). Since probes coming from different microarrays generally differ in terms of length, origin, and sequence quality, parameters and cut-off thresholds can be adjusted in order to always obtain the reasonably best possible results according to each platform's specific characteristics and user needs. The probe to gene mapping step has the advantage of enhancing data homogeneity since all microarray platforms will be annotated using the same gene list (i.e. the same genomic background represented by the FASTA file with gene sequences uploaded during the initial step).

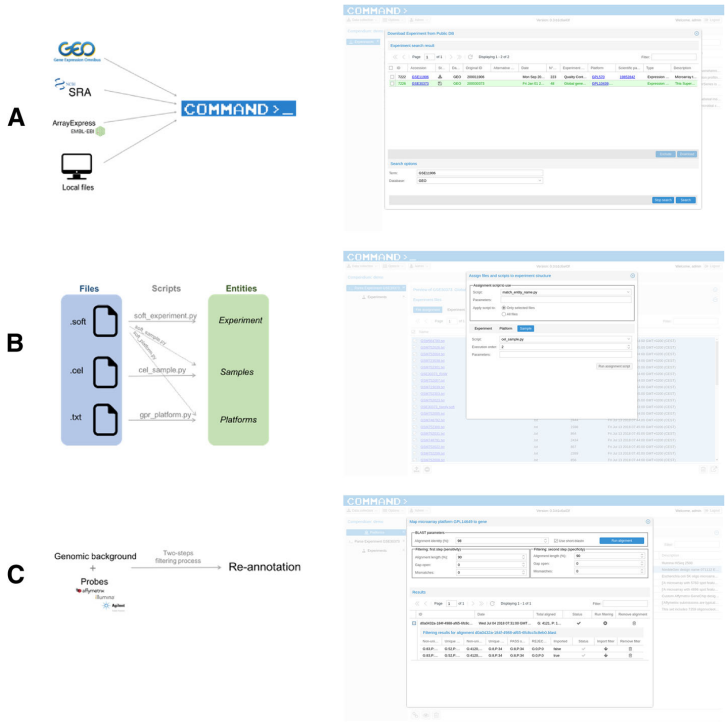


Fig. 4 a The download experiment dialog. From this dialog it is possible to search experiment from GEO, SRA or ArrayExpress and download all the related files. **b** The file assignment dialog. From this dialog users assign scripts to experiment files in order to parse the information. **c** Probe to gene mapping dialog. This dialog allows users to perform the alignment and the two-step filtering used to re-annotate microarrays by mapping probes to genes

Moreover, annotating the microarray with the latest available data is often preferable since it might improve the expression data interpretation [20, 21]. If probe sequences are not available, or relying on the default annotation is more appropriate, it is possible to manually associate probes to genes using, for example, the manufacturer annotation (gene identifiers). All the parameters and re-annotation are stored on COMMAND>_, so that the procedure is completely reproducible.

In case of RNA-seq platforms, these steps are not necessary since the imported measurements (raw counts) are directly related to genes of the defined genomic background without the need for reporters like probes as for microarray experiments. Once FASTQ files are downloaded and associated to samples, the user will need to create the index file for the genomic background to be used in the alignment program. A FASTA file with gene sequences imported by the user would be automatically created and put in the experiment directory to be used as target for the index creation script. By default, FASTQ files will be then trimmed using Trimmomatic [22] and expression level quantified using Kallisto [23]. Users that wish to use different programs, could copy them to the COMMAND>_ directory and to write a Python wrapper script to use them.

The three steps described in the workflow are specific for gene expression data, but would be the same in case of other kind of quantitative data. For example, exon or small-RNA sequencing could easily be used by adopting a different genomic background, thus uploading a FASTA file with exons or small-RNAs sequences respectively. The importance of the genomic background definition lies in the fact that it establishes exactly what is measured by the imported experiments. Considering that the genomic background should not change during data collection, not all quantitative data are equally suitable for being imported in COMMAND>_. For example, metagenomic experiments for which Operational Taxonomic Units (OTUs) change from sample to sample (and even more from experiment to experiment) would not be an ideal type of quantitative data to be collected.

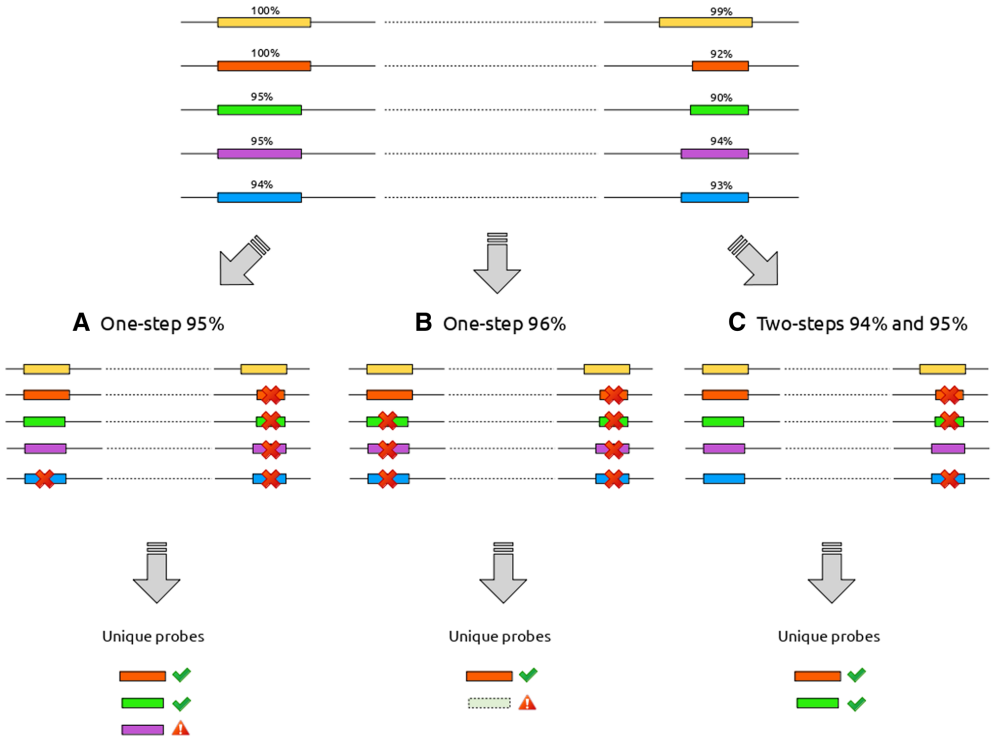


Fig. 5 Advantages of two-step filtering. The five probe sequences (yellow, orange, green, purple and blue) all aligns twice with different scores. Starting from the assumptions that, for this specific example case, a reasonable threshold is 95% or more and that for a sequence to be considered aligned uniquely the score difference between the alignments of the same probe should be more than 3%, we consider three different scenarios. The leftmost one (scenario **a**) presents the case of a single-step filtering with a threshold equal to 95%. In this case the probe selected as unique ones are the orange, green and purple one. While the first two probes are compliant with our assumptions, the purple one should be discarded since it aligns twice with a score of 95 and 94% respectively. To avoid this situation we might try do adjust the threshold raising it to 96% as depicted in the scenario **b**. The problem in this case is given by the fact that we would miss the green probe. Using a two-step filtering, as shown in scenario **c**, avoids both this unwanted situations since the first filter (at 94%) would retain both the purple probe alignments (that will be discarded since the alignment is not unique) and the blue probe that will then be removed by the second filter at 95% leaving only the orange and the green probes as expected

Comparison with similar tools

To the best of our knowledge there are no other tools that offer all the options COMMAND>_ does. Nevertheless, we will report the main differences between COMMAND>_ and other similar tools (see Table 1). GEOquery [24] is a package written for the R programming language (http://www.R-project.org/) that allow R users to easily connect, retrieve, parse and extract expression data from GEO ready to be used in downstream analysis. The ArrayExpress [25] R package works similarly to GEOquery but for the ArrayExpress database. GEOmetadb [26] and SRAdb [27] both allow the user to query GEO and SRA within the R environment, but they require to download an SQLite file that contains the totality of GEO/SRA metadata. Compendiumdb [28] is an R package framework used to parse and store expression information into a relational database that can be queried from within the R environment for subsequent analysis. VirtualArray [29] is another R package used to combine raw data from diverse microarray samples (or experiments) and generates a combined object for further analysis. It also implements several batch effect removal methods but it is not available for the latest Bioconductor version. Microarray retriever [30] is a web-application used to query and download expression data from both GEO and ArrayExpress, but is currently unavailable.

The main difference between all these tools, except for Microarray retriever, and COMMAND>_ is that all of

Table 1 Functionalities comparison between COMMAND>_ and other tools used to collect gene expression data from public databases

Tool	R	Local	GUI	GEO	AE	SRA	DB	ANN	Search	Note
COMMAND>_	NO	YES	YES	YES	YES	YES	YES	YES	YES	
GEOquery	YES	NO	NO	YES	NO	NO	NO	NO	NO	
ArrayExpress	YES	NO	NO	NO	YES	NO	NO	NO	YES	
GEOmetadb	YES	NO	NO	YES	NO	YES	NO	NO	YES	requires the dowload of an sqlite file with meta information
SRAdb	YES	NO	NO	YES	NO	YES	NO	NO	YES	requires the dowload of an sqlite file with meta information
compendiumdb	YES	NO	NO	YES	NO	NO	YES	NO	NO	
virtualArray	YES	YES	NO	YES	YES	NO	NO	NO	NO	allow to normalize data and correct for batch effect, available only for Bioconductor <= 2.14
Microarray retriever	NO	NO	YES	YES	YES	NO	YES	NO	YES	unavailable

Each column represent one functionality, respectively: R (the program is an R package), local (the program allow to use local data), GUI (the program provides a Graphical User Interface), GEO (the program connects to GEO), AE (the program connects to ArrayExpress), SRA (the program connects to SRA), DB (the program provides a database to store expression data), ANN (the program allows to annotate probes), Search (the program allow to perform queries using free text besides accession id) and Note (the program has special features or limitations)

them are Bioconductor packages and run within the R programming environment. The advantages are that Bioconductor is a strong and reliable environment and different packages can be used in combination to perform a vast amount of different analysis. Despite being a great tool for data analysis, R and Bioconductor are not meant for data retrieval, and management of large amount of data can be problematic since R programs, without specific packages such as parallel, are by default single-threaded process, the data are completely stored in RAM and thus don't easily scale to handle large datasets.

COMMAND>_ has been developed with the specific goal of simplifying this part. It relies on a relational database and a task queue system such as PostgreSQL and Celery respectively to easily scale when number of experiments grows significantly. In this regard they might be thought more as complementary tools with R to be used to analyse the datasets collected using COMMAND>_. In COMMAND>_ many operations can be done using the Graphical User Interface (GUI) such as the re-annotation tool which allows the user to produce an optimized annotation instead of relying on default ones. It is important to highlight that the re-annotation step allows perfect reproducibility of the analysis since all parameters are stored within COMMAND>_. Finally, despite being a graphical tool offering a friendly user experience, COMMAND>_ gives the same flexibility of a command-line environment to manage all possible situations through its Python editor.

Case study

In order to demonstrate COMMAND>_ functionalities, we present several case studies available within the on-line documentation. Moreover, we used it here for searching, downloading, parsing, re-annotating and exporting a collection of small airway samples from patients affected by Chronic Obstructive Pulmonary Disease (COPD) [31]. The original study is a collection of 273 samples from three Affymetrix microarray experiments retrieved from the Gene Expression Omnibus (GEO): GSE8545, GSE20257 and GSE11906. We start retrieving the GEO experiments used in the study using the "Download Experiment From Public Database" dialog with the GSE Series ID as term and GEO as database (Fig. 4a). Before starting to parse the experiments we need to import the gene sequences to be used for the probe mapping step. The parsing procedure starts by selecting one experiment and pressing the "Parse/Import experiment" button. The parsing interface is divided into three collapsible sections: the top one shows the experiment data preview, the middle one contains the experiment files browser and the assignment tool used to couple parsing scripts and experiment files, while the bottom section is the Python editor. Having the original probes is highly encouraged in order to take advantage of the probe to gene mapping functionality. Since probe sequences are not included in this experiment, we have to download them separately from the Affymetrix Support site and upload them into COMMAND>_ using the "Upload file" button. Once all files are in place we are ready to start assigning parsing scripts to the experiment files. Since we don't need to change any information related to the experiment entity we will start with platform-related files, i.e. HGU133Plus2_Hs_ENSG_probe_tab. The assignment procedure is itself based on the execution of a Python script, and in this way we can automatically assign a vast amount of files using user-defined rules. For this specific case we will tell COMMAND>_ to parse the "HGU133Plus2_Hs_ENSG_probe_tab" file using the "gpr_platform.py" script. To correctly parse the platform file we have to inform the "gpr_platform.py" about the field names to be used for the probe id and the probe sequence. The sample files assignment will proceed similarly by selecting all CEL files (we can use the filter by file names) and giving "cel_sample.py" as script to be used. This time we will use the "match_entity_name.py" assignment script in order to have COMMAND>_

to automatically couple CEL files with the corresponding samples (Fig. 4b). The last file to be assigned is the soft file that contains the meta-data for all entities, experiment, platform and samples. Once again we use the "assign_all.py" to assign "soft_experiment.py", "soft_platform.py" and "soft_sample.py" scripts to experiment, platform and samples respectively. After inspecting that all the assignments are correctly done we are ready to run the parsing scripts. Once the parsing is done we can inspect the results in the "Preview" section and import the experiment. We will have to repeat the same procedure for the remaining experiments.

Once that all the raw data are imported into the database we can map the probes for the GPL570 platform to the human genes we already imported. This fundamental step consists in two parts, the alignment and filtering of alignment results. For the alignment step we can chose a quite stringent identity threshold (such as 95% or 98%) since both probes and genes belong to the same species. The two-step filtering thresholds are set to 95% alignment length, 0 gap and 3 mismatches for the sensitivity step and 98% alignment length, 0 gap and 1 mismatch for the specificity step (Fig. 4c). The chosen threshold captures the idea that probes might align (even if not perfectly) on more than one gene resulting in an unusable probe, and, require a higher minimum alignment quality for a probe to be considered reliable. As stated previously this wouldn't be possible using only a single filter (Fig. 5). In this specific case choosing different thresholds will result in little differences since probes and genes come from the same organism. This step is increasingly relevant when more and more probes are designed for a different organism than the one we are using as the genomic background and for which we have the gene sequences, such as might be the case for different strains of bacteria or different cultivars of plant crops.

After the alignment process is complete, we can set the filtering parameters and run the filtering. Once the filtering is done, we are able to import the probe to gene mapping. Finally, we can export the resulting raw data in both TSV and HDF5 file format.

Conclusion

In this paper we present COMMAND>_, a web-based application used to download, collect and manage gene expression data from public databases. COMMAND>_ relies on a DBMS for data persistence and a set of customizable Python scripts to extract only relevant information from public gene expression databases. COMMAND>_ is a multi-user application that allow teamwork via definition of groups of users with specific privileges on each of the defined gene expression compendia. Moreover, it eases the long-time maintenance of such gene expression compendia storing a system log with all the relevant information about the operations performed. COMMAND>_ is a tool in constant development with new features to be added with newer versions. It is easily extendable to readily manage new technology platforms as they appear, for new data formats to be parsed, and even for new quantitative data types to be imported. This is reflected in the software architecture as well as in the data model.

Availability and requirements

Project name: COMMAND>_.

Project home page: https://github.com/marcomoretto/command

Operating systems: any supporting Docker Compose or Python 3 (tested on Linux).

Programming languages: Python, Javascript.

Other requirements: full requirements list available at https://raw.githubusercontent.com/marcomoretto/command/master/requirements.txt

License: GNU GPL v3.

Any restrictions to use by non-academics: none.

Abbreviations

COMMAND>_: COMpendia MANagement Desktop; COPD: Chronic Obstructive Pulmonary Disease; DBMS: DataBase Management System; GUI: Graphical User Interface; ORM: Object Relational Mapping; OTU: Operational Taxonomic Unit

Acknowledgements

Authors would like to thank Andrea Cattani and Patrizio Majer for the IT support.

Funding

This research was supported by the Autonomous Province of Trento (Accordo di Programma P1611051I). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Availability of data and materials

Documentation is available at https://command.readthedocs.io and a running demo of COMMAND>_ is available at https://command.fmach.it:4242. Gene FASTA file used in the case study is available at https://drive.google.com/file/d/1TJnsGnWdSjxhSlxgSe_au3X0E2Dc7nsx/view. Probe TAB file used in the case study is available at the Affymetrix Support site http://www.affymetrix.com/Auth/analysis/downloads/data/HG-U133_Plus_2.probe_tab.zip. The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Authors' contributions

KE and MM conceived the project. MM wrote the code and drafted the manuscript. PS, ABVA and KE tested the software and edited the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]Unit of Computational Biology, Research and Innovation Centre, Fondazione Edmund Mach, via E. Mach 1, 38010 San Michele all'Adige, Italy. [2]Laboratorio Internacional de Investigación Sobre el Genoma Humano, Universidad Nacional Autónoma De México, 76230 Juriquilla, Querétaro, Mexico. [3]Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, 62210 Cuernavaca, Morelos, Mexico.

Received: 5 November 2018 Accepted: 22 January 2019
Published online: 28 January 2019

References

1. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 1995; 270:467–70.
2. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5:621–8.
3. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 2008;320:1344–9.
4. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, et al. Highly integrated Single-Base resolution maps of the epigenome in Arabidopsis. Cell. 2008;133:523–36.
5. Evans TG. Considerations for the use of transcriptomics in identifying the 'genes that matter' for environmental adaptation. J Exp Biol. 2015;218:1925–35.
6. Gligorijević V, Pržulj N. Methods for biological data integration: perspectives and challenges. J R Soc Interface. 2015;12. https://doi.org/10.1098/rsif.2015.0571.
7. Rung J, Brazma A. Reuse of public genome-wide gene expression data. Nat Rev Genet. 2013;14:89–99.
8. Garrett-Mayer E, Parmigiani G, Zhong X, Cope L, Gabrielson E. Cross-study validation and combined analysis of gene expression microarray data. Biostatistics. 2008;9:333–54.
9. Xia J, Gill EE, Hancock REW. NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data. Nat Protoc. 2015;10: 823–44.
10. Papatheodorou I, Fonseca NA, Keays M, Tang YA, Barrera E, Bazant W, et al. Expression atlas: gene and protein expression across multiple studies and organisms. Nucleic Acids Res. 2018;46:D246–51.
11. Fucile G, Biase DD, Nahal H, La G, Khodabandeh S, Chen Y, et al. ePlant and the 3D data display initiative: integrative systems biology on the world wide web. PLoS One. 2011;6:e15237.
12. Kuo T-C, Tian T-F, Tseng YJ. 3Omics: a web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data. BMC Syst Biol. 2013;7:64.
13. Kamburov A, Cavill R, Ebbels TMD, Herwig R, Keun HC. Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. Bioinformatics. 2011;27:2917–8.
14. Moretto M, Sonego P, Dierckxsens N, Brilli M, Bianco L, Ledezma-Tejeida D, et al. COLOMBOS v3.0: leveraging gene expression compendia for cross-species analyses. Nucleic Acids Res. 2016;44:D620–3.
15. Moretto M, Sonego P, Pilati S, Malacarne G, Costantini L, Grzeskowiak L, et al. VESPUCCI: exploring patterns of gene expression in grapevine. Front Plant Sci. 2016;7. https://doi.org/10.3389/fpls.2016.00633.
16. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. 2013;41:D991–5.
17. Leinonen R, Sugawara H, Shumway M. The sequence read archive. Nucleic Acids Res. 2011;39(Database):D19–21.
18. Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, et al. ArrayExpress update--simplifying data submissions. Nucleic Acids Res. 2015;43(Database issue):D1113–6.
19. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.
20. Yin J, McLoughlin S, Jeffery IB, Glaviano A, Kennedy B, Higgins DG. Integrating multiple genome annotation databases improves the interpretation of microarray gene expression data. BMC Genomics. 2010;11:50.
21. Barbosa-Morais NL, Dunning MJ, Samarajiwa SA, Darot JFJ, Ritchie ME, Lynch AG, et al. A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data. Nucleic Acids Res. 2010;38:e17–7.
22. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
23. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34:525–7.
24. Davis S, Meltzer PS. GEOquery: a bridge between the gene expression omnibus (GEO) and BioConductor. Bioinformatics. 2007;23:1846–7.
25. Kauffmann A, Rayner TF, Parkinson H, Kapushesky M, Lukk M, Brazma A, et al. Importing ArrayExpress datasets into R/Bioconductor. Bioinformatics. 2009;25:2092–4.
26. Zhu Y, Davis S, Stephens R, Meltzer PS, Chen Y. GEOmetadb: powerful alternative search engine for the gene expression omnibus. Bioinformatics. 2008;24:2798–800.
27. Zhu Y, Stephens RM, Meltzer PS, Davis SR. SRAdb: query and use public next-generation sequencing data from within R. BMC Bioinformatics. 2013;14:19.
28. Nandal UK, Kampen V, AH C, Moerland PD. Compendiumdb: an R package for retrieval and storage of functional genomics data. Bioinformatics. 2016;32:2856–7.
29. Heider A, Alt R. virtualArray: a R/bioconductor package to merge raw data from different microarray platforms. BMC Bioinformatics. 2013;14:75.
30. Ivliev AE, Hoen PAC 't, Villerius MP, Dunnen D, T J, Brandt BW. Microarray retriever: a web-based tool for searching and large scale retrieval of public microarray data. Nucleic Acids Res 2008;36 suppl_2:W327–W331.
31. Yi G, Liang M, Li M, Fang X, Liu J, Lai Y, et al. A large lung gene expression study identifying IL1B as a novel player in airway inflammation in COPD airway epithelial cells. Inflamm Res. 2018;67:539–51.

3.2 Artículo: PulmonDB

Las enfermedades pulmonares, en especial EPOC, han sido estudiadas por diversos laboratorios con diferentes plataformas, por lo cual decidimos usar la herramienta de COMMAND para construir un compendio que tuviera datos transcriptómicos de EPOC y FPI. Este trabajo fue realizado en conjunto con los desarrolladores de COMMAND, además de especialistas en enfermedades pulmonares para la construcción del vocabulario estructurado.

Como primer paso, decidimos analizar los experimentos que se encuentran en repositorios públicos de EPOC y FPI. Los resultados fueron evaluados y curados buscando que los experimentos tuvieran muestras de estas enfermedades y que al menos la información el estado de la enfermedad estuviera disponible. Después se procedió a usar COMMAND para descargar los datos crudos, realizar su anotación, generar los contrastes individuales, re-anotar las sondas en el genoma hg19 de cada plataforma y normalizar los datos uniformemente, siendo la primera vez que se utilizó para datos de humanos.

El proceso de este trabajo fue publicado en el 2019, en el cual se describe los pasos que se siguieron para crear PulmonDB, una base de datos de expresión genética para enfermedades pulmonares (EPOC y FPI). En conjunto con los autores se describió un manual de uso el cual nos permitió describir más detalladamente cada uno de los pasos y criterios que se ocuparon para crear pulmonDB.

Adicionalmente, en el artículo se usó pulmonDB, para encontrar genes diferencialmente expresados en tejido pulmonar entre controles, enfermos de EPOC y FPI. El artículo se encuentra anexado a continuación.

3.2.1 Contribución personal

Participé en la realización de PulmonDB en todas sus etapas, la página web, el paquete de R, el análisis de EPOC y FPI, la escritura del artículo, y en la generación de las figuras.

3.2.2 Publicación 3: *PulmonDB: a curated lung disease gene expression database*

SCIENTIFIC REPORTS
nature research

OPEN

PulmonDB: a curated lung disease gene expression database

Ana B. Villaseñor-Altamirano[1], Marco Moretto[2], Mariel Maldonado[3], Alejandra Zayas-Del Moral[4], Adrián Munguía-Reyes[3], Yair Romero[5], Jair. S. García-Sotelo[1], Luis A. Aguilar[6], Oscar Aldana-Assad[1], Kristof Engelen[2], Moisés Selman[3], Julio Collado-Vides[4,7*], Yalbi I. Balderas-Martínez[3,8*] & Alejandra Medina-Rivera[1*]

Chronic Obstructive Pulmonary Disease (COPD) and Idiopathic Pulmonary Fibrosis (IPF) have contrasting clinical and pathological characteristics and interesting whole-genome transcriptomic profiles. However, data from public repositories are difficult to reprocess and reanalyze. Here, we present PulmonDB, a web-based database (http://pulmondb.liigh.unam.mx/) and R library that facilitates exploration of gene expression profiles for these diseases by integrating transcriptomic data and curated annotation from different sources. We demonstrated the value of this resource by presenting the expression of already well-known genes of COPD and IPF across multiple experiments and the results of two differential expression analyses in which we successfully identified differences and similarities. With this first version of PulmonDB, we create a new hypothesis and compare the two diseases from a transcriptomics perspective.

A common way to study diseases is by using transcriptomic analysis, which can reveal components of the genome that are active and help us understand which biological processes are affected[1]. Over the years, transcriptomic profiles have been compiled and published in public repositories such as Gene Expression Omnibus (GEO)[2,3] and ArrayExpress[4]. Having a way to compare transcriptomic data from Chronic Obstructive Pulmonary Disease (COPD) and Idiopathic Pulmonary Fibrosis (IPF) will help to identify common and distinct molecular mechanisms for these two diseases. However, an overwhelming task is to integrate high-throughput data from public repositories, because of platform differences (resulting in batch effects), heterogeneous experimental conditions, and the lack of uniformity on experimental annotations. Wang *et al.* reviewed different approaches in which they discussed tools such as GEO2R[5], ScanGEO[6], ImaGEO[7], BioJupies[8]. These tools reuse public data, reanalyze it consistently, and integrate additional data. Even with these available tools, performing meta-analyses is still challenging[9]. In particular, for COPD and IPF, because the information from only a few experiments is available in these resources, such an analysis requires manual annotation by the user or inclusion of only curated GEO Datasets (also referred as GDS), and only none of them integrates microarray and RNA-Seq data, to our knowledge.

Therefore, we created a curated gene expression lung disease database, PulmonDB, to organize the currently large amount of expression data for both COPD and IPF. To accomplish this task, we used COMMAND > _, a web application previously used to create two successful transcriptomic compendia: one for bacterial genomes, COLOMBOS[10,11], and the second for grapevine VESPUCCI[12]. While there are other chronic respiratory diseases, such as asthma, cystic fibrosis, and pulmonary hypertension association, among others, given the biological similarities between COPD and IPF, we decided to focus the first version of PulmonDB on these two diseases. We integrated transcriptomic experiments from different sources and their curated annotations, and built an online web resource to facilitate the exploration of gene expression profiles for COPD and IPF creating new hypotheses, and to allow for the identification of co-expression patterns in further analyses.

[1]Laboratorio Internacional de Investigación sobre el Genoma Humano, UNAM, Juriquilla, Mexico. [2]Unit of Computational Biology, Research and Innovation Centre, Fondazione Edmund Mach, 38010, San Michele all'Adige, Italy. [3]Instituto Nacional de Enfermedades Respiratorias Ismael Cosío Villegas, Mexico City, Mexico. [4]Center for Genomic Sciences, UNAM, Cuernavaca, Mexico. [5]Facultad de Ciencias, UNAM, Mexico City, Mexico. [6]Laboratorio Nacional de Visualización Científica Avanzada, LAVIS, UNAM, Juriquilla, Mexico. [7]Department of Biomedical Engineering, Boston University, Boston, Massachusetts, USA. [8]CONACYT-Instituto Nacional de Enfermedades Respiratorias Ismael Cosío Villegas, Mexico City, Mexico. *email: collado@ccg.unam.mx; yalbibalderas@gmail.com; amedina@liigh.unam.mx

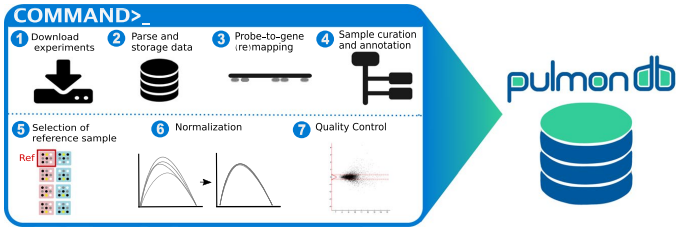


Figure 1. Flow chart of PulmonDB. PulmonDB was created using COMMAND by downloading, parsing and storing COPD and IPF public transcriptomic data into a MySQL database. Then, we remapped microarray probes to establish a uniform gene annotation, and we also created a controlled vocabulary for clinical and biological annotations for each sample. We created contrasts based on the original hypothesis, selecting a sample as the reference. Finally, the data were homogenized and subjected to a quality check.

Results

PulmonDB is a relational database implemented in MySQL with lung disease transcriptome measurements, re-annotated platform probes, and manually curated data with a controlled vocabulary designed for lung diseases (Fig. 1). Tables were created to describe each feature and to connect the information across experiments, samples, measurements, platforms, genes, and annotated information. The full database scheme is provided in Supplementary Fig. 1.

PulmonDB a curated gene expression lung disease database. PulmonDB is a curated gene expression database of human lung diseases, with RNA-seq and microarray data from different platforms that have been uniformly preprocessed and manually curated to add sample and experiment information. In addition, we developed a website to access and visualize homogenized data (http://pulmondb.liigh.unam.mx/), and we also developed an R package (https://github.com/AnaBVA/pulmondb) to download curated annotation and preprocessed data that can be used for further analysis in the R environment.

Our database has a total of 76 GSEs, corresponding to 4481 unique preprocessed GSM contrasts that used 26 different platforms or GPLs (platform ID from GEO) (Fig. 2C). PulmonDB contains different sample types, we searched for human gene expression experiments related to COPD and IPF without any restriction. Lung biopsies account for 37.8% of samples, and 33.2% are blood samples. However, different cell types can be found in PulmonDB: some of them are primary cells (*e.i.* alveolar macrophages, fibroblasts, alveolar epithelial cells, etc.), and others are cell lines (*e.i.* A549) (Fig. 2A). Of the samples, 34.9% correspond to COPD, 40.5% to control samples (30.9% healthy plus 9.6% match tissue), 17.2% to IPF, and 1.5% to other diseases (Fig. 2B and Supplementary Table 2). We separated control tissues into two groups, "healthy" individuals, as far as the authors are aware and "match_tissue_controls" which refers to tissue samples from a phenotypically healthy region of a patient who had a tumor removed (*e.i.* non-tumor tissue from a cancer patient).

Although other resources reuse and reanalyze GEO data using web interfaces[9], those tools are not specialized for lung diseases. Their limitations include the need for previous manual curation in each analysis, and they consider a small number of COPD and IPF experiments due to the fact that only curated GEO data are used. We designed a web interface that enables data exploration and visualization to facilitate lung disease analysis. This interface uses Clustergrammer[13] to visualize gene expression values and the creation of interactive heatmaps that allow data exploration. A valuable feature of Clustergrammer is to be connected to EnrichR[14], which provides pathway enrichment analysis. All these features together should help to generate new hypotheses about the pathologies of lung diseases to perform exploratory analyses, to visualize specific gene expression across public experiments for comparing results, and to generate new insights based on different data sets.

PulmonDB can recapitulate gene expression patterns expected in COPD and IPF. To show that PulmonDB can be used to recapitulate previously reported knowledge regarding COPD and IPF biology, we performed a literature search and manually selected relevant genes for each disease. We selected 19 genes related to IPF (not necessarily associated with gene expression in lung tissues) to visualize their gene expression: CCL18[15], CXCL12[16], CXCL13[17], collagens (COL1A1, COL1A2, COL3A1, COL5A2, COL14A1)[18], DSP[19], FAS[20], IL-8[21], MMP1[22], MMP2[23], MMP7[22], MUC5B[19], SPP1[24], PTGS2[25], TGFB1[26] and THY1[27]. Then, we selected eight IPF experiments performed with lung tissue biopsy samples (GSE32537, GSE21369, GSE24206, GSE94060, GSE72073, GSE35145, GSE31934), and using the PulmonDB website, we created a heatmap with the gene expression patterns and observed that the hierarchical clustering of these data separates IPF and control data sets (Fig. 3A, green and gray clusters at the bottom). For COPD, we curated 16 genes from the literature that were deemed relevant to this disease: HHIP[28,29], CFTR[30,31], PPARG[32], SERPINA1[33,34], JUN[35], FAM13A[36], MYH10[35], CHRNA5[37], JUND[35], JUNB[35], TNF[34], MMP9[34], MMP12[34], CHRNA3[37], TGFBR3[32], and GATA2[32]. We selected five experiments (GSE27597, GSE37768, GSE57148, GSE8581, GSE1122) performed on lung tissue biopsy samples from COPD patients and controls. Our hierarchical clustering analysis of the expression profiles using the PulmonDB interface allowed us to cluster patients and controls into two different groups (Fig. 3B), similar to the case of IPF. In conclusion, PulmonDB not only helps to recapitulate previously published work (Supplementary

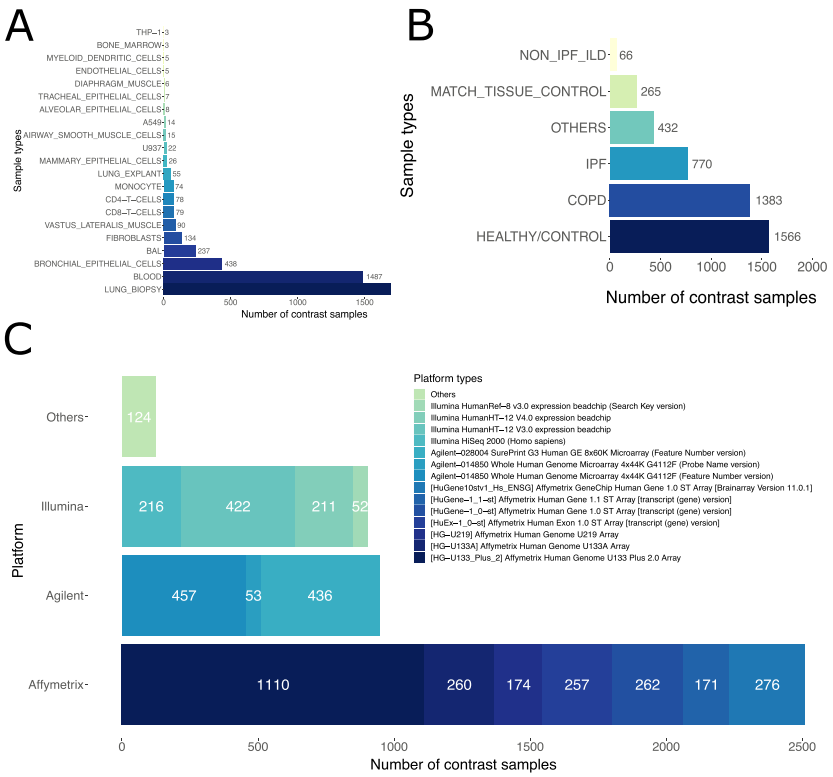


Figure 2. Summary of PulmonDB. (**A**) The number of contrast samples in PulmonDB per biological sample type. (**B**) The number of sample states found in PulmonDB. The color key below the bar chart shows the sectors for COPD patients, healthy/controls, IPF patients, match_tissue_controls (non-cancerous sample from a cancer patient), and other diseases (such as asthma). (**C**) The number of contrast samples measured using each platform (clustered by using Affymetrix, Agilent, Illumina, and other platforms with fewer samples).

Fig. 3) but also helps to verify gene expression stability across experiments. This may help to analyze concordance in different experiments, contrast study results, show implications of using different control groups, etc. We believe this resource can be used to drive, make decisions, and support new hypotheses in experimental laboratories for studying molecular or cellular disease mechanisms.

Differences and similarities in COPD and IPF. PulmonDB can be used not only to replicate previous knowledge but also to provide a framework to test new hypotheses. In this context, we set out to investigate the differences and similarities between COPD and IPF in lung tissue when compared to samples from healthy individuals (Fig. 4A). Using PulmonDB in the R environment, we selected contrasts where the sample was annotated as lung biopsy and the reference status as HEALTHY/CONTROLs (GSE52463, GSE63073, GSE1122, GSE72073, GSE24206, GSE27597, GSE29133, GSE31934, GSE37768) (Fig. 4B), and then using limma[38] we assessed differential gene expression between COPD and IPF. We identified 1781 differentially expressed genes (Supplementary Fig. 4). To have a visual representation of the differences between COPD and IPF, we selected the top 20 differentially expressed genes and visualized their expression using the PulmonDB website tool (Fig. 4C). We observed that data sets tend to cluster by test status; Fig. 4C shows IPF contrasts on the left (turquoise), control contrasts in the middle (blue), and COPD contrasts on the right (red). Genes are clustered in two groups (left panel, y-axis); the first gene group (I) is overexpressed in IPF while it is barely expressed or underexpressed in COPD contrasts. By comparison, the second gene cluster (group II) is overexpressed in COPD contrasts and underexpressed in IPF. To correlate similarities among samples, the 20 top differentially expressed genes were used (Fig. 4C, right panel); samples from the same disease group showed higher correlations and tended to have a null or negative correlation with the HEALTHY/CONTROL and the opposite disease (Fig. 4C). For example, FOSB and CXCL2 have opposite behaviors, as both genes are overexpressed in COPD and underexpressed in IPF. FOSB is part of the family of Fos genes that can dimerize with JUN family proteins to form the transcription factor complex AP-1, which is related to COPD[39]. CXCL2 is a chemokine secreted in inflammation that induces chemotaxis in

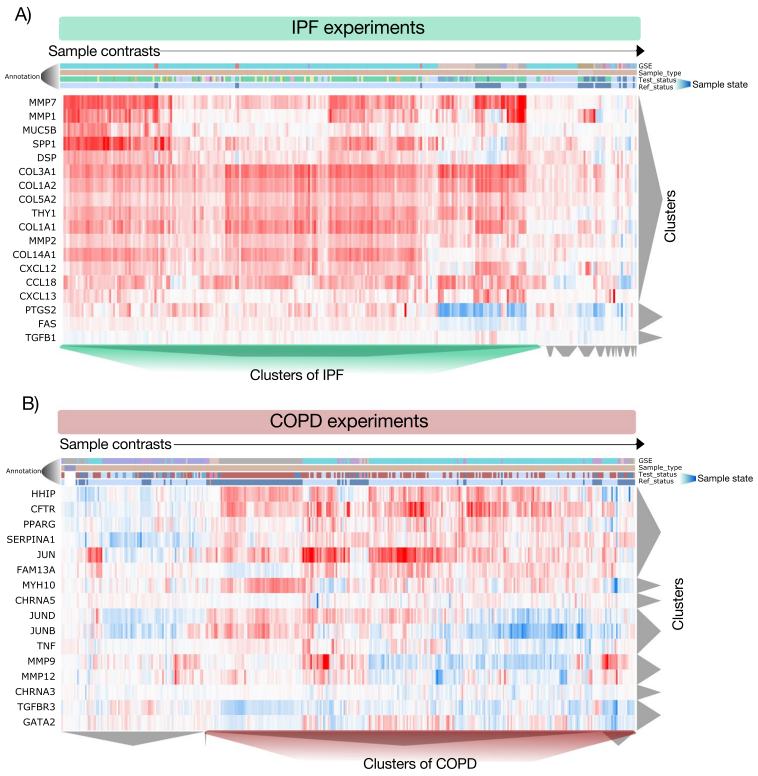


Figure 3. IPF and COPD well-known disease-associated genes. In both heatmaps, rows are genes, and columns are sample contrasts. Both were hierarchically clustered. The first annotation row represents their GSE IDs. The second annotation row is the sample type, LUNG_BIOPSY samples, in light brown. The third and the fourth annotation rows are sample states, the third annotation row represents the test state, and the fourth annotation row is the reference state. (**A**) IPF genes reported being relevant in the literature (CCL18[15], CXCL12[16], CXCL13[17], COL1A1, COL1A2, COL3A1, COL5A2, COL14A1[18], DSP[19], FAS[20], IL-8[21], MMP1[22], MMP2[23], MMP7[22], MUC5B[19], SPP1[24], PTGS2[25], TGFB1[26] and THY1[27]). The IPF experiments selected were GSE32537 (pink), GSE21369 (purple), GSE24206 (blue), GSE94060 (grass-green), GSE72073 (lemon yellow), GSE35145 (green), and GSE31934 (yellow). The third and the fourth annotation rows are sample states: light blue, MATCH_TISSUE_CONTROL; dark blue, HEALTHY/CONTROL; turquoise, IPF samples; and grey, NON_IPF_ILD. (**B**) COPD genes reported being relevant in the literature (HHIP[28,29], CFTR[30,31], PPARG[32], SERPINA1[33,34], JUN[35], FAM13A[36], MYH10[35], CHRNA5[37], JUND[35], JUNB[35], TNF[34], MMP9[34], MMP12[34], CHRNA3[37], TGFBR3[32], and GATA2[32]). The COPD experiments selected were GSE27597, GSE37768, GSE57148, GSE8581, and GSE1122. The third and the fourth annotation rows are sample states: light blue, MATCH_TISSUE_CONTROL; dark blue, HEALTHY/CONTROL; red, COPD samples.

neutrophils[40,41]; these cells are predominant in COPD, and they are key mediators in tissue damage[42]. While neutrophils are also important in IPF, we observed their underexpression in this disease.

We also asked the opposite question, *i.e.*, whether we could identify which genes that are shared between these two diseases. We assigned a weight to COPD and IPF expression to perform limma contrasts (Fig. 4D), which enabled us to identify when both diseases drove a differential expression profile. We selected the 20 top differentially expressed genes and visualized their expression patterns using PulmonDB website tool, and we could see that a set of genes was consistently overexpressed or underexpressed in both COPD and IPF. In particular, VCAM1 and FCN3 are differentially expressed in COPD and IPF, with a similar trend in both diseases when compared with HEALTHY/CONTROLs. VCAM1 is the vascular cell adhesion molecule 1, and it is important in the immune response for mediating cellular adhesion in leukocytes[43]; it is overexpressed in these two diseases, suggesting infiltration of immune cells in both pathologies[44,45]. In contrast, FCN3 (or ficolin 3) is underexpressed in both diseases: this gene is a collagen-like protein associated with the innate immune defense, as it activates the lectin complement pathway[46], which has been shown to be important in pulmonary pathologies[47,48].

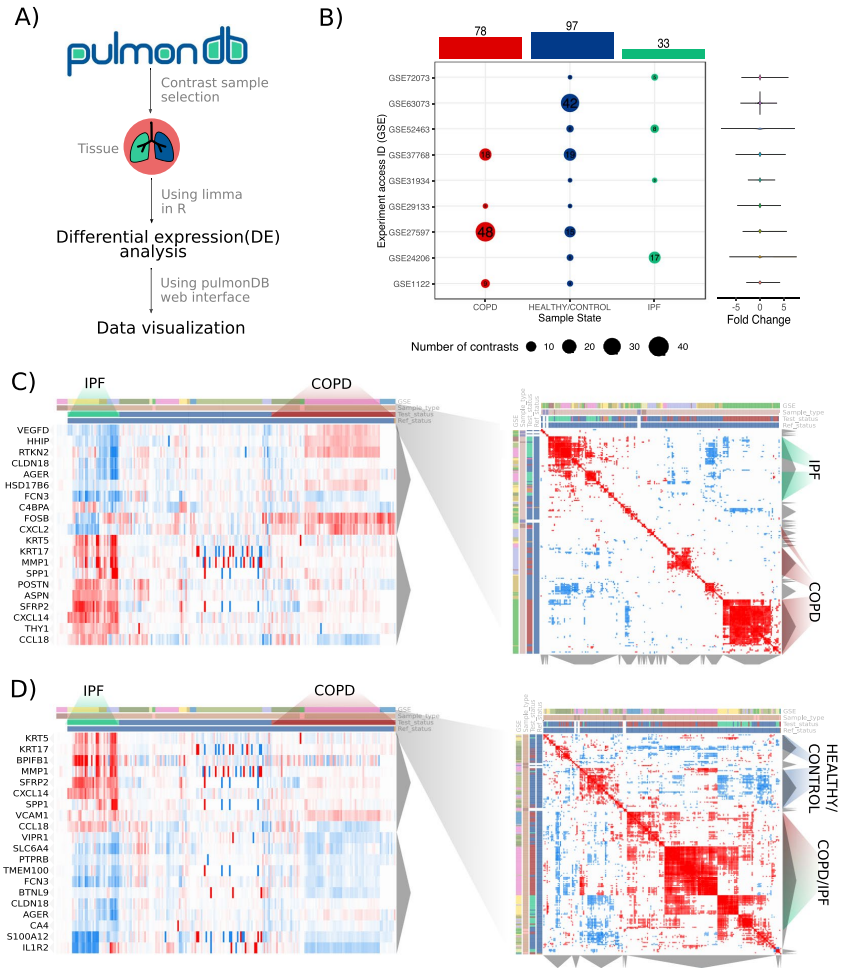


Figure 4. IPF and COPD differentially expressed and similarly expressed genes. (**A**) Flow chart of steps used for COPD and IPF differential expression analysis to evaluate transcriptomic differences and similarities. (**B**) Experiments selected for the analysis, following the criteria of being lung biopsy samples and contrasted with HEALTHY/CONTROL references. The colors represent the sample state: COPD, red; HEALTH/CONTROL, blue; IPF, turquoise. At the top, the bar graph is the total sum of contrasts, rows are the GSE experiments, and each dot is the number of contrasts per experiment from COPD, HEALTHY/CONTROL, or IPF subjects. On the right side, we can see the distributions in violin plots for all sample contrasts per experiment. (**C**) Differentially expressed genes between COPD and IPF. (**D**) Similar genes between COPD and IPF. In both (**C, D**) columns are sample contrasts, rows are genes, the first covariate is colored by each corresponding experiment, the second covariate is the sample type (in this case, lung tissue is shown in light brown), the third row is the test status, and the fourth is the reference status. Columns are ordered by test status and genes by hierarchical clusterization. The right heatmap is the correlation among sample contrasts, and the covariates are the same.

As a result, PulmonDB assisted our analysis of COPD and IPF analogous and antagonist genes and can thus be used to dissect common molecular mechanisms, because both lung diseases are present under heterogeneous conditions with progressive and irreversible phenotypes mainly caused by smoking and by aging, plus both diseases entail cellular matrix remodeling. Furthermore, the differential gene signatures between COPD and IPF might explain the particularities of each disease.

Discussion

The present methodology had been previously applied for the study of bacterial and grapevine gene expression in different experiments and conditions, allowing for the integration of data from a diverse origin. Here we prove this methodology can also be applied to human data to exploit publicly available resources better, we hope these methods will be taken by other teams to create databases to help understand relevant diseases in other tissues.

PulmonDB can help the scientific community to study which genes have a distinct expression profile in COPD and IPF, explore experiments across technologies and platforms, identify interesting expression patterns across different diseases, generate new hypotheses, and find relationships among clinical or experimental variables. This database also enables comparisons of an updated collection of expression profiles already homogenized for their analyses of specific diseases. Additionally, having different lung diseases (COPD and IPF) in the same database creates the opportunity to observe their similarities and differences. In the future, we aim for PulmonDB to grow and include more diseases. To our knowledge, there is no other resource for transcriptomic analysis focused on the same lung diseases; for this reason, we believe researchers of different backgrounds can use and benefit from the information contained in PulmonDB, by using the web interface and the R package.

An integrated comparable collection of homogenized values with controlled vocabulary describing biological and technical characteristics will facilitate further comparative analyses, such as the study of profiles in COPD and IPF, exploration of experiments across technologies and platforms, identification of interesting coexpression patterns across different diseases, the generation of new hypotheses, and determination of relationships among clinical or experimental variables.

This project sets the foundation to integrate transcriptomics data of other respiratory diseases or related phenotypes and thus facilitates the identification of common and divergent pathways that lead to a pathological state. PulmonDB platform will be expanded in the future to include other lung diseases.

Methods

Platform and metadata. Most of the metadata was obtained from GEO. For specific cases, the platform information (.cdf file) was obtained from the Affymetrix website (http://www.affymetrix.com/site/mainPage.affx). Additional information (*e.g.*, clinical data, source of the biological sample), was obtained either from metadata or manually curated from the original papers.

Inclusion criteria for transcriptome data. The experiments currently included in PulmonDB are listed in Supplementary Table 2.

We used two main resources to download raw data and preprocessed counts, GEO and Recount2.

Gene expression omnibus. Using GEO[2,3], we searched datasets related to COPD and IPF for gene expression data. The following queries were used to retrieve the experiments:

(“*pulmonary disease, chronic obstructive*”[*MeSH Terms*] *OR COPD*[*All Fields*]) *AND* “*Homo sapiens*”[*porgn*] *AND* (“*gse*”[*Filter*] *AND* (“*Expression profiling by array*”[*Filter*] *OR* “*Expression profiling by high throughput sequencing*”[*Filter*])).

(*Idiopathic pulmonary fibrosis*[*All Fields*] *AND* “*Homo sapiens*”[*porgn*] *AND* (“*gse*”[*Filter*] *AND* (“*Expression profiling by array*”[*Filter*] *OR* “*Expression profiling by high throughput sequencing*”[*Filter*])).

GEO experiments were manually curated, abstracts and related articles were revised, and only datasets confirmed as having COPD and/or IPF samples were considered. In order for an experiment to be included in PulmonDB we used the following criteria: The data set had to be original, samples had to be unique, raw data had to be public and available, platform information must had sequence probes, and custom platforms must have had information to link raw expression signal with the probe sequence. Otherwise, data sets were not taken into account for PulmonDB.

Recount2. Recount2 is an online resource with RNA-seq human experiments already preprocessed using Rail-RNA alignment and summarized by gene and exon counts[49]. We used the keywords “IPF” and “COPD” separately in Recount2 to retrieve counts form RNA-seq.

Compendium creation. The compendium creation process was done as previously described in COLOMBOS and VESPUCCI[10,12]. The platform was developed in bacteria and later employed in grapevine, but in this paper, we used COLOMBOS for the first time in human data. After we selected the datasets using the experiment ID from GEO (GSE), we worked on COMMAND>_[50].

COMMAND. COMMAND stands for COMpendia MANagement Desktop, it is a web application tool that provides a framework to facilitate and perform the following steps: (1) download data from selected experiments, (2) parse files and store data in database form, (3) probe-to-gene (re)mapping process, (4) sample curation and annotation with a controlled vocabulary, (5) selection of references and sample experiments to determine contrasts, (6) homogenization (and normalization) of data, and (7) perform data quality control (Fig. 1). This software can be used for any transcriptomic data[50].

In more detail, each experiment with a GSE ID, also referred to as a data set, was normalized independently without performing background correction, as explained in[11]. We defined a contrast for each sample with a GSM ID (sample ID from GEO) by using a unique control reference sample per data set. The sample contrast per gene was defined as the log ratio between the expression value in the test condition (*i.e.*, IPF, COPD) and the

expression value in the reference condition (*i.e.*, healthy, untreated, smokers without COPD) (Fig. 1, step 5). This gives every comparison an interpretable biological meaning when combined with extensive manual curated annotation. The condition properties describing the contrasts were then structured in a condition-controlled vocabulary tree. Finally, all contrasts were homogenized, resulting in direct comparable log ratios across all experiments; this information later became part of the final compendium of expression data (Supplementary Fig. 2).

PulmonDB uses a controlled vocabulary to describe sample metadata. A controlled vocabulary is required to create databases with homogeneous and standard information. For PulmonDB, we created a controlled vocabulary organized in a hierarchical structure that contains terms to annotate transcriptome experiments in lung diseases. We defined classes describing the main categories and terms that can be found in experiments, with some of them as mandatory features (*e.i.* sample type, sample status, and platform). Some non-IPF or non-COPD diseases were included in the controlled vocabulary because the original experiments used them.

Once the controlled vocabulary was established, each article related to the experiment was manually curated, and whenever it was necessary, new terms were added, making the vocabulary flexible and allowing for the inclusion of other diseases to our database in the future. Complete definitions of the terms are provided in Supplementary Table 1.

Experiment annotation. Each sample was manually annotated using the controlled vocabulary; when necessary, the vocabulary was updated to include new features. The information was curated by experts who reviewed the associated articles and protocols to retrieve data such as age, sex, ancestry, stage of disease or treatment, DLCO (the diffusing capacity of the lung for carbon monoxide, a common functional test), etc., from either GEO or the associated paper.

Homogenization and quality control. As described before, data homogenization was done with COMMAND>_[11,12]. This step was performed on raw data without background correction, as it has been shown to retrieve more errors[51–53]. A nonlinear model was applied to homogenize raw data. We used RMA Quantile for Affymetrix samples and loess fit for the other platforms. The next step was to summarize probes per transcript using RMA median polish summary from Affymetrix or with data averaged across replicates for the other platforms. After performing the homogenization step, low-quality microarrays were identified using MA plots and histograms of raw and homogenized data.

Website implementation. PulmonDB has a web interface that uses Clustergrammer (https://clustergrammer.readthedocs.io/index.html)[13] to visualize gene expression contrasts. Clustergrammer has a frontend in javascript and a backend in python, supporting an interactive web application for gene expression exploration. The PulmonDB web interface requires one or several GSE identifiers and more than two gene names to generate interactive heatmaps.

In addition, Clustergrammer is connected with EnrichR (http://amp.pharm.mssm.edu/Enrichr/)[14], an integrative web application tool for enrichment analysis that helps the user explore not only potentially differentiated genes but also enriched pathways, facilitating the discovery of transcriptomic signature patterns in lung diseases or related phenotypes.

COPD and IPF comparative analysis. We used limma 3.40.0 in a Rstudio environment 3.6.0 for our comparative analyses, and the GSE ID was included in the linear model. Then, two contrasts were created: (1) "COPD – IPF", for obtaining differentially expressed genes between COPD and IPF, and (2) "(COPD + IPF)/2 – CONTROL", for genes similarly expressed between COPD/IPF and CONTROL. Differential gene expression analyses were adjusted for multiple testing using the false discovery rate (FDR) method, also referred to as Benjamini & Hochberg adjustment. We applied a cutoff of the adjusted p-value < 0.05, and after sorting based on the log fold change, the top 20 genes were obtained.

Data availability

PulmonDB is accessible (http://pulmondb.liigh.unam.mx/) and through an R package (https://github.com/AnaBVA/pulmondb).

Received: 12 September 2019; Accepted: 5 December 2019;
Published online: 16 January 2020

References

1. Qian, X., Ba, Y., Zhuang, Q. & Zhong, G. RNA-Seq technology and its application in fish transcriptomics. *OMICS* **18**, 98–110 (2014).
2. geo. Home - GEO - NCBI. Available at: https://www.ncbi.nlm.nih.gov/geo/. (Accessed: 21st July 2019)
3. Clough, E. & Barrett, T. The Gene Expression Omnibus Database. *Methods Mol. Biol.* **1418**, 93–110 (2016).
4. EMBL-EBI. ArrayExpress <EMBL-EBI. Available at: https://www.ebi.ac.uk/arrayexpress/. (Accessed: 21st July 2019)
5. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.* **41**, D991–5 (2013).
6. Koeppen, K., Stanton, B. A. & Hampton, T. H. ScanGEO: parallel mining of high-throughput gene expression data. *Bioinformatics* **33**, 3500–3501 (2017).
7. Toro-Domínguez, D. *et al.* ImaGEO: integrative gene expression meta-analysis from GEO database. *Bioinformatics* **35**, 880–882 (2019).
8. Torre, D., Lachmann, A. & Ma'ayan, A. BioJupies: Automated Generation of Interactive Notebooks for RNA-Seq Data Analysis in the Cloud. *Cell Syst* **7**, 556–561.e3 (2018).
9. Wang, Z., Lachmann, A. & Ma'ayan, A. Mining data and metadata from the gene expression omnibus. *Biophys. Rev.* **11**, 103–110 (2019).

10. Moretto, M. *et al*. COLOMBOS v3.0: leveraging gene expression compendia for cross-species analyses. *Nucleic Acids Res.* **44**, D620–3 (2016).
11. Engelen, K. *et al*. COLOMBOS: access port for cross-platform bacterial expression compendia. *PLoS One* **6**, e20938 (2011).
12. Moretto, M. *et al*. VESPUCCI: Exploring Patterns of Gene Expression in Grapevine. *Front. Plant Sci.* **7**, 633 (2016).
13. Fernandez, N. F. *et al*. Clustergrammer, a web-based heatmap visualization and analysis tool for high-dimensional biological data. *Sci Data* **4**, 170151 (2017).
14. Chen, E. Y. *et al*. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128 (2013).
15. Cai, M. *et al*. CCL18 in serum, BAL fluid and alveolar macrophage culture supernatant in interstitial lung diseases. *Respir. Med.* **107**, 1444–1452 (2013).
16. Antoniou, K. M. *et al*. Expression analysis of angiogenic growth factors and biological axis CXCL12/CXCR4 axis in idiopathic pulmonary fibrosis. *Connect. Tissue Res.* **51**, 71–80 (2010).
17. Vuga, L. J. *et al*. C-X-C motif chemokine 13 (CXCL13) is a prognostic biomarker of idiopathic pulmonary fibrosis. *Am. J. Respir. Crit. Care Med.* **189**, 966–974 (2014).
18. Jenkins, R. G. *et al*. Longitudinal change in collagen degradation biomarkers in idiopathic pulmonary fibrosis: an analysis from the prospective, multicentre PROFILE study. *Lancet Respir Med* **3**, 462–472 (2015).
19. Allen, R. J. *et al*. Genetic variants associated with susceptibility to idiopathic pulmonary fibrosis in people of European ancestry: a genome-wide association study. *Lancet Respir Med* **5**, 869–880 (2017).
20. Huang, S. K. *et al*. Histone modifications are responsible for decreased Fas expression and apoptosis resistance in fibrotic lung fibroblasts. *Cell Death Dis.* **4**, e621 (2013).
21. Yang, L. *et al*. IL-8 mediates idiopathic pulmonary fibrosis mesenchymal progenitor cell fibrogenicity. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **314**, L127–L136 (2018).
22. Rosas, I. O. *et al*. MMP1 and MMP7 as potential peripheral blood biomarkers in idiopathic pulmonary fibrosis. *PLoS Med.* **5**, e93 (2008).
23. García-Alvarez, J. *et al*. Membrane type-matrix metalloproteinases in idiopathic pulmonary fibrosis. *Sarcoidosis Vasc. Diffuse Lung Dis.* **23**, 13–21 (2006).
24. Pardo, A. *et al*. Up-regulation and profibrotic role of osteopontin in human idiopathic pulmonary fibrosis. *PLoS Med.* **2**, e251 (2005).
25. Parra, E. R., Lin, F., Martins, V., Rangel, M. P. & Capelozzi, V. L. Immunohistochemical and morphometric evaluation of COX 1 and COX-2 in the remodeled lung in idiopathic pulmonary fibrosis and systemic sclerosis. *J. Bras. Pneumol.* **39**, 692–700 (2013).
26. Martinez, F. J. *et al*. Idiopathic pulmonary fibrosis. *Nat Rev Dis Primers* **3**, 17074 (2017).
27. Sanders, Y. Y. *et al*. Thy-1 promoter hypermethylation: a novel epigenetic pathogenic mechanism in pulmonary fibrosis. *Am. J. Respir. Cell Mol. Biol.* **39**, 610–618 (2008).
28. Zhou, X. *et al*. Identification of a chronic obstructive pulmonary disease genetic determinant that regulates HHIP. *Hum. Mol. Genet.* **21**, 1325–1335 (2012).
29. Chang, W.-A., Tsai, M.-J., Jian, S.-F., Sheu, C.-C. & Kuo, P.-L. Systematic analysis of transcriptomic profiles of COPD airway epithelium using next-generation sequencing and bioinformatics. *Int. J. Chron. Obstruct. Pulmon. Dis.* **13**, 2387–2398 (2018).
30. Rab, A. *et al*. Cigarette smoke and CFTR: implications in the pathogenesis of COPD. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **305**, L530–41 (2013).
31. Campbell, J. D. *et al*. A gene expression signature of emphysema-related lung destruction and its reversal by the tripeptide GHK. *Genome Med.* **4**, 67 (2012).
32. Hedström, U. *et al*. Bronchial extracellular matrix from COPD patients induces altered gene expression in repopulated primary human bronchial epithelial cells. *Sci. Rep.* **8**, 3502 (2018).
33. Lackey, L., McArthur, E. & Laederach, A. Increased Transcript Complexity in Genes Associated with Chronic Obstructive Pulmonary Disease. *PLoS One* **10**, e0140885 (2015).
34. Kotnala, S., Tyagi, A. & Muyal, J. P. rHuKGF ameliorates protease/anti-protease imbalance in emphysematous mice. *Pulm. Pharmacol. Ther.* **45**, 124–135 (2017).
35. Kim, W. J. *et al*. Comprehensive Analysis of Transcriptome Sequencing Data in the Lung Tissues of COPD Subjects. *Int. J. Genomics Proteomics* **2015**, 206937 (2015).
36. Yun, J. H. *et al*. Transcriptomic Analysis of Lung Tissue from Cigarette Smoke-Induced Emphysema Murine Models and Human Chronic Obstructive Pulmonary Disease Show Shared and Distinct Pathways. *Am. J. Respir. Cell Mol. Biol.* **57**, 47–58 (2017).
37. Matsson, H. *et al*. Targeted high-throughput sequencing of candidate genes for chronic obstructive pulmonary disease. *BMC Pulm. Med.* **16**, 146 (2016).
38. Ritchie, M. E. *et al*. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
39. Mroz, R. M., Holownia, A., Chyczewska, E. & Braszko, J. J. Chronic obstructive pulmonary disease: an update on nuclear signaling related to inflammation and anti-inflammatory treatment. *J. Physiol. Pharmacol.* **59**(Suppl 6), 35–42 (2008).
40. Kim, D. & Haynes, C. L. Neutrophil chemotaxis within a competing gradient of chemoattractants. *Anal. Chem.* **84**, 6070–6078 (2012).
41. Larsson, K. Aspects on pathophysiological mechanisms in COPD. *J. Intern. Med.* **262**, 311–340 (2007).
42. Hoenderdos, K. & Condliffe, A. The neutrophil in chronic obstructive pulmonary disease. *Am. J. Respir. Cell Mol. Biol.* **48**, 531–539 (2013).
43. Ley, K. & Huo, Y. VCAM-1 is critical in atherosclerosis. *The Journal of clinical investigation* **107**, 1209–1210 (2001).
44. Nakao, A., Hasegawa, Y., Tsuchiya, Y. & Shimokata, K. Expression of cell adhesion molecules in the lungs of patients with idiopathic pulmonary fibrosis. *Chest* **108**, 233–239 (1995).
45. Davis, B. B. *et al*. Leukocytes are recruited through the bronchial circulation to the lung in a spontaneously hypertensive rat model of COPD. *PLoS One* **7**, e33304 (2012).
46. Garred, P., Honoré, C., Ma, Y. J., Munthe-Fog, L. & Hummelshøj, T. MBL2, FCN1, FCN2 and FCN3-The genes behind the initiation of the lectin pathway of complement. *Mol. Immunol.* **46**, 2737–2744 (2009).
47. Pandya, P. H. & Wilkes, D. S. Complement system in lung disease. *Am. J. Respir. Cell Mol. Biol.* **51**, 467–473 (2014).
48. Eisen, D. P. Mannose-binding lectin deficiency and respiratory tract infection. *J. Innate Immun.* **2**, 114–122 (2010).
49. Collado-Torres, L. *et al*. Reproducible RNA-seq analysis using recount2. *Nat. Biotechnol.* **35**, 319–321 (2017).
50. Moretto, M., Sonego, P., Villaseñor-Altamirano, A. B. & Engelen, K. First step toward gene expression data integration: transcriptomic data acquisition with COMMAND>_. *BMC Bioinformatics* **20**, 54 (2019).
51. Irizarry, R. A. *et al*. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264 (2003).
52. Fujino, N. *et al*. Gene expression profiles of alveolar type II cells of chronic obstructive pulmonary disease: a case-control study. *BMJ Open*, **2**, (2012).
53. Golpon, H. A. *et al*. Emphysema lung tissue gene expression profiling. *Am. J. Respir. Cell Mol. Biol.* **31**, 595–600 (2004).

Acknowledgements

We are thankful to the colleagues who help us installing and maintaining the PulmonDB server, particularly Luis Alberto Aguilar -Bautista and members of Laboratorio Nacional de Visualización Científica Avanzada, México. We are grateful to Miguel Negreros for discussing the concepts for curation and Orlando Santillán for his insights for parsing GEO data. We thank Alejandra Castillo and Carina Uribe for technical assistance. We thank Mauricio Guzmán and Centro Cultural Cine y Arte, particularly Renata Campuzano and Diego Morales for graphical and design assistance. Y.I.B.-M. acknowledges the Cátedras CONACyT program. We also acknowledge Catalina Frank, José Antonio Alonso and the Manuscript Writing Training Team (CEMAI for its Spanish acronym) of CONACyT for their help with the structure, reviews and constructive criticism of this research paper. A.M.-R.'s laboratory is supported by a CONACYT grant [269449], Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica – Universidad Nacional Autónoma de México (PAPIIT-UNAM) grant [IA206517-IA201119] and Estímulos a Investigaciones Médicas "Miguel Alemán Valdés"; J.C.-V., A.M.-R., Y.I.B.-M., and M.S., further acknowledge CONACYT "Fronteras de la Ciencia" support [Project 15]. A.B.V.-A. is a doctoral student from the Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM) and has received CONACyT fellowship CVU 557690.

Author contributions

A.B.V.-A. prepared and created all figures. A.B.V.-A., Y.I.B.-M. and A.M.-R. wrote the main manuscript text. Mariel M., A.M.-R., Y.R. and Y.I.B.-M. manually curate the data. A.B.V.-A., Marco M., E.K. and A.Z.-D.M. downloaded, processed and analysed data, and created the database. A.B.V.-A. and O.A.-A. created the R package. A.B.V.-A., J.S.G.-S., O.A.-A. and L.A.A. built, modified and created the web interface. M.S., J.C.-V., Y.I.B.-M. and A.M.-R. jointly supervised this work. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-019-56339-5.

Correspondence and requests for materials should be addressed to J.C.-V., Y.I.B.-M. or A.M.-R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Capítulo 4

Robustez en PulmonDB

PulmonDB, en su artículo, mostró la capacidad de replicar conocimiento previamente generado, así como poder ser usado para integrar experimentos de diferentes laboratorios e identificar genes que varían entre EPOC y FPI, pero también aquellos genes que se alteran en ambas enfermedades (Villaseñor-Altamirano et al., 2020).

La comunidad de Turing Way (Arnold et al., 2019) ha definido términos para poder describir de una manera más precisa el trabajo que se realiza. La figura 4.1 representa la definición de los conceptos de una manera muy visual. En donde se define *Reproducibilidad* como la capacidad de generar los mismos resultados cuando se usan los mismos datos y el mismo análisis, *Replicabilidad* cuando los datos son distintos pero el análisis es el mismo, *Robustez* cuando los datos son los mismos pero se ocupa un análisis disinto, y *Generalizado* cuando los datos y el análisis cambian.

En este capítulo, se busca demostrar la *robustez* de PulmonDB, usando los mismos datos pero con análisis distintos. Para ello se ocuparán los datos normalizados de PulmonDB y se contrastarán con el flujo de trabajo clásico usado para microarreglos y RNA-seq.

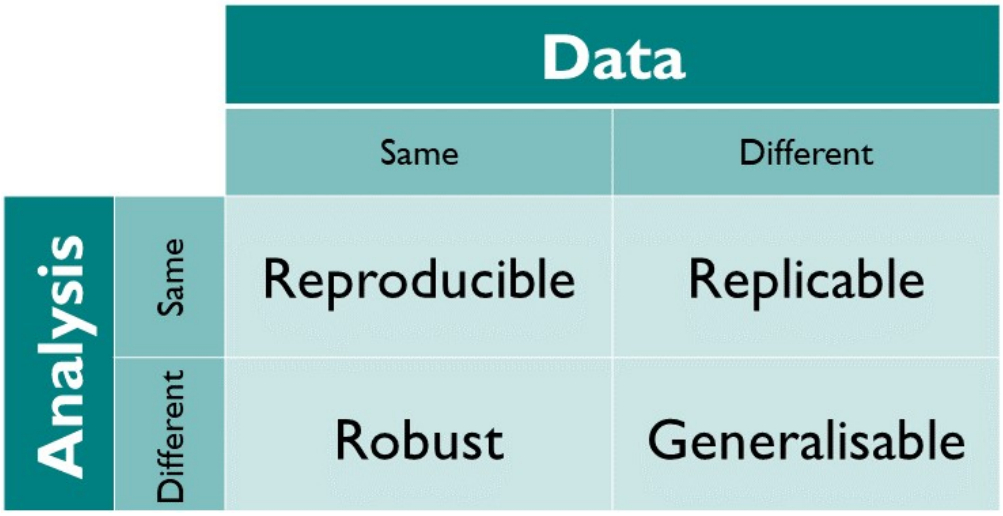


Figura 4.1: Definiciones de reproducibilidad, replicabilidad, robustez y generalizado. Figura, tomada de la sección *Definitions* el libro de *Turing Way* (Arnold et al., 2019).

4.1 Metodología

Se seleccionaron experimentos que 1) estuvieran en PulmonDB, 2) que las muestras fueran de tejido pulmonar, 3) que tuvieran grupos de controles y pacientes con EPOC. Estos experimentos seleccionados fueron: **GSE1122, GSE1650, GSE27597, GSE37768,GSE47460, GSE57148, y GSE8581** (Véase Tabla 4.1).

Se siguió el flujo de trabajo clásico para microarreglos, en donde se usa *Robust Multi-array Average* (RMA) para normalizar los datos crudos, los cuales se obtuvieron usando la función de `getGEOSuppFiles` de la paquetería de `GEOquery`.

La normalización es el proceso que permite comparar las intensidades de las sondas del microarreglos entre las diferentes muestras. El algoritmo de RMA se usa en los datos de *Affymetrix* y tiene los siguientes principios: Asume que las intensidades siguen una misma distribución, ajusta las intensidades utilizando un modelo de fondo, normaliza los datos mediante cuantiles en donde se obtiene la misma distribución en todos los microarreglos, transforma las intensidades en logaritmos de base 2, y finalmente se condensa la expresión de las sondas utilizando *median polish*, dando como resultado la expresión del gen o de los grupos de sondas (Irizarry et al., 2003). El algoritmo de *median polish* es robusto y no se ve afectado por los valores extremos, tiene como finalidad normalizar los datos utilizando tanto

Tabla 4.1: Información de los experimentos usados. EPOC: Enfermedad Pulmonar Obstructiva Crónica; AAD: Deficiencia de alfa-1-antitripsina; ILD: Enfermedades intersticiales

Experimento	Muestras	Plataforma	Año	Categoría
GSE1122	15	Affymetrix	2004	Normal , Enfisema, AAD
GSE1650	30	Affymetrix	2004	Normal o medio Enfisema, severo Enfisema
GSE27597	72	Affymetrix	2011	Normal , EPOC y Enfisema
GSE37768	38	Affymetrix	2016	Normal , EPOC
GSE47460	582	Agilent, Agilent	2013	Normal , EPOC, ILD
GSE57148	189	Illumina	2015	Normal , EPOC
GSE63073	42	Rosetta	2014	Normal , EPOC
GSE8581	58	Affymetrix	2008	Normal , EPOC

columnas como filas. Es un procedimiento iterativo que busca acercar las medianas calculadas por filas y por columnas a cero con la finalidad de escalar los valores y está implementado como parte de RMA (Gimond, 2021, Irizarry et al. (2003)). Para utilizar RMA, se uso la función `rma` del paquete de `affy` (Gautier et al., 2004).

Para los microarreglos de `Agilent` se siguió el flujo de trabajo presentado en el paquete de `limma` (Ritchie et al., 2015). Este flujo inicia con la lectura de archivos usando la función `read.maimages`, y se calcula la expresión usando la mediana con el atributo `source = "agilent.median"`. Después se utilizó la función de `backgroundCorrect` para ajustar las intensidades siguiendo el mismo modelo de fondo que en RMA (Ritchie et al., 2007). Se normalizó los datos con cuantiles usando la función `normalizeBetweenArrays` y finalmente se transformaron los valores en logaritmo base 2.

El experimento `GSE57148` son datos de RNA-seq, y el número de *lecturas* por transcrito, se tomó de `recount2` siguiendo su *vignette*. Los datos de *Recount2* han sido previamente procesados usando el flujo de trabajo de Rail-RNA para alinear las *lecturas* (Collado-Torres et al., 2017).

Para obtener los datos de PulmonDB, se usó el Paquete de R (*PulumonDB*)[https://github.com/AnaBVA/PulmonDB]. Estos datos ya se encuentran normalizados utilizando cuantiles y en contrastes de muestras.

Después usando `limma` (Ritchie et al., 2015), se realizó por separado el análisis de expresión diferencial para cada experimento. Se comparó EPOC contra controles sin incluir variables adicionales, esto tanto para los datos normalizados utilizando el flujo de trabajo típico como usando los datos de PulmonDB.

Con los resultados de cada experimento, se graficaron los cambios de incremento en las muestras de EPOC (o *FC* por sus siglas en inglés de *fold-change*) en logaritmo base 2 (*logFC*) y se calculó su correlación para determinar la similitud de los resultados.

Estos pasos son representados en la figura 4.2.

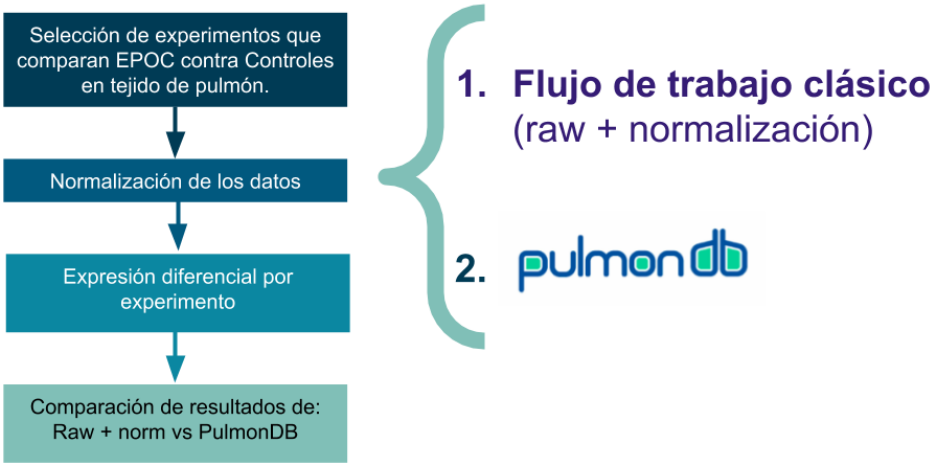


Figura 4.2: Diagrama de la metodología usada para explorar la robustez de PulmonDB. Como primer paso fue la selección de experimentos que contrastan muestras de EPOC contra muestras controles en tejido pulmonar. Después la obtención de datos normalizados, para seguir 1) el flujo de trabajo clásico se usaron los datos crudos y se normalizaron utilizando RMA, 2) los datos normalizados de PulmonDB se obtuvieron usando su paquetería de R. Luego se realizó la expresión diferencial por experimento y después se compararon los datos obtenidos de RMA con los que se obtuvieron de PulmonDB.

4.2 Resultados

El diagrama de dispersión nos ayuda a describir el comportamiento de dos variables. Cuando ambas variables no presentan ninguna relación se observan como una nube de puntos, mientras que una linea recta indica una relación lineal(Taylor, 1990). El coeficiente de

correlación va de `1` a `-1`, indicando alta correlación cuando el coeficiente se acerca al valor de `1`, si el valor es negativo indica una correlación negativa (Taylor, 1990).

El diagrama de dispersión se realizó para cada experimento que se seleccionó, los análisis se realizaron de manera separada para cada uno y la correlación se calculó usando los datos de *logFC* comparando los resultados normalizados de la manera típica contra los resultados de PulmonDB. Como ejemplo se presenta el gráfico del experimento **GSE27597**, el cual tuvo un coeficiente de correlación de `R = 0.95`. En el gráfico cada gen se representa con un punto y se observa como los genes que se encuentran sobre expresados en los resultados analizados con el flujo de trabajo clásico, también se encuentran sobre expresados en PulmonDB y viceversa para los genes subexpresados. La cual indica una gran similitud entre los resultados obtenidos de PulmonDB y los resultados curdos normalizados, como se observa en la Figura 4.3.

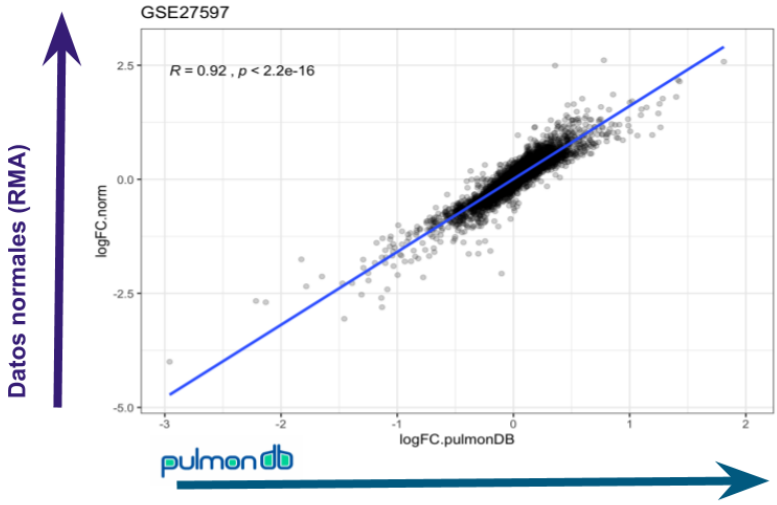


Figura 4.3: Diagrama de dispersión entre los datos normalizados utilizando el flujo de trabajo típico y los contrastes de muestras normalizados en PulmonDB para la muestra GSE27597. En el eje X se observan los resultados de *logFC* de PulmonDB, en el eje Y los resultados de *logFC* usanndo el flujo de trabajo típico. En color azul se encuentra la regresión linear y el coeficiente de correlación se indica en la esquina superior izquierda.

Para todos los experimentos la correlación fue aceptable, en promedio el coeficiente de correlación fue de `R = 0.808`. La más baja fue de `R = 0.61`, que se obtuvo para el experimento **GSE8581** y la más alta fue de `R = 0.95` para el experimento **GSE47469**. En la Figura 4.4, se observan los gráficos de dispersión para cada experimento como se mostró

anteriormente para la Figura 4.3.

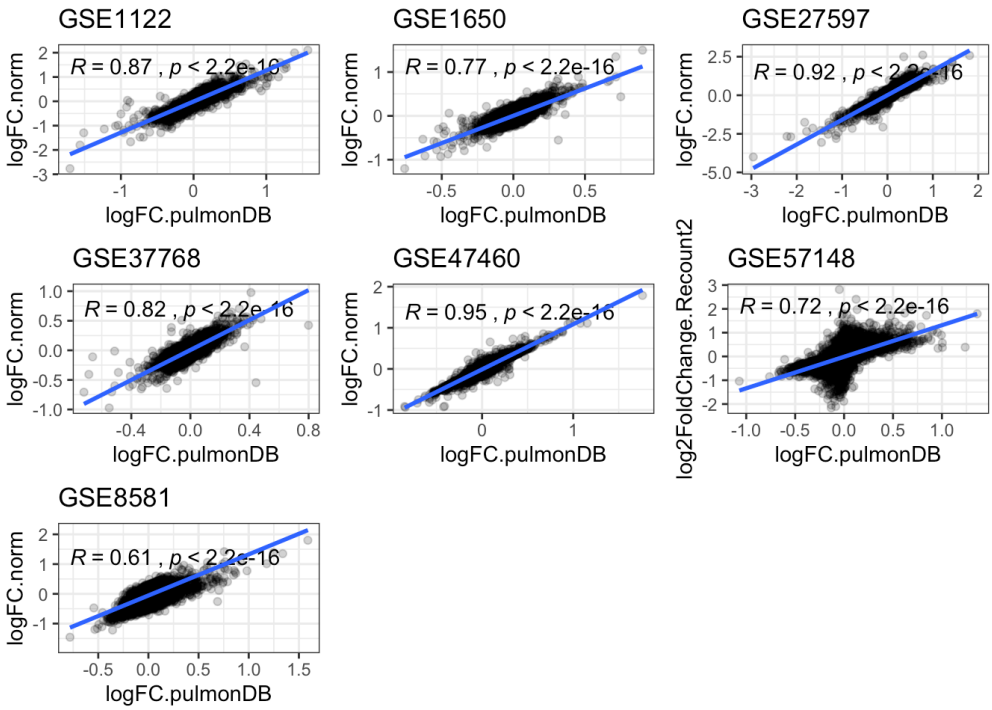


Figura 4.4: Diagrama de dispersión para todos los experimentos selecionados. El eje X tiene los resultados de *logFC* para PulmonDB, el eje Y los resultados de *logFC* obtenidos usando el flujo de trabajo clásico. La línea azul representa la regresión lineal.

La correlación entre el *logFC* de PulmonDB y los datos pre-procesados usando el flujo de trabajo clásico, muestran como los contrastes individuales preservan una estructura similar en los datos. Cuando los genes se reportan sobreexpresados usando el flujo de trabajo clásico, también se encuentran sobreexpresados en PulmonDB, y viceversa. Lo que demuestra que los datos de PulmonDB son robustos, capaces de regresar resultados similares usando contrastes individuales.

4.3 Conclusión

Los datos de PulmonDB son contrastes individuales, los cuales se calcularon contrastando una muestra usada como referencia contra todas las muestras restantes del experimento. Esto permite tener una valor relativo que indica el cambio de la expresión el cual ha sido normalizado usando cuantiles en cada *contraste individual*.

La normalización en PulmonDB se hace con base en cuantiles, de manera similar que el algoritmo de RMA, sin embargo en PulmonDB no se realiza una corrección de fondo por lo que era importante determinar la robustez de los datos en PulmonDB.

Utilizando gráficas de distribución y coeficientes de correlación, se pudo comprobar que los resultados de PulmonDB se comportan de manera similar a los resultados que se pueden obtener con el flujo de trabajo clásico para los experimentos previamente seleccionados.

Capítulo 5

Heterogeneidad en EPOC y FPI

5.1 Introducción

El pulmón es un tejido complejo, conformado por diferentes estructuras entre las cuales resaltan los bronquios, bronquiolos y alveolos. Cada estructura tiene diferente composición celular, las **vías respiratorias** (*airways*) se caracterizan por tener células basales, secretorias (*club cells*), ciliadas, *goblet, tuft*, ionocitos, serosas, mucosasas, entre otras (Travaglini et al., 2020). Mientras que los **alvéolos** están compuestos principalmente por células alveolares tipo I, tipo II, fibroblastos (en el intersticio pulmonar), macrófagos, etc (Zepp and Morrisey, 2019, Franks et al. (2008), Travaglini et al. (2020)) (Vease 5.1).

Adicionalmente, el pulmón es un tejido altamente conectado por **capilares vasculares** por los cuales el oxígeno es llevado a todos los tejido del organismo. La vascularización pulmonar también aporta heterogeneidad en el pulmón, sobretodo cuando se obtienen muestras para los análisis transcriptómicos (Zepp and Morrisey, 2019; Franks et al., 2008, Travaglini et al. (2020)). Por lo que en este capítulo se calculará la composición celular de muestras de la EPOC (GSE57148) y FPI (GSE52463). Esto nos permitirá determinar tipos celulares clave para las enfermedades y observar la composición por muestra, la cual podría estar implicada en la heterogeneidad de los datos.

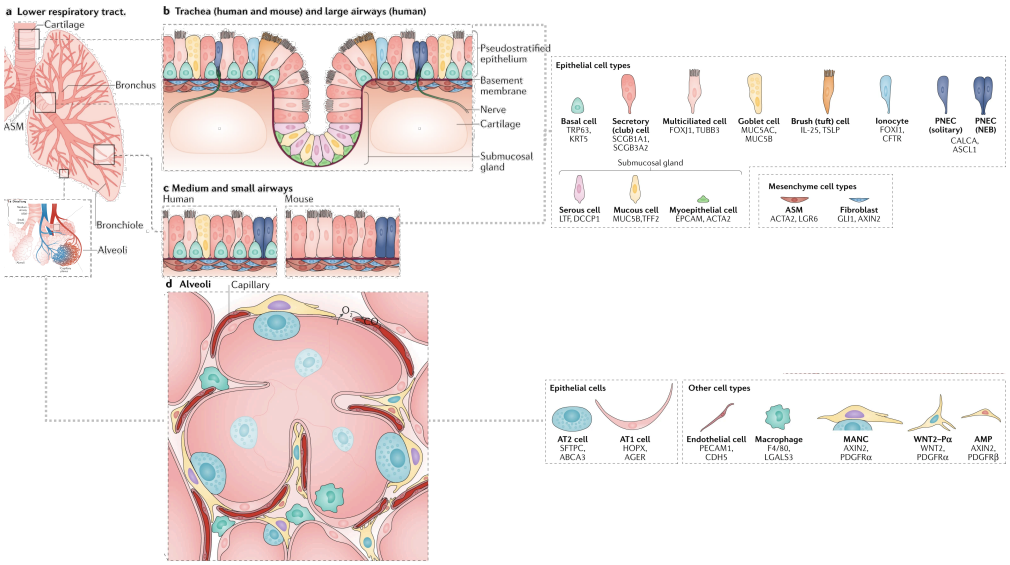


Figura 5.1: Tipos celulares en el pulmón, modificada de (Zepp and Morrisey, 2019). a) representación del pulmón que se encuentra dividido en b) Tráquea y vías aéreas largas (Bronquios), c) Vías aéreas medias y bajas (Bronquiolos) y d) alvéolos, cada una de las regiones con diferente composición celular.

5.1.1 Tecnologías de secuenciación

Cuando se mide la expresión de RNA por secuenciación o microarreglos normalmente lo que se mide es una expresión promedio de todas las células que fueron capturadas en esa muestra. En el capítulo **??** se discute a detalle cada tecnología, su uso y el flujo de trabajo general para cada una.

En general, si la muestra se toma de tejido, la expresión que se obtiene es un promedio de todas las células que se encuentran en dicho tejido. Por ejemplo, de una muestra tomada del pulmón se obtiene la expresión de células alveolares, macrófagos, células epiteliales, células ciliadas, células *goblet* etc. Lo mismo sucede para las muestras de sangre o de cualquier otro tipo, a esto se le conoce como *"bulk"* (Véase capítulo 5.2).

Actualmente existen tecnologías que nos permiten medir la expresión de una sola célula, a esto se le llama *single-cell RNA-seq* (*scRNA-seq*) para conocer más sobre está tecnología se puede consultar la revisión que se adjuntó en el capítulo **??**. Los datos de *scRNA-seq* se pueden usar para calcular la proporción de células que componen las muestras de secuenciación *bulk*, a estos algoritmos se le conoce como de *deconvolución*.

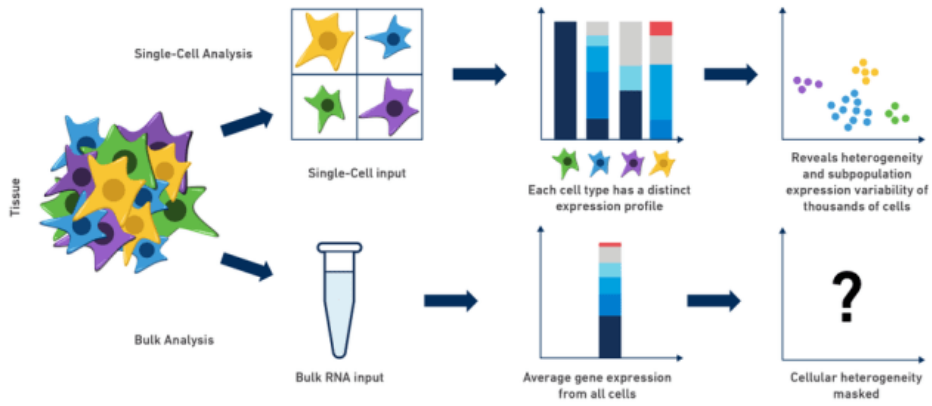


Figura 5.2: Figura de secuenciación *bulk* contra *single-cell* obtenida de (Sheila-10x, 2017). Se representa un tejido compuesto de diversas células, las cuales son analizadas con *single-cell* (arriba) o *bulk* (abajo) y como la expresión genética obtenida de ambas tecnologías representa la expresion de un tipo celular o un promedio de todo.

5.1.2 Métodos de deconvolución

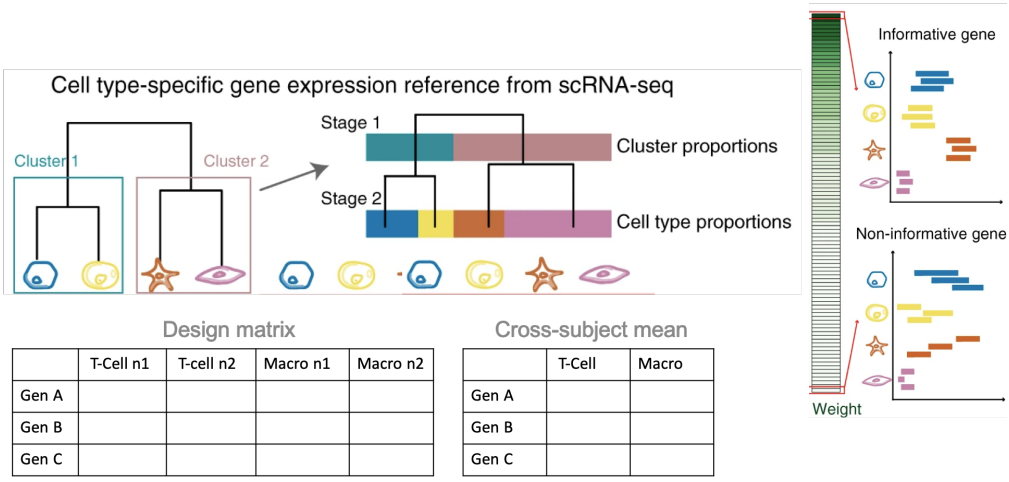
Existen métodos que permiten inferir la proporción de células en datos de *bulk RNA-seq* a partir de *scRNA-seq*, a estos métodos se les conoce como deconvolución. *Cobos, et al.* diferencian dos tipos de métodos de deconvolución, aquellos que requieren un panel de referencia *scRNA-seq* y aquellos que no lo requiere (Cobos et al., 2020).

De los métodos que no requieren panel de referencia *scRNA-seq* se encuentra CIBERSORT, el cual predice la composición celular usando el perfil de expresión de cada tipo celular ocupando regresión de máquinas de vectores de soporte lineales (o *lineal support vector regression*) (Newman et al., 2015).

Otro de los métodos disponibles de deconvolución es *MUlti-Subject SIngle Cell* (MuSiC) (Wang et al., 2019) para calcular la composición celular de muestras evaluadas con *RNA-seq* para EPOC y FPI. MuSiC, entra dentro de los métodos que ocupan panel de referencia *scRNA-seq* para estimar las proporciones en *bulk RNA-seq* identificando genes marcadores de los diferentes tipos celulares (tanto entre individuos como entre células). A estos genes informativos se les da un peso para priorizar e inferir la similitud de las células usando un árbol basado en agrupamiento jerárquico, después se calcula la proporción usando mínimos cuadrados no negativos ponderados (*weighted least-squares regression o W-NNLS*) y esto se

repite varias veces (Wang et al., 2019) (Véase la figura 5.3).

MuSiC está dentro de los mejores métodos de deconvolución que usan datos de *scRNA-seq* para calcular las proporciones de datos *bulk*, tanto en las evaluaciones realizadas por los creadores de MuSiC (Wang et al., 2019) como por otros autores recientemente (Cobos et al., 2020). Sin embargo, es importante considerar que una limitante de MuSiC es que todos los tipos celulares deben encontrarse en la matriz de referencia *scRNA-seq*, si el tipo celular no se encuentra en los datos *scRNA-seq* de referencia, no calculará de manera correcta la proporción celular (Wang et al., 2019). Dado su buen funcionamiento, decidimos utilizar MuSiC para calcular la composición celular de muestras evaluadas con RNA-seq para EPOC y FPI.



	T-Cell n1	T-cell n2	Macro n1	Macro n2
Gen A				
Gen B				
Gen C				

	T-Cell	Macro
Gen A		
Gen B		
Gen C		

Figura 5.3: Figura general de MuSiC modificada de (Wang et al., 2019). El primer paso es calcular similitudes usando agrupamiento jerárquico (clustering jerárquico). Después se obtienen las proporciones priorizando los genes con mayor peso. Esto se hace iterativamente hasta obtener la proporción de todos los tipos celulares.

5.2 Metodología

En este análisis se utilizó el paquete MuSiC (Wang et al., 2019), siguiendo su tutorial y descargando su paquetería de R `MuSiC`.

MuSiC require 1) datos de *scRNA-seq* y 2) datos *bulk* de RNA-seq. Para los datos de *scRNA-seq* se usó el experimento de (Reyfman et al., 2019) publicado en GEO con el ID

Tabla 5.1: Datos de bulk RNA-seq

GSE	Enfermedad	SRA	Muestras	Fecha	Pais
GSE52463	IPF	SRP033095	15	Nov 18, 2013	EUA
GSE57148	COPD	SRP041538	189	Abril 28, 2014	Sur Corea

GSE122960, el cual contiene 8 muestras de pulmón tomadas de donadores sanos y 8 muestras de pulmón de pacientes con FPI. Sin embargo para la deconvolución solo solo se seleccionaron las muestras de los donadores sanos.

Los datos seleccionados para esté análisis fueron los **GSE52463** y **GSE57148** los cuales se encuentran descritos en la Tabla 5.1. Estos datos fueron seleccionados por ser datos de *RNA-seq* y experimentos con muestras de EPOC y FPI contra sanos. El número de *lecturas* se obtuvieron de `recount2` (Collado-Torres et al., 2017) siguiendo su *vignette*.

Para definir la similitud en los tipos celulares se usó la función `music_basis`, agrupando por tipos ceulares obtenidos en el panel de referencia. Los cuales fueron estimados usando los genes por tipo celular (*design matrix*) o la abundancia relativa del promedio entre individuos (*mean of RA*). El árbol de dendograma se calculó con distancia euclidiana y e usó un corte en el árbol, que daba como resultado 7 grupos asociados a los tipos celulares.

Para estimar las proporciones se usó `music_prop`, el panel de *scRNA-seq* con los 7 grupos previamente calculados y los datos de *bulk* RNA-seq tanto para la EPOC como para la FPI, el cálculo se realizó por separado para cada enfermedad. Se calcularon los cambios estadísticos con una prueba no paramétrica tipo *Wilcoxon* por tipo celular. También se calculó la varianza por tipo celular y se graficó por grupo (control y FPI/EPOC) utilizando la paquetería de R `ggpubr` (Kassambara, 2018), y se calculó su significancia con una prueba estadística no paramétrica tipo *Wilcoxon*.

Esta parte del trabajo fue realizada en colaboración con el Dr. **Lukas Simon** quien se encontraba en el *Institute of Computational Biology, Helmholtz Zentrum München*. El subconjunto de datos, así como la asignación del tipo celular fue proporcionada por el Dr. Lukas Simon.

5.3 Resultados

5.3.1 Definición de *clusters*

Como se mencionó anteriormente, MuSiC trabaja con árboles usando agrupamiento jerárquico para definir similitud en tipos celulares. Como primer paso se graficó el dendograma que muestra similitud en los tipos de células. A partir de este árbol de similitud se seleccionaron siete grupos (*clusters*) los cuales son:

- C1 = 'Lymphatic' (azul)
- C2 = 'Fibroblast' (morado)
- C3 = 'NK cell' (verde)
- C4 = 'Endothelium' (rojo)
- C5 = 'Ciliated' (aqua)
- C6 = 'T cell' and 'Mast cell' (rosa)
- C7 = 'Type 1', 'Secretory', 'Transformed epithelium', 'Type 2','Macrophages' and 'B cell' (amarillo-verde).

En la Figura 5.4, se observan los dos árboles jerárquicos que se obtuvieron, se decidió cortar el árbol para obtener 7 grupos los cuales se indican de color. El grupo 7 (C7), está enriquecido con células que conforman el alveolo, las cuales son células epiteliales y células inmunes (Macrófagos y células B). De manera notable, las células T y los mastocitos tuvieron más similitud y se agruparon juntas en el grupo 6 (C6). Mientras células del mismo linaje no tuvieron tanta similitud como las células T, B y NK que comparten el mismo progentor (progenitor común linfoide o *common lymphoid progenitor*) (Kitamura and Ito, 2005).

5.4 Composición celular de las muestras en IPF

Utilizando MuSiC (Wang et al., 2019) se calcularon las proporciones celulares en las muestras de *bulk* RNA-seq. La Figura 5.5 muestra los resultados obtenido para los datos de la FPI, en

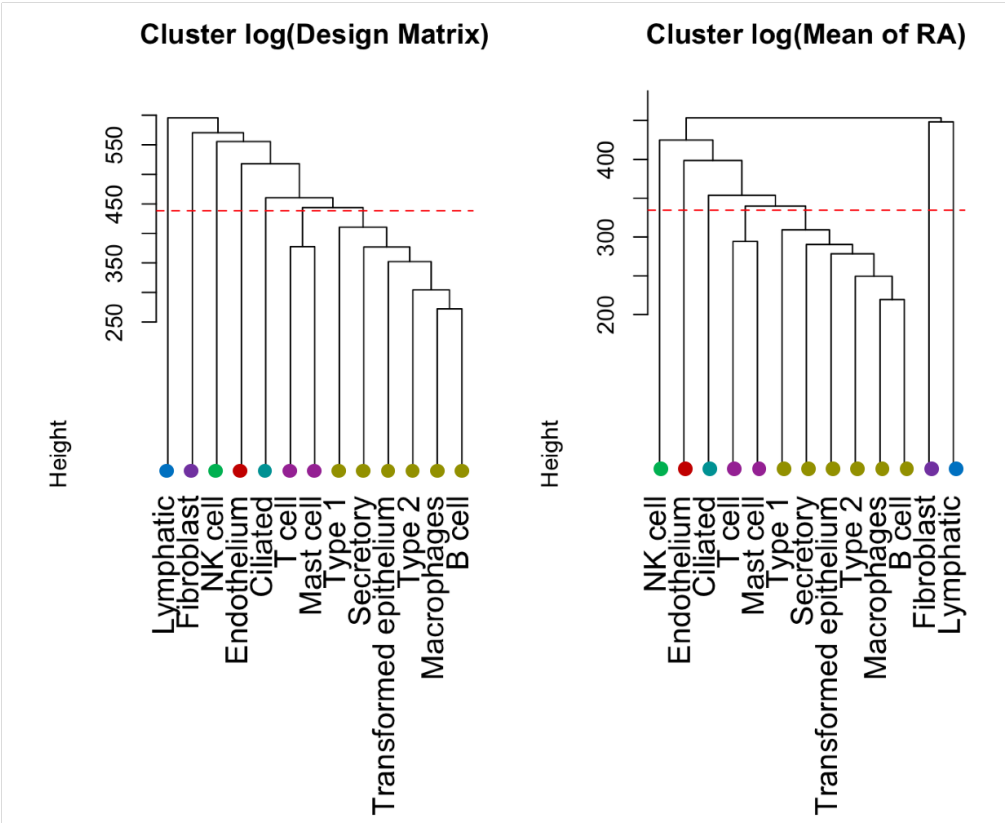


Figura 5.4: Árbol con agrupamiento jerárquico. Se calculó similitud entre tipos celulares para la matriz completa (*Cluster log (Design Matrix)*) o para la matriz del promedio entre individuos (*Cluster log(Mean of RA)*).Los 7 clusters o grupos (línea punteada roja) están indicados por colores: C1: *Lymphatic* (azul), C2: *Fibroblast* (morado), C3: *NK cell* (verde), C4: *Endothelium* (rojo), C5: *Ciliated* (aqua), C6: *T cell* y *Mast cell* (rosa), C7: *Type 1*, *Secretory*, *Transformed epithelium*, *Type 2*, *Macrophages* y *B cell* (amarillo-verde).

las columnas se encuentran los tipos celulares y cada fila tiene los resultados de para cada paciente (indicado en verde muestras de FPI y en ocre los controles).

Las muestras de los pacientes fueron agrupados jerarquicamente, dividiendo las muestras en dos grupos (indicados por el número uno y dos en la Figura 5.5). El grupo 1 tiene muestras controles (con excepción de una muestra de FPI), mientras que el grupo 2 tiene mayor enriquecimiento en pacientes con FPI.

Para medir los cambios en proporciones celulares, hicimos una prueba estadística no paramétrica (*Wilcoxon test*) por tipo celular (Véase Figura 5.6). IPF muestra cambios significativos con un p-valor menor a 0.05 en células alveolares tipo 1 y tipo 2 y fibroblastos. Mostrando una disminución en células alveolares tipo 1 y tipo 2 y un enriquecimiento en fibroblastos caracteristico de está enfermedad (Ramos et al., 2001). Esto también se ha observado en análisis anteriores en ratones (Xie et al., 2018) y en humanos (Adams et al., 2020).

A pesar que las células endoteliales no pasaron el corte de p-valor menor a 0.05, en estudios anteriores, *Adams et al* reportaron un perfil transcripcional ectópico muy similar a las células endoteliales vasculares (VE) en los bronquios, y vías respiratorias de pacientes con FPI. En el estroma, conformado por la pleura y tejido intersticial, los autores encontraron fibroblastos invasivos y miofibroblastos aberrantes cuando se comparó con controles (Adams et al., 2020) como se observa también en nuestros resultados.

Las células epiteliales alveolares de tipo II (AT-II) tienen una proporción reducida en las muestras de FPI (Figura 5.6). Lo cual correlaciona con revisiones anteriores, en donde se observa apoptosis en células epiteliales (Uhal, 2008; Barbas-Filho et al., 2001). El incremento en apoptosis de células AT-II, así como la activación de los miofibroblastos son considerados como agentes clave para la iniciación de FPI (Uhal, 2008).

Se sabe que las células AT-II dañadas en combinación con otros factores ambientales y genéticos desencadenan la enfermedad para luego comenzar un proceso repetitivo de curación-lesión. Estos daños ocasionan la activación del sistema inmune como macrófagos y neutrófilos, así como la activación de los miofibroblasotos, produciendo la proliferación de

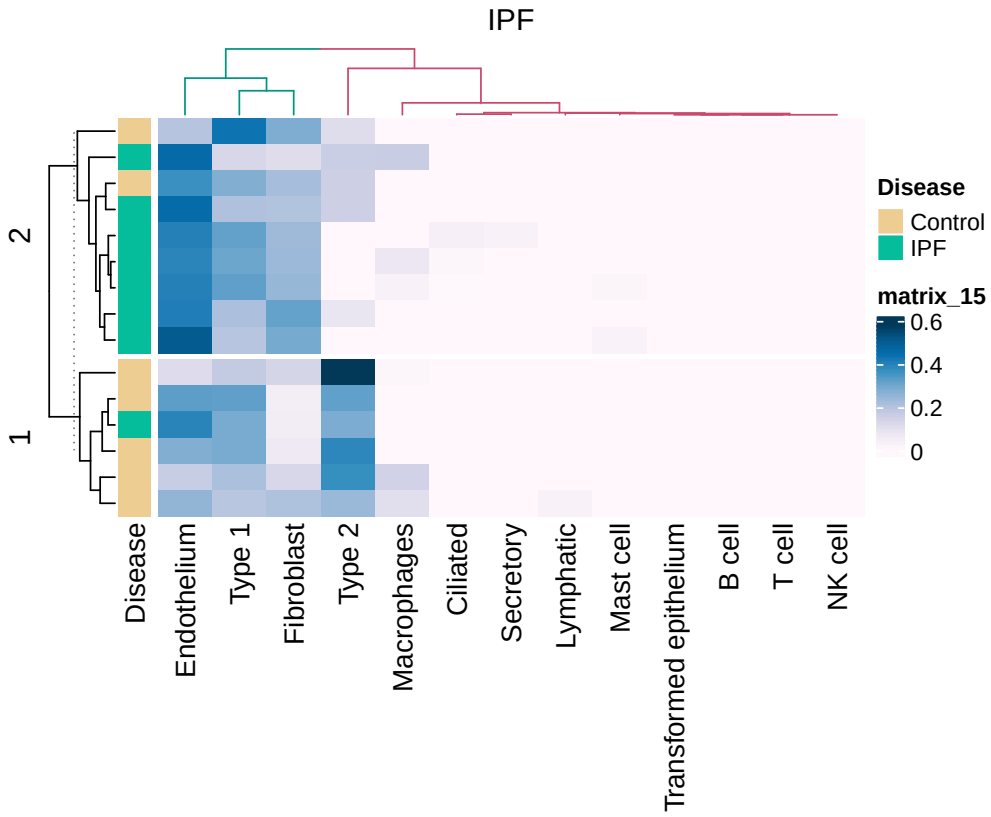


Figura 5.5: Proporción de tipos celulares en FPI obtenidos por deconvolución. Porcentajes de tipos célulares para el experimento **GSE52463**, el cual tiene muestras control (covariable en ocre) y de FPI (covariable en verde). El porcentaje está indicado en azul, entre más obscuro mayor proporción y más claro indica menor proporción. Tanto las muestras como los tipos celulares están agrupados jerárquicamente.

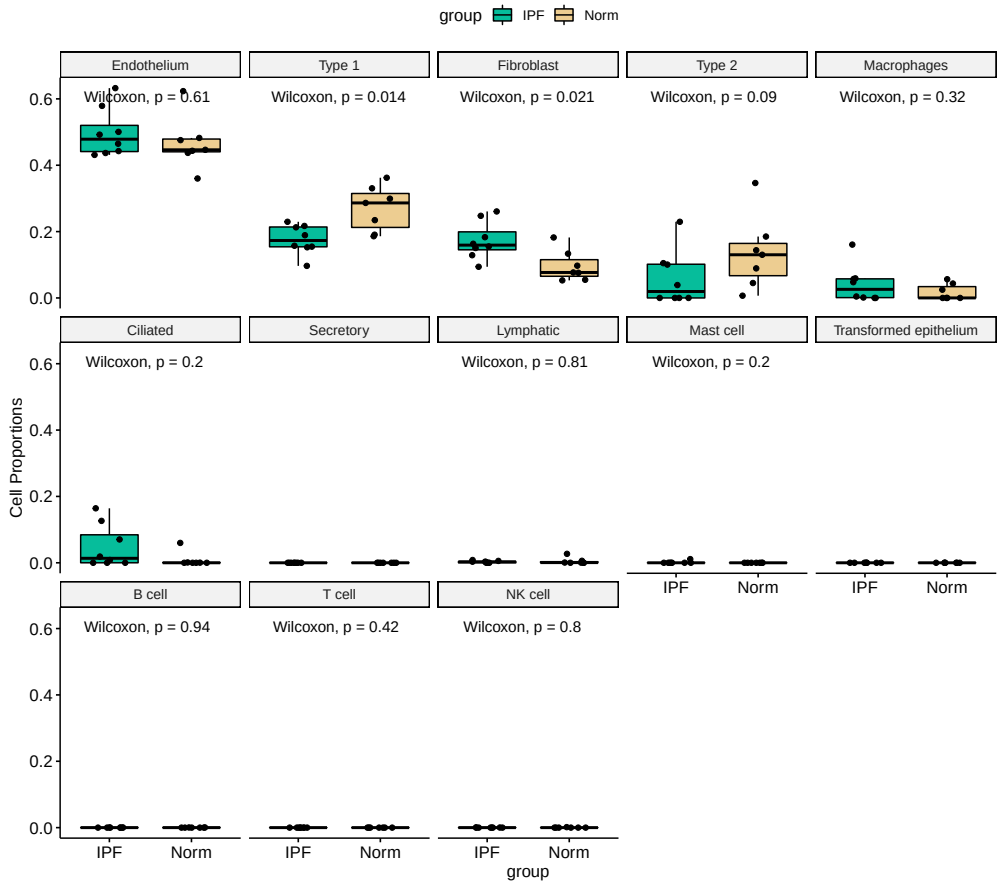


Figura 5.6: Comparaciones entre las proporción de tipos celulares en FPI obtenidos por deconvolución. El p-valor de la prueba Wilcoxon se muestra en cada gráfica. Verde es el grupo de pacientes con FPI y ocre los controles.

células epiteliales que terminan en apoptosis, resultando en la FPI (Camelo et al., 2014).

A demás de conocer la proporción celular y los cambios que existen en la enfermedad, nos interesó conocer si la proporción celular por paciente tiene varianza significativa que explicara cambios en la expresión transcripcional. Para ello se calculó la varianza por tipo celular y se graficó por grupos (Figura 5.7). El tipo celular con más varianza son las células AT-II y endotelio, sin embargo la prueba estadística no muestra un cambio significativo. Para un estudio más profundo se sugiere aumentar las muestras en un futuro, así como estudiar los cambios de proporción celular entre distintos experimentos.

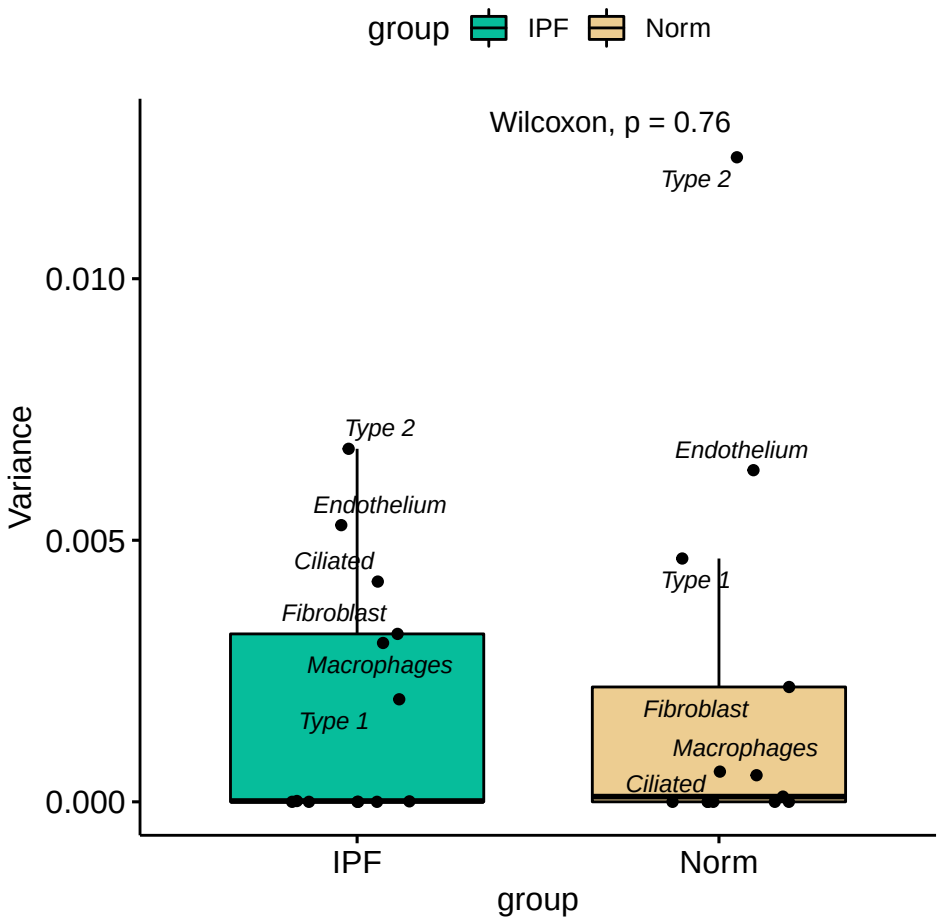


Figura 5.7: Varianza de las proporciones celulares para los datos de la FPI. La varianza entre pacientes calculada por tipo celular y mostrada por grupo. El p-valor de la prueba Wilcoxon se muestra en la gráfica. Verde es el grupo de pacientes con FPI y ocre los controles.

5.5 Composición celular de las muestras en EPOC

Similarmente como se realizó para FPI, se realizaron los análisis para la EPOC. En la Figura 5.8, se puede observar las proporciones calculadas para las muestras de la EPOC. El experimento para la EPOC **GSE57148**, tiene 189 muestras, a comparación del experimento que se usó para la FPI (n = 15). En la agrupación jerárquica de las muestras, la Figura 5.8 muestra dos grupos principalmente, el primero tiene se observa con mayor cantidad de muestras control, mientras que el 2 grupo tiene más muestras de la EPOC.

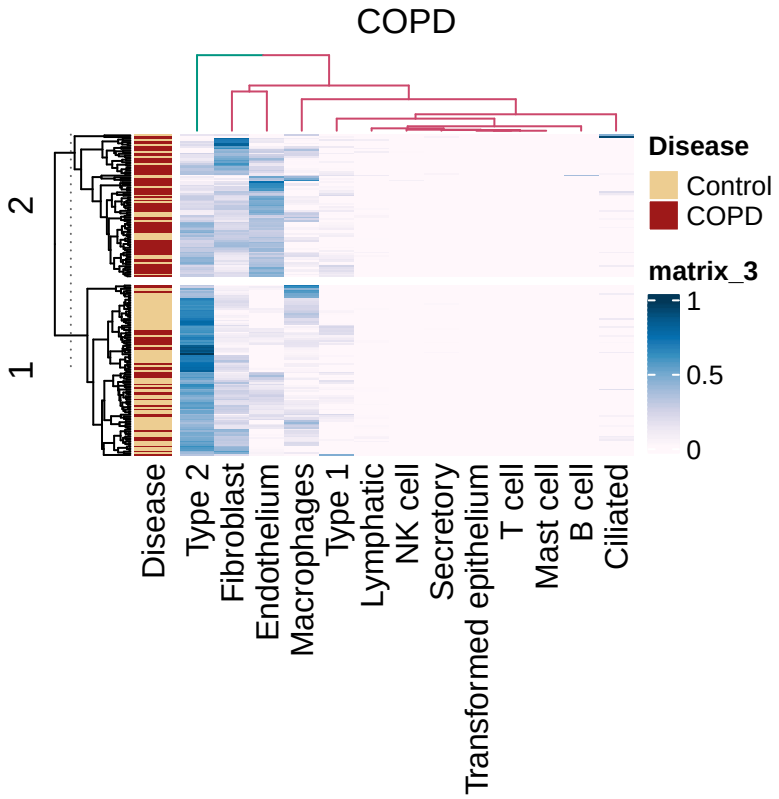


Figura 5.8: Proporción de tipos celulares en EPOC obtenidos por deconvolución. Porcentajes de tipos celulares para el experimento **GSE57148**, el cual tiene muestras control (covariable en ocre) y de la EPOC (covariable en rojo). El porcentaje está indicado en azul, entre más obscuro mayor proporción y más claro indica menor proporción. Tanto las muestras como los tipos celulares están agrupados jerárquicamente.

En la deconvolución, las células epiteliales AT-II y los macrófagos muestran una ligera reducción en las muestras con EPOC con cambios significativos de p-valor menor a 0.05 (Figura 5.9). Sin embargo, las células del endotelio aumentaron sus proporciones en

pacientes con la enfermedad en comparación con los controles con cambios estadísticamente significativos (Figura 5.9).

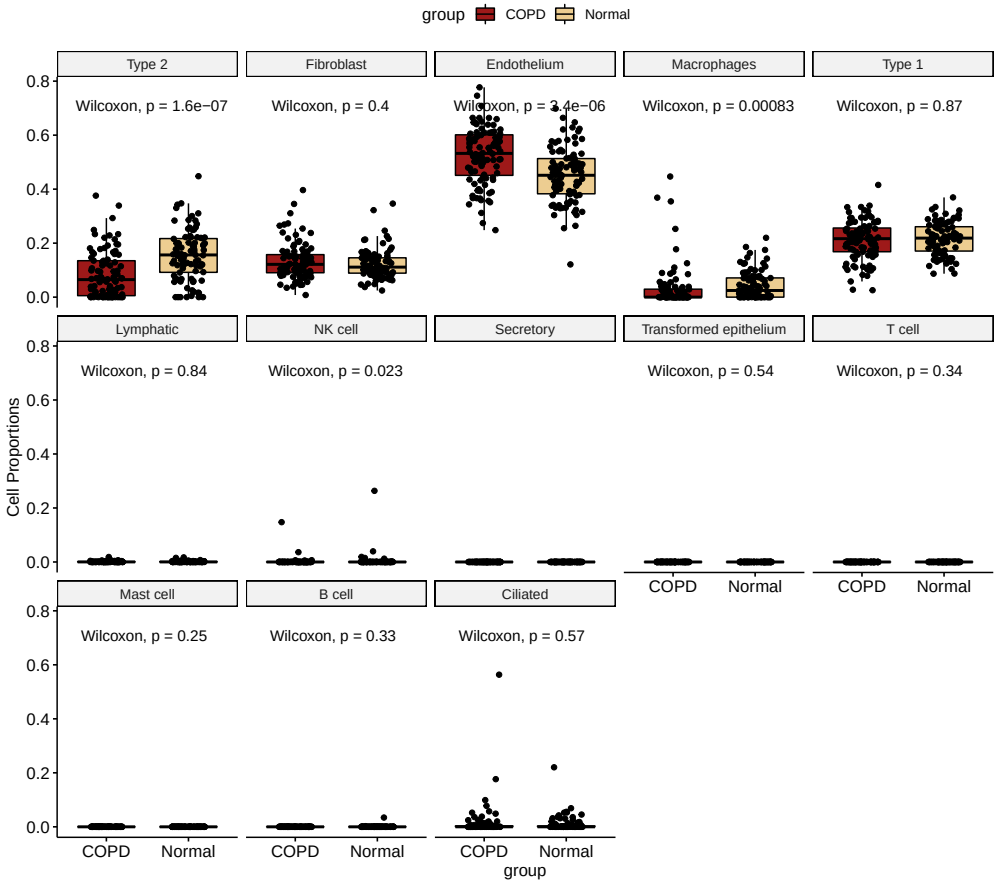


Figura 5.9: Comparaciones entre las proporción de tipos celulares en COPD obtenidos por deconvolución. El p-valor de la prueba Wilcoxon se muestra en cada gráfica. Verde es el grupo de pacientes con FPI y ocre los controles.

La EPOC se caracteriza por enfisema y bronquitis. Se desconoce la causa exacta, pero el cigarrillo, el humo de leña, la edad, la predisposición genética, entre otros, se han relacionado con EPOC, que conduce en una respuesta inflamatoria exacerbada. Como consecuencia, los neutrófilos y macrófagos son reclutados y secretan cantidades excesivas de proteasas, generando un patrón de destrucción alveolar (De Rose et al., 2018). Además, la apoptosis mediada por mayores cantidades de especies reactivas de oxígeno (ROS) que desencadena señales de daño oxidativo del DNA se ha relacionado con EPOC (Olajuyin et al., 2019). Lo que ayuda a explicar la disminución en AT-II.

Las células epiteliales AT-II son productoras de surfactante pulmonar, barrera de las vías

respiratorias, progenitor células de tipo I (AT-I) y encargadas de la autorrenovación en respuesta a lesiones (Olajuyin et al., 2019, Zepp and Morrisey (2019)). Por lo tanto, la ausencia de AT-II puede generar implicaciones importantes en la regeneración pulmonar después de una lesión, como es el caso de la EPOC.

Las células endoteliales muestran un enriquecimiento en las muestras de pacientes con EPOC (Figura 5.9), fenómeno que se ha reportado anteriormente y discutido sobre la relevancia del endotelio pulmonar en la EPOC. El endotelio pulmonar es clave para llevar acabo la transferencia de gases en la respiración, y las células endoteliales son las encargadas de formar la vasculatura. Durante el desarrollo de la EPOC, la migración transendotelial permite a células del sistema inmune como los neutrófilos migrar al pulmón a través de la unión con las células endoteliales, proceso que se encuentra elevado en EPOC (Green and Turner, 2017) por lo que nuestros resultados coinciden con la literatura.

En nuestros resultados los fibroblastos no mostraron cambios estadísticamente significativos (Figura 5.9). Sin embargo, en estudios anteriormente reportados muestran como los fibroblastos son clave para la reparación de lesiones, pero en la EPOC, las lesiones del tejido pulmonar no se reparan de manera adecuada, en parte por la reducción de la capacidad de los fibroblastos (Togo et al., 2008, Kulkarni et al. (2016)).

Para evaluar la variabilidad entre muestras, se realizó la medición de la varianza entre pacientes por tipo celular. Los resultados se graficaron en la Figura 5.10), mostrando variabilidad similar entre muestras control y pacientes con la EPOC. De manera similar a los resultados de FPI, el análisis podría mejorarse si se añadieran más experimentos para compara variabilidad de muestras entre diferentes estudios.

5.6 Conclusiones

La expresión transcriptómica de FPI muestra aumento en la proporción de fibroblastos y células epiteliales AT-I y AT-II. Para las muestras de la EPOC, existe un aumento de células endoteliales y una disminusión en macrófagos y células epiteliales AT-II, lo que recapitula a etiología de cada enfermedad. Sin embargo no se observaron cambios notorios en la variación

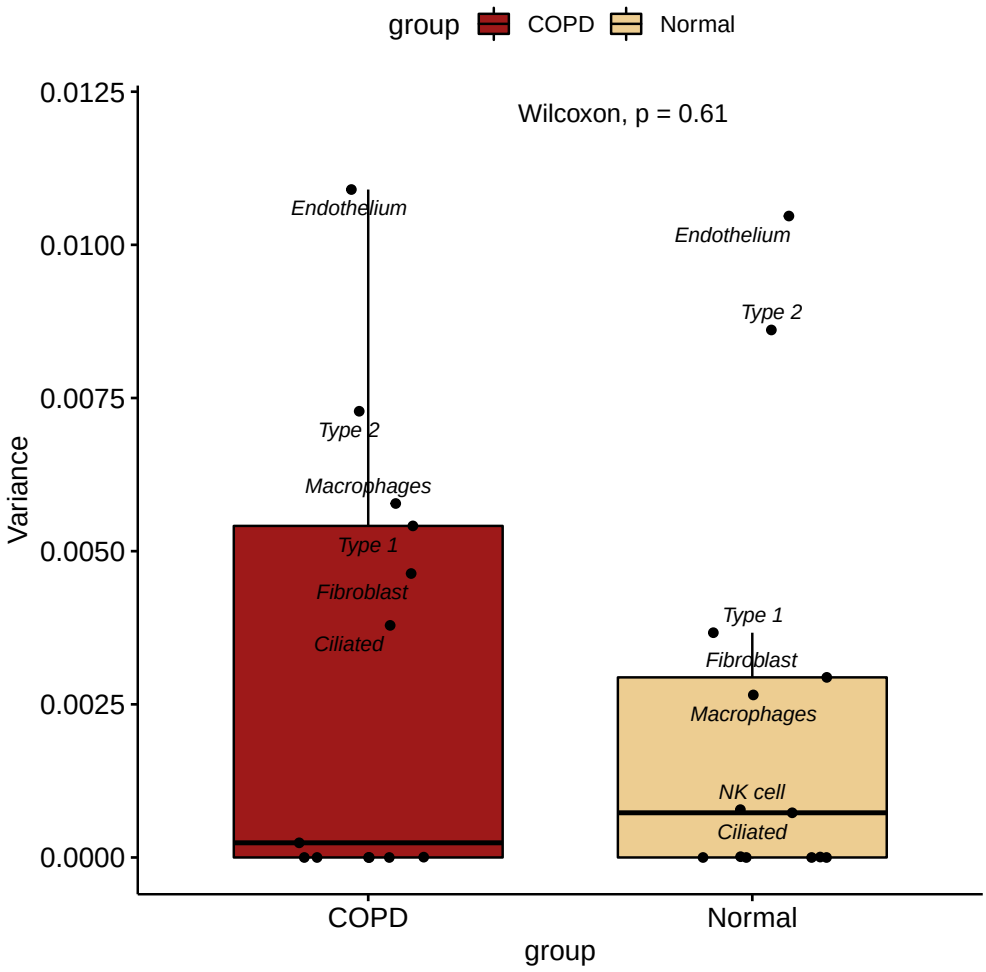


Figura 5.10: Varianza de las proporciones celulares para los datos de la EPOC. La varianza entre pacientes calculada por tipo celular y mostrada por grupo. El p-valor de la prueba Wilcoxon se muestra en la gráfica. Verde es el grupo de pacientes con FPI y ocre los controles.

por muestra cuando se analizó los experimentos de manera individual. Concluimos que la variabilidad en la expresión genética entre muestras del mismo estudio no sé puede explicar por la varianza entre proporciones celulares para los experimentos análizados. Se requiere un análisis que contemple varios experimentos de la misma enfermedad y se comparen las proporciones de las muestras para concluir si la proporción de células cambia entre experimentos.

Capítulo 6

Regulación de HHIP

6.1 Introducción

El gene *hedgehog interacting protein* (HHIP) tiene relevancia en EPOC debido a las variantes que se han encontrado en pacientes con esta enfermedad (Bártholo et al., 2019, Zhou et al. (2012), Zhou et al. (2012)). Como resultado del analysis de expression genética donde se utilizó PulmonDB, HHIP se encontró sobreexpresado en pacientes con EPOC. Por lo que se decidió estudiar la regulación de HHIP en EPOC

6.1.1 Vía de Hedgehog (HH)

La vía canónica de Hedgehog (HH) se activa cuando *soni hedgehog* (SHH), el ligando de *patched-1* (PTCH1), se une y promueve la activación de GLI3/2 (Briscoe and Thérond, 2013, di Magliano and Hebrok (2003)). Estos factores transcripcionales promueven la expresión de genes blanco, entre los que se encuentra HHIP. La proteína de HHIP regula negativamente la vía al competir con SHH por el ligando PTCH1. Adicionalmente, la vía de HH converge y participa con otras vías y factores de transcripción, como es el caso de JUN/API (Aberger and i Altaba, 2014).

La vía de HH se ha relacionado con EPOC, en donde se observa un fenotipo enfisematoso cuando se altera la vía HH en fibroblastos, los cuales afectan la estructura alveolar. Dicha

activación de la vía HH, cambia según la ubicación en el pulmón, presentando mayor activación en las vías aéreas y disminuyendo en los alveolos (Wang et al., 2018).

Anteriormente, se ha reportado la disminución de mRNA y de proteína en pacientes con EPOC en tejido pulmonar y la confirmación de polimorfismos de un solo nucleótido (*Single Nucleotide Polymorphism o SNP*) en regiones promotoras y *enhancers* cercanas a HHIP (Zhou et al., 2012). La variante rs1542725 está asociada a pacientes con EPOC y mostró tener mayor afinidad de pegado para el factor transcripcional Sp3, el cual se considera represor. Concluyendo que el incremento de afinidad para Sp3 promueve la represión transcripcional de HHIP (Zhou et al., 2012).

Diversos modelos murinos se han diseñado en donde se altera la vía de HH, ya sea alterando PTCH1, o HHIP, los cuales muestran un fenotipo enfisematoso. Sin embargo, el modelo de PTCH1 no presenta inflamación por lo que modela la enfermedad parcialmente (Tam et al., 2019). Mientras que el modelo murino haploinsuficiente de HHIP (+-) muestra enfisema espontánea e incremento de estrés oxidativo en los pulmones (Lao et al., 2016).

Estos modelos murinos no corresponden a lo observado en los datos de expresión en pacientes con EPOC. Se encontró sobreexpresión de HHIP en epitelio bronquial de pacientes con EPOC y una posible interacción microRNA-mRNA (Chang et al., 2018). Además, en un artículo recientemente publicado en medRxiv, se presentó el análisis de *scRNA-seq* en muestras de tejido pulmonar de 17 pacientes con EPOC y 15 donadores control. Los autores encontraron una población de células epiteliales AT-II que sobreexpresan HHIP (Sauler et al., 2020).

6.2 Metodología

6.2.1 Expresión de HHIP y componentes associados

Se utilizó el paquete de PulmonDB en R (Villaseñor-Altamirano et al., 2020) para descargar los datos de expresión en 8 estudios que evaluan la expresión genética en tejido pulmonar de grupos de EPOC contra controles (GSE1122, GSE1650 , GSE27597, GSE37768, GSE47460, GSE57148, GSE63073, GSE8581) con la función `genesPulmonDB`.

Obtuvimos los genes asociados a la vía de HH usando la base de datos de Kyoto Encycopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000). Se recuperaron 51 componentes y se descargó su expresión usando PulmonDB (Villaseñor-Altamirano et al., 2020). La expresión de los componentes de la vía de HH se graficaron en un heatmap.

6.2.2 Módulos de regulación

Se utilizó pySCENIC (Van de Sande et al., 2020) para calcular módulos de regulación utilizando la expresión de todos los genes compartidos (genes = 14,393) en los 8 estudios selecionados (muestras = 738). Los motivos usados para el análisis fueron los que la herramienta pySCENIC propone. Como resultado tuvimos 25,156 regulones, de los cuales se selecionaron 293 en los cuales HHIP esta presente. Los regulones tienen información sobre el factor transcripcional, el motivo del factor transcripcional y los genes blanco.

6.2.3 Variantes genéticas de la literatura

Este trabajo se hizo en colaboración con la estudiante de medicina de la Universidad Anáhuac México, *Mayra Padilla*, quien buscó variantes asociadas a EPOC en la literatura científica. Con la finalidad de encontrar variantes asociadas a EPOC que tengan efecto en la regulación de la expresión en HHIP. La búsqueda se realizó en Pubmed buscando el gen HHIP y la EPOC, se curaron manualmente las variantes que aparecen en los artículos científicos.

Adicionalmente se usaron variantes asociadas a EPOC o asociadas a la capacidad pulmonar (como FEV1). Las variantes se tomaron de la región reguladora de HHIP (1000 bp arriba del inicio de la transcripción). Para ello, se usó una base de datos creada por la estudiante de genómicas *Lucía Ramírez*, quien unió las variantes reportadas en GTEx, ClinVar, RegulomDB y GWAS (Figura 6.1).

6.2.4 Variantes regulatorias

Se utilizó `variation-scan` de RSAT (Santana-Garcia et al., 2019) para identificar variantes regulatorias que alteran los motivos de regulación en las regiones regulatorias (*e.i.* promotores). La base de motivos que se utilizó fue JASPAR de RSAT “core nonredundant

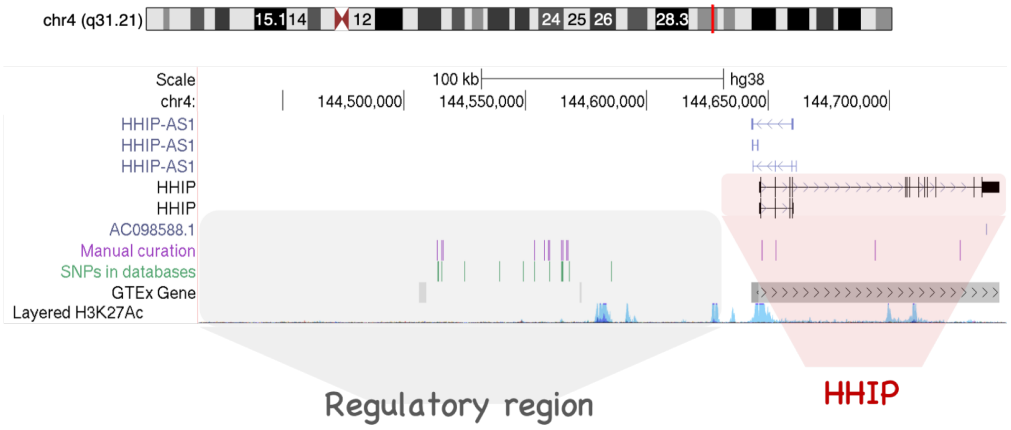


Figura 6.1: Descripción de HHIP y variantes asociadas a la EPOC.

vertebrates (2020)". Se tuvo como resultado 172 motivos únicos que se alteran con valores mayores a 10 en la relación de p-values. La relación de p-values es un valor que refleja los cambios en afinidad debido a la variante.

Se utilizaron los motivos resultantes de `variation-scan` para selecionar los regulones donde participa HHIP. Selecionamos las matrices que estuvieran presentes en los regulones de pySCENIC y en los motivos con cambio de afinidad por variantes asociadas a EPOC.

6.3 Resultados

En este trabajo se encontraron cambios en la expresión de HHIP. Específicamente, los pacientes con EPOC presentan mayor expresión de HHIP con respecto a los controles y los pacientes con FPI (Figura 6.2 A). La mayoría de los componentes asociados a la vía de HH no muestran cambios, sin embargo DHH y IHH muestran sobreexpresión en EPOC. La agrupación jerárquica de las muestras da como resultado grupos donde se observa mayor concentración de pacientes con FPI, otro grupo con controles, otro de pacientes con EPOC y oros grupos con mezcla de EPOC-FPI y EPOC-Controles (Figura 6.2 B).

Actualmente, con el desarrollo de tecnologías como *single-cell RNA-seq* (scRNA-seq) se puede obtener la expresión de los genes por tipo celular, mejorando la resolución. En un estudio de scRNA-seq con pacientes, *HHIP* muestra sobreexpresión en la EPOC para células dendríticas, alveolares, B, alveolares tipo I, basales, club, limpfáticas, fibroblastos y mesoteliales. Sin

embargo, células cilidadas, goblet, y capilares B muestran sobreexpresión en controles (Sauler et al., 2020, http://www.copdcellatlas.com/). Esto puede provocar diferencias en la expresión de *HHIP* de todo el tejido. Utilizar datos de scRNA-seq podría ayudar a mejorar nuestro entendimiento sobre *HHIP* y la vía HH en la EPOC.

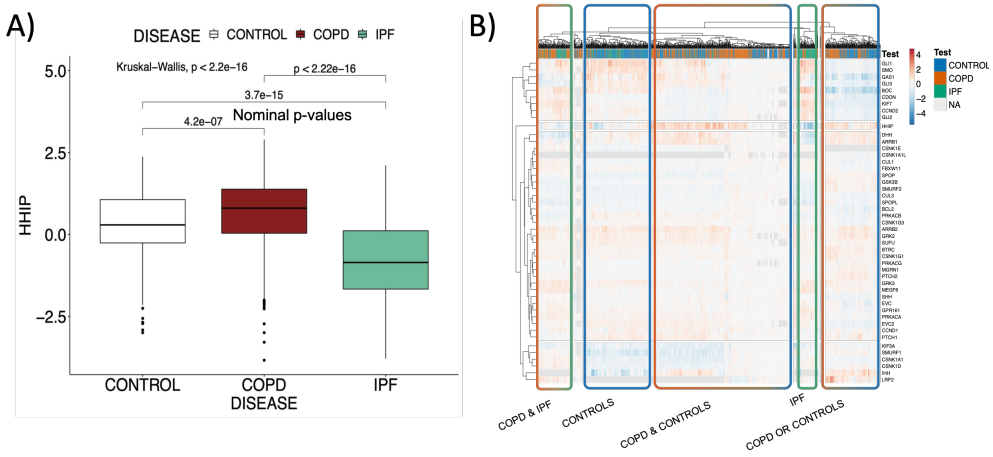


Figura 6.2: Expresión de HHIP en datos de PulmonDB. Muestras de tejido pulmonar de los 8 estudios selecionados se usaron para graficar la expresión de HHIP. Blanco representa a las muestras control, rojo a las muestras de EPOC y verde a muestras con FPI.

Para encontrar posibles reguladores de HHIP se utilizó la herramienta de pySENIC (Van de Sande et al., 2020), la cual agrupa genes con expresión similar y busca posibles factores transcripcionales que estén regulando dicho módulo de co-expresión. pySENIC predice los factores transcripcionales HLF, ATF3, IRF1, FOS, JUN, JUND, entre otros. Los cuales regulan módulos donde se encuentra HHIP (como ejemplo se muestran los regulones de *ATF3* y *HLF* Figura 6.3).

Se usaron las variantes asociadas a la EPOC y/o a sus fenotípos (ejemplo FEV1, FEV1/FVC%) curadas manualmente y las que se encontraron en bases de datos (Figura 6.1) para analizar cambios de afinidad en las matrices de pegado en los factores transcripcionales previamente identificados con pySCENIC, esto se realizó con la herramienta *Variation-scan* de RSAT (Santana-Garcia et al., 2019).

Las dos matrices de Jaspar que muestran cambio de afinidad por variantes asociadas a EPOC (calculado con *Variation-scan*) y que son reguladores de los regulones donde se encuentra HHIP (calculado con pySCENIC) son: *HLF* (MA0043) y *ATF3* (MA0018) (Figura

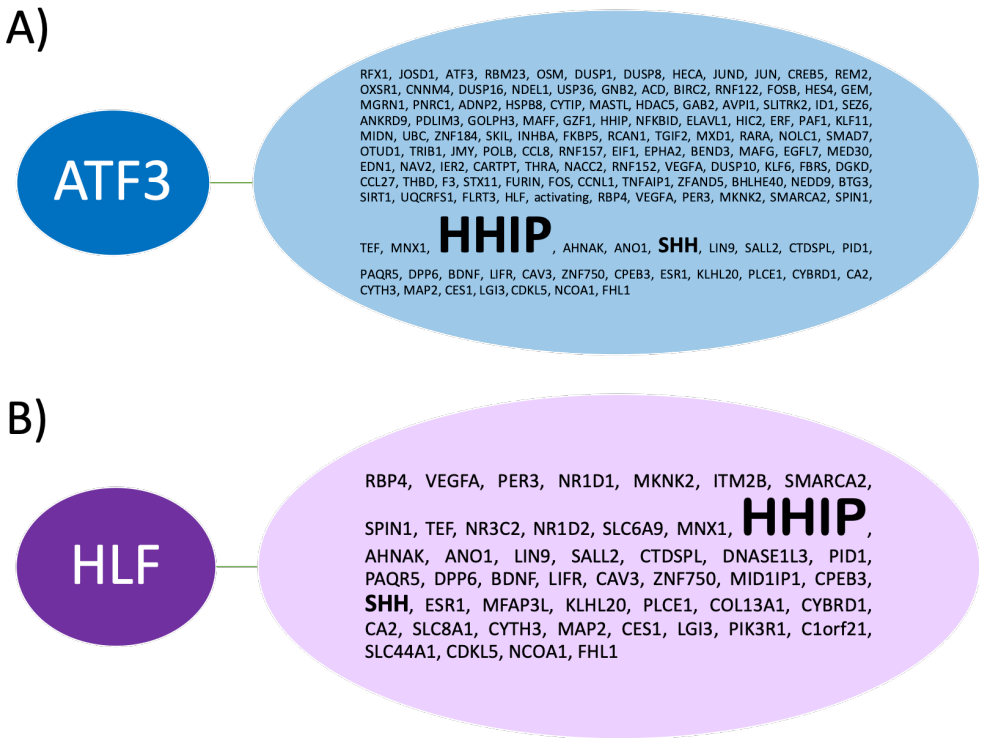


Figura 6.3: Regulones de *ATF3* (MA0018) y *HLF* (MA0043). Los regulones de *ATF3* y *HLF* también cuentan con variantes que modifican el motivo transcripcional. *ATF3* y *HLF* son factores transcripcionales, usando pySCENIC se calculó los módulos de regulación de dichos genes. Los genes mostrados en la figura son los modulos de regulación resultantes de pySCENIC respectivamente.

6.3). Las dos variantes que generan el cambio de afinidad son **rs1032295** y **rs12509311** respectivamente.

El factor transcripcional HLF mostró mayor afinidad en su matriz cuando el alelo T de la variante rs1032295 está presente en el genoma, el cual se asocia a la relación de FEV1/FVC% (Lutz et al., 2015, Kim et al. (2013), https://www.ebi.ac.uk/gwas/variants/rs1032295) mientras que el alelo menor, G, se reporta como un alelo de proteción contra la EPOC (Cortes et al., 2020, https://www.treewas.org/snp/rs1032295) (Figura 6.4).

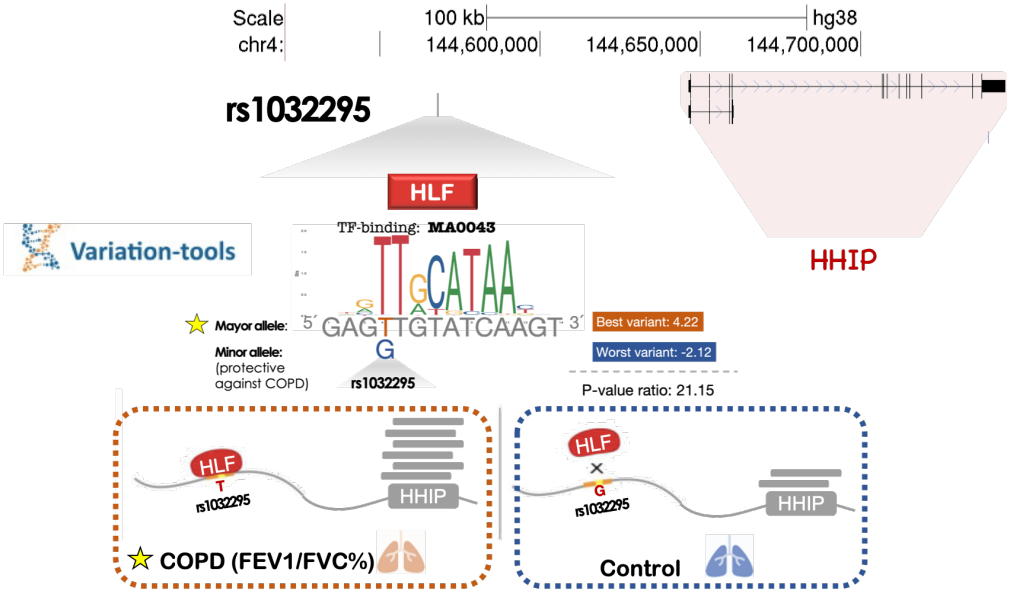


Figura 6.4: Motivo de *HLF* (MA0043) y la variante **rs1032295**. Se muestra la ubicación en la región río arriba de *HHIP* (chr 4, 144513432); los resutltados de *Variation-scan* para cada variante; y su interpretación en el contexto de pacientes con la EPOC (o *COPD* por sus siglas en inglés).

La variante **rs12509311** genera un cambio de afinidad en el motivo de *ATF3* (MA0018), el alelo que se encuentra mayormente en la población es C, se le conoce como alelo mayor. Mientras que el alelo con menor frecuencia en la población es T, también conocido como alelo menor. La variante **rs12509311** (T) se ha asociado con la relación FEV1/FVC% post broncodilatador pero no a EPOC (Lutz et al., 2015, https://www.ebi.ac.uk/gwas/variants/rs12509311). En la base de datos Treewas se reporta el alelo de riesgo (T) como alelo de proteción contra la EPOC (Cortes et al., 2020, https://www.treewas.org/snp/rs12509311). Sin embargo, en un estudio usando población china, se encontró que la variante **rs12509311**

(T) está asociada también con FEV1/FVC% en pacientes con EPOC. Los autores discuten las diferencias entre los resultados, como población distinta, poder estadistico y genética, a demás de discutir como la relación FEV1/FVC% cambia en distintas poblaciones. Adicionalemente, los autores no encuentran relación entre la variante **rs12509311** (T) y la susceptivilidad en la EPOC ni la severidad (Zhang et al., 2017).

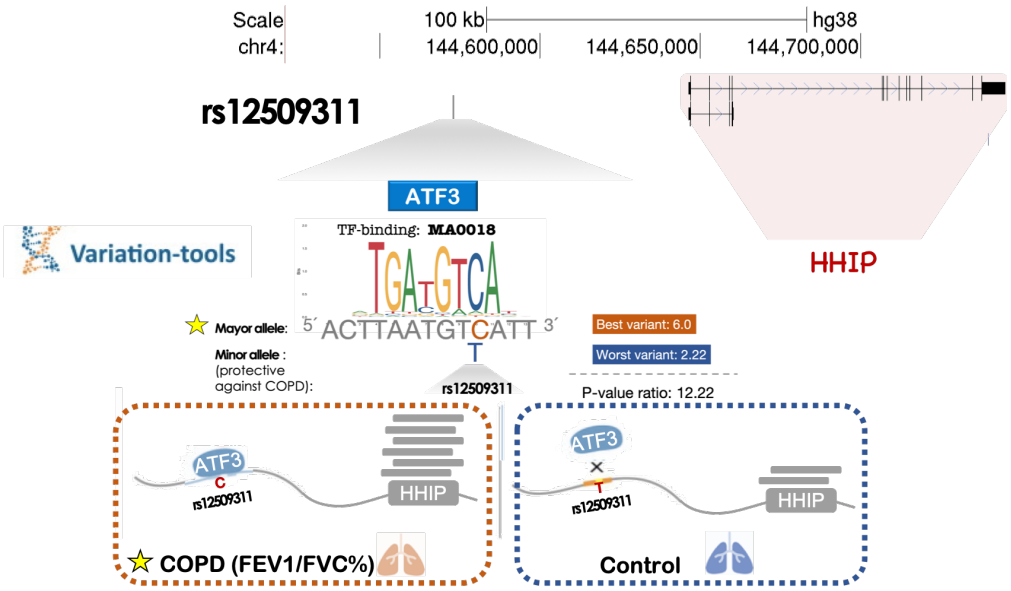


Figura 6.5: Motivo de *ATF3* (MA0018) y la variante **rs12509311**. Se muestra la ubicación en la región río arriba de *HHIP* (chr 4, 144557510); los resutltados de *Variation-scan* para cada variante; y su interpretación en el contexto de pacientes con la EPOC (o *COPD* por sus siglas en inglés).

6.4 Conclusiones

HHIP podría estar inhibiendo la vía de HH en tejido pulmonar de pacientes con la EPOC ya que se encontró sobreexpressión de *HHIP* pero no de los componentes de la vía.

Las variantes **rs12509311** y **rs1032295** pueden estar implicadas en la sobreexpresión de *HHIP*, confiriendo mayor afinidad a la matriz de los factores transcripcionales *ATF3* (MA0018) y *HLF* (MA0043) en pacientes con la EPOC.

Las variantes asociadas a la EPOC o a sus fenotípos pueden estar implicadas en la regulación de la expresión de genes en la EPOC. El contexto genético en la EPOC puede contribuir a

la regulación de genes importantes en la enfermedad. Y PulmonDB facilitó la obtención de los datos de expresión para el análisis.

6.5 Poster en ISMB

La estudiante de medicina *Mayra Padilla* presentó estos resultados en el poster titulado **Gene regulation of Hedgehog interacting protein (HHIP) in Chronic Obstructive Pulmonary Disease** para la International Society for Comutational Biology (ISMB) 2020(https://doi.org/10.7490/f1000research.1118486.1).

Capítulo 7

Conclusiones

Las enfermedades pulmonares, en especial la EPOC y FPI son enfermedades crónicas, complejas y que causan gran impacto en la vida de los pacientes por lo que diversos grupos se han dedicado al estudio de estas enfermedades (Selman et al., 2019, Meiners et al. (2015), Petty (2006), Lutz et al. (2015), **?**).

PulmonDB, mediante el uso de la interfaz web y el paquete R facilitará realizar comparaciones y análisis entre la EPOC y la FPI.

PulmonDB puede ayudar a la comunidad científica a estudiar los perfiles de expresión compartidos y únicos para la FPI y la EPOC, explorar experimentos a través de diferentes tecnologías y plataformas e identificar patrones de expresión que ayuden a generar nuevas hipótesis.

Usando PulmonDB encontramos genes expresados tanto en la EPOC como en la FPI que podrían estar asociados a los fenotipos compartidos en las dos enfermedades (*e.i* procesos inflamatorios y de envejecimiento).

También se encontró genes característicos de cada enfermedad que podrían estar participando en la etiología de cada una y ser usados como marcadores.

Los contrastes usados en PulmonDB recapitulan los perfiles de expresión calculados con una normalización clásica (*e.i.* RMA).

La heterogeneidad en los datos de expresión para la EPOC no se debe a la variación en la proporción celular.

Se encontró sobreexpresión del gen *HHIP* en los pacientes con la EPOC usando PulmonDB, contrastando lo reportado en la literatura.

Variantes en la región reguladora del gen *HHIP* podrían estar involucrados en la alteración de la regulación de dicho gen cambiando la afinidad en los factores de trascripción *ATR3* y *HLF*.

7.1 Perspectivas

PulmonDB demostró ser de gran valor por su anotación manual, vocabulario controlado y su capacidad de encontrarlo en un mismo lugar. La base de datos generó interés en la comunidad científica que estudia la EPOC y FPI. Sin embargo, la complejidad de los contrastes individuales que se plantean son limitantes, tanto para análisis posteriores como para su interpretación. Como futuras perspectivas, PulmonDB podría mejorar está situación con el uso del algoritmo *Remove Unwanted Variation* RUV (Risso et al., 2014; Jacob et al., 2016) porque los datos no serían contrastes como actualmente se manejan los datos en PulmonDB.

Adicionalmente, PulmonDB se encuentra en una base de datos relacional, la cual es lenta para implementar *queries* en diversas tablas, esto se podría mejorar migrando la base de datos a un tipo no relacional como *elasticsearch*.

Para mejorar y agilizar la curación manual, PulmonDB podría implementar curación semi automática, que permitiera a los curadores tener una experiencia más agradable y permitir actualizar de nuevos datos de manera más sencilla.

Las posibilidades de PulmonDB como recurso para analizar enfermedades pulmonares, permite su crecimiento en distintas dimensiones. Pueden incluirse enfermedades pulmonares asociadas a EPOC y FPI como asma, o enfermedades de relevancia actual como COVID-19, sin afectar el propósito fundamental de PulmonDB. A demás que PulmonDB puede crecer

aumentando el tipo de datos que contiene, es decir, podría agregarse información genética, epigenética, curación de artículos, etc. dada la flexibilidad de PulmonDB como base de datos. Adicionalmente PulmonDB podría almacenar datos de *single-cell* RNA-seq de enfermedades pulmonares, creciendo como un *HUB* para datos especializados de enfermedades pulmonares como la EPOC y FPI.

EPOC es una enfermedad compleja y heterogénea, para poder entender mejor está enfermedad y proponer nuevas terapias es importante entender la fuente de variación. Es necesario poder subclasificar a los pacientes para que obtengan el tratamiento más adecuado. Para ello, un análisis de sub-clasifición y detección de biomarcadores podría ayudar, seguido de validación experimental.

Por otro lado, tener acceso a datos de transcriptómica y genotipo en el mismo individuo mejoraría el análisis de regulación de variantes.

Apéndice: Técnicas de secuenciación y expresión génica a nivel ómico

La transcriptómica se enfoca en caracterizar la expresión de los genes que se encuentran activados en las células y existen técnicas que permiten caracterizar el perfil transcriptómico de forma masiva.

Para poder obtener el perfil de expresión genético de un organismo, se utilizan tecnologías como microarreglos o secuenciación masiva de RNA, ambas tecnologías tienen ventajas y desventajas y ambas se siguen ocupando. Históricamente, los microarreglos surgieron primero, La secuenciación de RNA es actualmente la más usada y cuenta con mayor resolución, lo que permite a los investigadores poder tener un perfil trancriptómico con más precisión.

Contribución personal

En el capítulo 10 del libro titulado **"Best Practices in Genetic and Genomic Research: Rigor, Reproducibility, and Protocols for the Laboratory and Classroom"** mi contribución personal fue en la realización de la figura 1, redacción de las secciones de *Splicing detection*, *Public databases*, *Reproducibility across studies* y *Conclusion and remarks*, así como en la revisión de todo el capítulo.

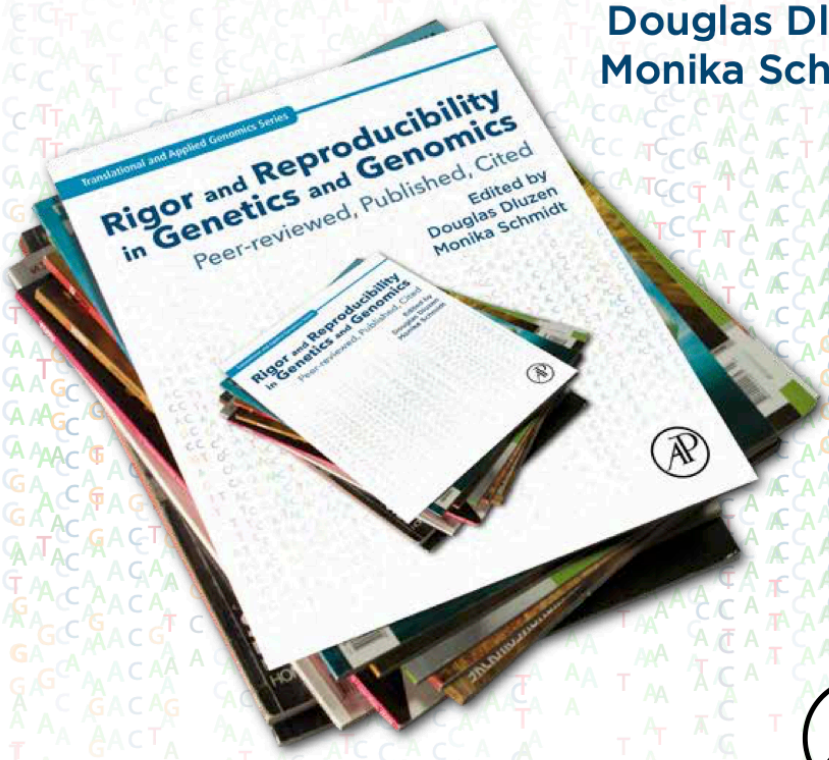
Publicación 1: *Review of 'omics-level gene sequencing and gene expression techniques*

Translational and Applied Genomics Series

Rigor and Reproducibility in Genetics and Genomics

Peer-reviewed, Published, Cited

Edited by
Douglas Dluzen
Monika Schmidt



Outline

Section 1: Introduction
Chapter 1: Review of reproducibility in genetic studies
Chapter 2: Replication Efforts in Genetic Publishing
Chapter 3: Rigor in the classroom and in the mentor/mentee relationship
Section 2: Genotyping
Chapter 4: Genome-wide association studies (GWAS): what are they, when to use them?
Chapter 5: Best practices in GWAS and pitfalls to avoid
Chapter 6: GWAS learning and training activities
Chapter 7: DNA sequencing for genotyping and best practices for comparative genomics
Chapter 8: Statistical approaches for rigorous genome sequence analyses and genotype imputations
Chapter 9: DNA sequencing activities; classroom case studies
Section 3: Next-Generation Sequencing and Gene Expression
Chapter 10: Review of 'omics-level gene sequencing and gene expression techniques
Chapter 11: Guidelines and important considerations for 'omics-level studies
Chapter 12: Best practices for statistical analysis of 'omics data
Chapter 13: Approaches to validate gene expression studies
Chapter 14: 'Omics misconceptions, classroom activities and case studies
Section 4: Epigenetic Analyses
Chapter 19: Review of DNA methylation and other omics data resources
Chapter 20: Best practices for ATAC-seq and its data analysis - assigned
Chapter 21: Best methods for combining DNA, RNA, and methylation data
Chapter 22: Teaching epigenetics in the classroom
Section 5: Gene Editing Technologies
Chapter 23: Review of current gene editing technologies, including CRISPR
Chapter 24: Best strategies to design and implement CRISPR-based genetic analysis
Chapter 25: CRISPR classroom activities and/or case studies

Chapter 10: *Review of gene expression using microarray and RNA-seq*

Ana B. Villaseñor-Altamirano[1], Yalbi I. Balderas-Martínez[2] & Alejandra Medina-Rivera[1]

[1]Laboratorio Internacional de Investigación sobre el Genoma Humano, UNAM, Juriquilla, Mexico
[2]Instituto Nacional de Enfermedades Respiratorias Ismael Cosío Villegas, Mexico City, Mexico

1. Introduction

This chapter will discuss the different technologies used to determine the expression levels of genes in a given sample using high throughput techniques.

High throughput techniques refer to those approaches that allow for assessing and identifying thousands of elements in one experiment. A great example is determining mRNA expression levels from the whole genome of an organism, which without high-throughput techniques will need to be done individually per gene.

Transcriptomics was born with the advent of high throughput techniques, particularly next generation sequencing methods. Other areas have benefited from this development, such as genomics and epigenomics.

In the following sections, we will review two major techniques applied in transcriptomics: Microarrays and RNA-sequencing, plus the most recent development of single cell RNA-seq (Figure 1). Microarrays are a classic technique that measures specific genes based on oligomers attached to a plate, representing the genes of interest, this was the first high throughput technique for measuring gene expression. Sequencing refers to the action of determining the nucleotide composition of a chain of DNA. This DNA can come directly from the genome of an organism (genome sequencing) or could have come from RNA (transcriptome sequencing) through retro transcription (cDNA).

RNA-sequencing has extended its application thanks to the development of new technologies such as long read sequencing and single cell sequencing. One of the challenges of high throughput techniques is reproducibility, as variation can come from many external variables: laboratory and platform (either microarray or sequencing). In the following sections, we will discuss these challenges and how to overcome them.

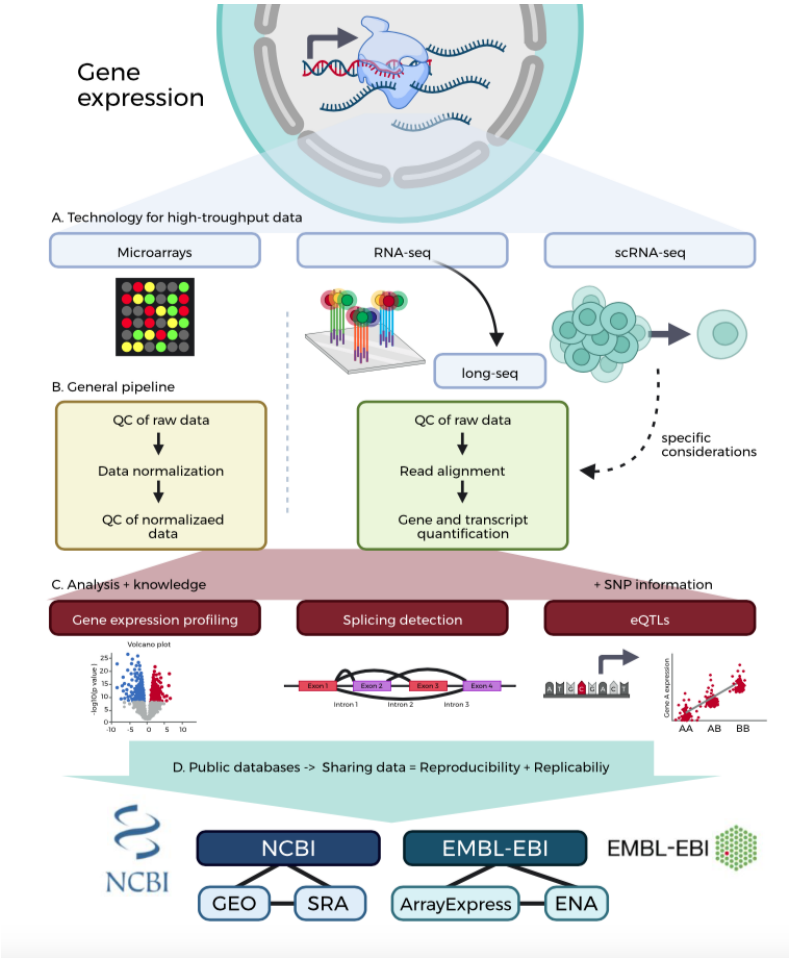


Figure 1: Overview of transcriptomic technologies, their general analysis pipelines, and applications. A) Microarrays and RNA-seq are the major transcriptomic technologies, recently RNA-seq has been paired with new approaches to allow for single cell RNA-seq. B) General analysis pipelines corresponding to each of the transcriptome technologies. C) Major applications of transcriptomics to generate biological knowledge. Gene expression profiling aims to identify which genes are relevant for a specific condition. Splicing detection aims to identify the different isoforms present in specific conditions. eQTL analysis aims to identify the relation between genetic variation and gene expression. D) Published transcriptomic data is available through different public resources so it can be realised or re-proposed for new studies.

2. High Throughput techniques to assess gene expression

a. Microarrays

The most popular technology to assess gene expression up today is microarrays. They were developed in the ‘90s, although before there were other technologies to study individual transcripts.

In the ‘80s, low-throughput Sanger sequencing was used to sequence individual transcripts called expressed sequence tags (ESTs), fragments of mRNA sequences obtained from randomly selected clones of cDNA libraries. With ESTs in 1991, it was possible to obtain the first brain transcriptome using 600 sequences (Adams et al., 1991). Later, in 1995, the protocol of serial analysis of gene expression (SAGE) was released (Velculescu et al., 1995), with the advantage of analyzing thousands of transcripts. The protocol starts extracting mRNA from a sample and obtaining cDNA with reverse transcriptase, cDNA is digested with restriction enzymes to produce “tag fragments,” these fragments are concatenated and sequenced, and later quantified with computer programs to assess the occurrence of individual tags. Other methods, such as northern blot, nylon membrane arrays, and reverse transcriptase quantitative PCR, became popular. However, these techniques only could retrieve a fraction of the transcriptome (Lowe et al., 2017).

The first classic paper showing microarray technology's potential was published in 1995 (Schena et al., 1995). Shena M. et al. used 45 complementary cDNA probes of Arabidopsis thaliana to hybridize with mRNA simultaneously, showing that using robotic printing would make it feasible to scale up the fabrication process. The technology was improved, and today it is possible to cover up until 6 million probes in only one chip. In general, microarray technology consists of measuring specific transcripts' abundances through mRNA hybridization with complementary probes in an array. More specifically, RNA samples obtained from cell culture, tissues, bacterial growth, or others are reverse transcribed to produce complementary cDNA and then labeled with fluorescence.

Microarrays are built with probes, which are spotted or directly synthesized onto a glass or silicon surface. Probes are usually localized at the 3’ region of the transcript, requiring a priori knowledge about the genome. Hybridization between sample cDNA and microarray probes produces a fluorescence intensity that correlates with the abundance of the transcripts. There are two ways to compare the experimental versus reference sample: using one or two-color arrays. On the first option, there is only one microarray by sample. On the second, there is one microarray, and samples are labeled with different fluorescent tags (Cy3, Cy5).

Microarrays can cover entire genomes or just a subset of genes (e.g., all related to apoptosis), and custom microarrays can be designed. In general, splicing is not covered, but some versions can detect them (see splicing section below). Applications are mainly on basic research and clinical practice for diagnosis or decision making (Govindarajan et al., 2012), but it can also be used on agriculture.

The protocol for analyzing microarray data usually starts with data that contains the intensity calculations. It is possible to use the manufacturer software. An alternative and popular option is to use open software such as Bioconductor packages (Gentleman et al., 2004; Huber et al., 2015), which has a variety of packages developed for different steps and platforms. Bioconductor has also created a microarray workflow analysis (arrays) so anyone can follow the basic steps. Analyzing microarray data requires at least the following steps (Figure 1B):

1. **Quality control of raw data:** First, it is necessary to perform a quality control analysis, this step is critical as introducing bad quality data will lead to bad results. Some exploratory analysis as boxplots of data intensity and principal component analysis can highlight issues in the experimental procedure. Depending on the figures, we might want to exclude some samples before the normalization step.

2. **Data normalization:** This step is essential to make the samples comparable between them and to eliminate technical variation. The normalization steps will change depending on the platform. For Affymetrix chips, RMA (robust-multiarray average) is the most common algorithm used, it performs three main steps: i) Background adjusted correction, ii) Quantile normalization, and iii) Summarization with median polish which results in log2 transformed values (Irizarry et al., 2003). For Agilent, i) Background correction is calculated using median, and ii) Global loess normalization on individual arrays or quantile normalization between arrays needs to be calculated before performing differential gene expression analysis (Ritchie et al., 2007).

3. **Quality control of normalized data and gene annotation:** After normalization, it is recommended to make a boxplot again and compare it with those obtained using raw data. Here it is important to verify that medians have the same expression level. We can also check if there are batch effects or errors introduced by different times or places when performing the experiment (see batch effects section below). Annotation files will be required to assign the probe with their respective gene.

4. **Define the contrast model:** After quality control has passed, it is necessary to define which samples (conditions or treatments) will be compared, this can be done using limma (Ritchie et al., 2015). The matrix is a table that describes which sample corresponds to which experimental condition or treatment. Then a contrast matrix is used to describe the comparisons between the groups. If there are additional covariates (*i.e.* age, time, sex, etc.), the model can be used to describe them(Ritchie et al., 2015; Wang et al., 2012).

5. **Differential gene expression:** Limma package implements the Empirical Bayes model to fit a linear model and moderates the standard errors of the estimated log-fold changes. It also returns the list of genes with their log-fold change and the adjusted p-value that will help us filter those genes that are changing the most. With the list of genes, the functional relevance can be evaluated through a Gene Ontology enrichment analysis using a package like enrichR.

Some disadvantages can be summarized as follows. When analyzing microarray data, the main challenge is the technical noise, and in most of the studies, there are few replicates to help disentangle this noise. Different methods have been proposed to deal with this problem

in each step of quality control, normalization, and differential expression analysis in the past years, and most of them have already been implemented on Bioconductor packages{REF}. Another problem is the accuracy of the expression measurements due to the background hybridization on transcripts with low abundance or inclusive saturation for those transcripts with high abundance. Probes are different in their hybridization properties, and there could be mismatches between the probe and the target molecule, so that specificity could be improved.

There are some differences between microarrays and RNA-seq, e.g., they have a limited range of sensitivity in the low and high ends. Sequence variation in the probe region could significantly impact how well the probe works, reduced hybridization by DNA polymorphisms could give results that are not real. Removing the affected probes can reduce the impact, but then information can be lost. This issue is particularly relevant in species where the genome variation is unknown.

b. Sequencing

i. Next Generation Sequencing

Next Generation Sequencing (NGS) refers to the series of platforms that started being developed in the mid 2000s. These platforms had a significant impact on the Human Genome Project, enabling cost reduction with throughput increase.

Before NGS, the major sequencing technique was Sanger, a technique still considered the gold standard, as it can produce long sequences with low error rates. Sanger sequencing is based on DNA synthesis, initially used dideoxynucleotides that will terminate transcription, termination bases will have a particular dye that will be read using electrophoresis capillaries (Sanger et al., 1977) . This technique is still widely used to confirm genetic variants, particularly for disease research or diagnosis.

NGS sequencing platforms are based on parallelization, where thousands or millions of sequencing reactions are happening simultaneously. In 2000, Massively Parallel Signature Sequencing Lynx technologies launched the first NGS platform that was then bought by Illumina. Next, a technique based on pyrosequencing, where DNA polymerase activity was measured based on the detection of pyrophosphate(Ahmadian et al., 2006), was commercialized as the 454 platform in 2005. Also during that year, Solexa released a platform with technology based in sequencing by synthesis using a reversible dye terminator, Solexa was bought then by Illumina in 2007. Over the years, different technologies and innovations were incorporated into NGS techniques. Platforms vary on their cost, throughput, and read length. Several reviews have been written further describing the details of the different chemistries used in commercial sequencing platforms(Goodwin et al., 2016; Levy and Myers, 2016; Zhang et al., 2011), and we encourage the reader to revise these papers.

Currently, sequencing technologies can be categorized into two major groups:

1. Sequencing by ligation: As the name states, this procedure involves ligating hybridized labeled probes and anchor sequences to the DNA. Labeled probes are degenerated sequences that will bind to the template and promote ligation that will release the label. SOLiD (Valouev et al., 2008) and BGISEQ (Huang et al., 2017) sequencing platforms are the two major representatives of this

strategy. However, SOLiD has been discontinued and BGISEQ is still under development and has not been launched for its commercialization yet.

2. Sequencing by synthesis: This term is used to describe sequencing methods that rely on DNA-polymerase. Major commercial platforms used in NGS based on this technology are Illumina and Ion Torrent.

In general, NGS requires that genetic material to be used is amplified, having more initial material to read allows for the correct detection of the sequenced bases. The major downside of NGS is the high error rates and the read lengths (currently ~200 on average and going from 50-400bp), read length refers to the number of nucleotides that can be determined from one sequence, the reads that come from a single sequencing assay will be of the same size on average. Error rates refers to the possibility of the sequencer incorrectly determining a base pair, this can be compensated by raising the number of sequences, this is known as sequencing depth or coverage, which is the number of times (or average of times) a unique sequence represents a nucleotide from the genome, i.e., ten unique sequencing reads can contain a nucleotide from the genome and confirm its identity.

NGS applications are broad, the technology has been adapted to characterize not only an organism genome (Lupski et al., 2010), but also RNA (transcriptomics) (Wang et al., 2009), or regulatory sequences (epigenomics)(Johnson et al., 2007). Particularly, RNA sequencing has enabled the detection of transcripts without pre-determining targets, as it was required for microarrays. This unbiased approach has allowed the detection of new transcripts (Weirick et al., 2016), and specifically it has facilitated the study of isoforms (Hardwick et al., 2019), novel noncoding RNAs (Shi et al., 2013), and genomic variants.

Adaptations have been made to RNA capture protocols to facilitate the identification of transcription start sites, one of these modifications is the Cap Analysis of Gene expression (CAGE) (Shiraki et al., 2003). The aim of CAGE is to select full transcripts for sequencing by selecting the 5' cap found in RNAs. CAGE has been extensively used in combination with sequencing platforms by the FANTOM consortium(Forrest et al., 2014) to identify transcription start sites across human and mouse tissues and cell types.

In general, RNA-seq data analysis has the following steps (Figure 1B):

1. **Quality control of sequencing data:** As described above, each sequencing platform will have particular sources of noise, for this reason, is important to check the general quality of the obtained sequences: base quality, CG content, presence of enriched k-mers (usually, adapters from the sequencing platforms), read length, sequence duplication levels, etc. FastQC(Andrews and Others, 2010) is an easy to use software that can facilitate this task.

2. **Read alignment:** Alignment refers to the task where each read obtained from sequencing is "aligned" to the reference genome, to identify its source. Many aligners are available for RNA-seq analysis, and development is still being made in this area (Arora et al., 2020). Take into account each tool has been designed to help answer a particular question, be sure you understand which aligner will fit your research better, some of the most commonly used aligners are bowtie2 (Langmead and Salzberg, 2012), kallisto (Bray et al., 2016), STAR (Dobin et al., 2013), and Salmon (Patro et al., 2015).

3. **Gene and transcript quantification:** The number of times sequenced reads align to a gene or transcript can be interpreted as the expression. Some aligners, like kallisto, Salmon, and STAR can also perform this task. These counts are usually normalized to allow for comparisons. Some commonly used normalization procedures are Reads per Kilobase Million (RPKM), Fragments Per Kilobase (FPKM), and Transcripts Per Million (TPM). A benchmark on methods to assess performance of RNA quantification methods is presented in (Teng et al., 2016).

4. **Differential Gene expression:** Once we quantify gene expression, one of the analyses usually performed is comparing expression of genes or transcripts across different samples. Many tools are available to perform these comparisons: edgeR (Robinson et al., 2010), DESeq2 (Love et al., 2014), and limma (Ritchie et al., 2015) are some of the most common ones.

5. **Additional Analysis:** RNA-seq data can provide the opportunity to answer many biological questions. Co-expression networks can give information regarding gene regulation (Moerman et al., 2019). Clustering analysis can give insights into shared pathways or mechanisms between genes or samples (Zhao et al., 2018). Comparing gene expression data across species using orthology information can give insight into key conserved pathways or mechanisms (Barbosa-Morais et al., 2012).

There are some available pipelines to ease the task of integrating these analysis steps such as RASflow (Zhang and Jonassen, 2020), VIPER (Cornwell et al., 2018), and BioJupies (Torre et al., 2018). Additionally, Galaxy, a web platform, provides the possibility of using RNA-seq analysis pipelines without installing software (Taylor et al., 2007).

ii. Long Read Sequencing

There is a great diversity of RNA molecules in a given cell, there are messenger RNAs that can come from different isoforms and noncoding RNAs that can vary in length between small (~10-40bps)(Boyd, 2008) to long (1-10kbs)(Zampetaki et al., 2018), particularly some messenger RNAs can be in the range of 1 to 2 kbs(Harrow et al., 2012). Messenger and long noncoding RNAs can have alternative splicing, increasing RNA diversity.

One of the major limitations of NGS platforms, as Illumina, is the necessity for RNA to be converted to cDNA and cut it to lengths of around 300bps to enable library preparation(Pease and Sooknanan, 2012). This has imposed a limit on our capacity to study full transcripts and hence isoforms.

Studies aimed to identify isoforms using NGS technologies have been enabled through transcriptome analysis tools(Merino et al., 2019). Some of these tools are focused on matching reads with similarities, like a puzzle, to identify full transcripts, this is known as transcriptome assembly, and has become a key analysis framework to study non-model organisms as it does not require having a sequenced reference genome(Hölzer and Marz, 2019). Nevertheless, there are limitations to these analyses, as the challenge comes from the difficulties in identifying correctly the exons that belong together in the same molecule vs. separating isoforms with differential exons, long read sequencing technologies can overcome these challenges.

The third generation of sequencing technologies refers to the ones that enable long read sequencing in the range of kilobases. The two primary technologies currently used in this field are PacBio, which generates read lengths of up to 15Kbs, and Oxford Nanopore which generate reads of >30Kbs. One additional advantage of Oxford Nanopore is that its technology allows for direct RNA sequencing, without the need to convert to cDNA before, reducing biases caused by this step.

However, one of the downsides of long read sequencing technologies is the low throughput these technologies have, while NGS technologies offer millions of reads per run, long read technologies reach at most 20Gb with Oxford Nanopore and PacBio is currently limited to up to 700,000 reads. Furthermore, long read sequencing still genders high positive error rates, PacBio currently reports ~15% error rates compared to ~0.1% observed in Illumina, this has limited long read sequencing usage.

Long read sequencing is continuously being improved, which will be translated into new applications in the future. One of the current developments to reduce the high error rate in Oxford Nanopore Technology is to add redundancy by sequencing the same chain repeatedly, which is the same rationale behind genome coverage. One of the strategies used is Circular Consensus Sequencing (CCS)(Travers et al., 2010) where, through circularization of the template, the sequence is read twice, and a consensus is created. However, CCS conveys a reduction in the length of the obtained reads, and new algorithms are being developed to propose new barcode designs that will allow for redundancy without affecting sequence length(Ezpeleta et al., 2017). Other computational approaches to reduce error rates in long read sequencing incorporate data from short read RNA-seq data(Choudhury et al., 2018; Wyman and Mortazavi, 2019).

The baseline protocol to analyze long read sequencing data includes the following general steps:

1. **Identify full length transcripts:** Transcripts are identified based on read clustering(Gordon et al., 2015), in this way, similar reads can be integrated to confirm one transcript. Genome information can be taken into account to correct transcript annotation(Wyman and Mortazavi, 2019).

2. **Quantify transcripts and genes:** Reads are assigned to transcripts, and the number of reads per transcript is converted to Transcript per Million to normalize.

Even with their current limitations, long read sequencing has proven to be useful for sequencing full length transcripts in human cell lines(Tilgner et al., 2013), to detect variation in isoform expression in human primary cells(Byrne et al., 2017) and survey full length transcripts in non model organisms(Ye et al., 2019), computational methods will continue to be developed to improve data analysis(Wyman et al., 2020). Nevertheless, RNA-seq based on NGS technology will remain the standard for the following years, particularly in clinical applications, as the technology has matured, and computational methods to analyse it are well established.

3. Applications

Regardless of the method used to measure gene expression, detection of RNA molecules can have diverse applications. In the next section, we will present an overview of the most common ones: Gene expression profiling, splicing detection, expression quantitative trait loci assessment and the novel application to identify single cell transcriptional profiles (Figure 1C).

a. Gene Expression Profiling

Gene expression profiling refers to the measurement of gene expression activity in thouthands of genes simultaneously. This can be done in samples coming from tissues, cell cultures, etc., and it has been primarily used in research, but it has shown to have clinical potential (Claussen et al., 2016).

Depending on the tissue, cell type, treatment or health/disease condition cells will change gene expression, this means different sets of genes will be expressed or repressed in order to allow for each cell to contend with the requirements of the environment.

Gene expression profiling studies usually aim to identify sets of genes that changed expression or that are unique to a particular state. In order to be able to identify changes in gene expression across samples it is important to properly plan and identify the relevant questions we want to answer. High-throughput methods can have high technical variability and this variability can have an impact on the results and could cause interpretation errors.

As described above, there are several platforms that can be used to measure gene expression (see section 1 and 2 of this chapter), but each of them has its particular pros and cons, so defining the question is very important for selecting the best one for your specific interest. For example, the best sequencing set up to detect long non coding RNAs is not the best for detecting polyadenylated messenger RNAs that code for proteins, in one case the best fit could be a combination of Illumina sequencing with PacBio, while for the other one an Illumina approach with poly A capturing would suffice.

Once a platform that matches best the requirements is selected, is time to plan the experiment or sampling that is required. Here is very important again to remember we have technical variability that can come from the platforms we select, this includes also sample processing that can cause batch effects. In this point it is important to avoid processing in different days or moments samples that have different biological features that are part of our question. For example, treatment samples and controls should try to be processed at the same time, same day, same place, to avoid technical differences that could be interpreted as biological ones. Nonetheless, sometimes to accommodate these requirements is not possible, so there are ways to contend with batch effects, we discuss this further in section 6b of this chapter.

In general, gene expression profiling analysis has the following steps (for recommendations specific to sequencing or microarrays, please see sections 1 and 2) :

1. Quality control of gene expression data. Depending on the platform used (sequencing or microarray) is important to evaluate that the obtained measurements can be used to proceed.

2. Gene and transcript quantification: Each platform will provide in a different way the measurement of gene or transcript quantification, microarrays represent this as intensity and sequencing as number of reads sequenced for that gene or transcript. These measurements constitute the first gene expression profile. However, these have to be corrected for any potential confounders or batch effects (see section 6), so they can be correctly interpreted or compared.

3. Differential Gene expression: Using corrected gene expression profiles it is possible to compare gene expression across samples.

4. Visualization: Heat Maps have been one of the classic displays used to show gene expression profiles. A heatmap is a matrix that contains expression values per gene per sample/condition.

Reproducibility can be a challenge in gene expression profiling analysis. As described before, batch effects are a common downside to these experiments, and this can affect reproducibility. Nevertheless, meta analysis approaches can be used to account for variations that come from the experimental origin (*i.e.* different laboratory and/or year), allowing to compare results that come from different studies. For a more detailed description of these procedures, please see section 6b of this chapter.

Differential gene expression has been widely used through the years to answer biological and clinical questions. Currently, gene expression profiling has been included as an standar analysis in major databases like UK biobank(Sudlow et al., 2015) and The Cancer Genome Atlas (Cancer Genome Atlas Research Network et al., 2013), and have been used to reach a better understanding of the biology of human traits and diseases.

Moreover, gene expression profile data is widely available for other organisms, which has opened the door for comparative genomics analysis, where conservation of gene expression levels is assessed. Comparative genomics gene expression has shown that in general mRNA patterns are shared across species in a given tissue(Brawand et al., 2011). However, when splicing patterns are compared it is possible to observe that there is an organism specific variation(Barbosa-Morais et al., 2012).

Gene Expression Profiling has been widely standardized, allowing it to be used to inform clinical practice (Szalat et al., 2016), and it has become one of the baseline analyses in transcriptomics.

b. Splicing detection

A key finding acquired from genome sequencing is the fact that cellular complexity has poor correlation with the number of protein-coding genes. Later on it was found that alternative splicing (AS) is used as a mechanism to create diversity which may lead to increased complexity(Blencowe, 2006). AS is a biological mechanism that generates multiple isoforms from a single gene by selecting different exons, and retaining introns. This mechanism allows

for protein diversity and even opposite function being coded in the same genomic loci(Li et al., 2014).

Multiple isoforms from the same gene have been recognized in eukaryotes genomes such as vertebrates, plants, and fungi, even unicellular organisms (e.i. *Sphaeroforma artica*)(Grau-Bové et al., 2018), this process is highly conserved and has been tracked to the last eukaryotic common ancestor (LECA)(Csuros et al., 2011). In addition, AS plays a crucial role in proteome diversity,92 to 94%(Wang et al., 2008) of protein-coding genes in humans go through this process, and has been shown to be tissue-specific(Modafferi and Black, 1999; Noh et al., 2006), relevant in organ development(Baralle and Giudice, 2017), and to have clinical applicability in human diseases such as cancer(Zhang et al., 2019). A large and old database to visually search for AS variations in humans is SpliceSeq(Ryan et al., 2012), a tool with RNA-seq data that facilitates visualization graphs. More recently, the same research team developed a resource with The Cancer Genome Atlas (TCGA) data to explore splicing events using a web based tool(Ryan et al., 2016).

The AS classification varies between authors which have described between five and eight types, the most common types being: inclusion or skipping of exons, intron retention, alternative 5' or 3' splice site events, differentially untranslated regions (UTRs) and alternative first and last exons, these classifications are furthered described by (Wang et al., 2015) and (Wang et al., 2008) respectively.

Moreover, exon skipping isoforms are particularly relevant in animals and have been shown to participate in regulating protein-protein interactions. Nonetheless, intron retentions have been associated with down-regulation by controlling nonsense-mediated decay (NMD), a process which regulates mRNA levels with premature stop codons(Baralle and Giudice, 2017).

Current research based on the identification of AS has been applied to cancer data in the TCGA consortium(Zhang et al., 2019). Using annotations from the SpliceSeq databases, Zhang, *et al.* were able to predict prognostic signatures for 31 cancer types using AS events by implementing a random forest survival model(Zhang et al., 2019). This shows the impact of AS in disease mechanisms and how it varies across individuals.

Alternative splicing in microarrays

Since microarrays became available as a technology to measure gene expression, scientists had interest to evaluate AS using high throughput technologies, therefore companies such as Affymetrix developed microarrays with specific probes that recognized spliced exons, introns, and exon's junctions. However, this technology is limited by probe design, as this is based on current reference genome annotations. Affymetrix tried to compensate for this by predicticting possible new splicing regions to help *de novo* discoveries, as consequence, probe sets where classify as core (based on RefSeq and GenBank annotation), extended (core probes plus EST and partial mRNA annotation) and full (extended plus *ab-initio* predictions)(Subbaram et al., 2010).

However, microarrays have small length probes, which are between 25 to 60 bp depending on the platform(Jaksik et al., 2015), and it is difficult to cover small exons with unique matches in the genome(Srinivasan et al., 2005) with a validation rate of 33% to 86%(Moore and Silver, 2008). Despite its disadvantages, microarrays were used to uncover

AS events in different conditions, such as tissue-specific variants(Clark et al., 2007), cancer(Lapuk et al., 2010), dioxin exposure(Villaseñor-Altamirano et al., 2019), etc.

The general pipeline for analyzing AS with microarrays is similar to the analysis of microarrays for gene expression:

1. **Quality control for raw data:** such as explained above (microarray section).
2. **Normalization:** such as explained above (microarray section).
3. **Quality control for normalized data:** such as explained above (microarray section).
4. **AS analysis:** Different methods have been developed to detect AS events using microarrays, some packages were developed such as Splicing Index(Srinivasan et al., 2005), FIRMA(Purdom et al., 2008), MADS(Xing et al., 2008), MiDAS(GeneChip, n.d.), ARH(Rasche and Herwig, 2010) among others.

Alternative splicing in RNA sequencing

RNA sequencing improved AS analysis dramatically since microarrays, and it has been used to confirm alternative isoforms and find new ones. This relatively unbiased technology gives the possibility to analyze transcriptome data without prior knowledge. However, using RNA-seq for AS analysis has challenges too. For example, aligning ambiguous reads from 100 to 150 bp length in some genomic regions is problematic because of its non-unique assignment(Hu et al., 2013). Transcript isoform abundance also partake in AS detection accuracy, highly expressed isoforms have comparable accuracy across different methods but it decreases with low abundance transcripts(Kanitz et al., 2015). In summary, RNA-seq enables a better analysis and understanding of AS events than microarrays but it has its own bias when studying differential splicing.

The general pipeline to analyse AS events with RNA-seq data starts by:

1. **Quality control of sequencing data:** explained above (sequencing section).
2. **Read alignment:** explained above (sequencing section). In addition, it is important to mention that RNA-seq aligning read methods can be classified depending on the reference used to identify genes:
    a. **Reference genome.** Which can detect novel transcripts but it is computationally expensive. First the reads are mapped to a reference genome (e.i. TopHat, STAR, HISAT2) and then another program is used to quantify transcript abundance (e.i. Cufflinks)(Ghosh and Chan, 2016).
    b. **Alignment-free or pseudoalignment** which is faster but only recognizes known transcripts (e.i. Kallisto, Salmon)(Bray et al., 2016).
    c. **De novo transcript assembly** which is helpful for unreliable reference genomes, unexisting references and distinguishing new transcripts (e.i. trinity, Oases,DiffSplice)(Mehmood et al., 2019; Sahraeian et al., 2017).

3. **Gene and transcript quantification:** explained above (sequencing seccion).

4. **Differential AS analysis:** After calculating read counts, we can perform differential analysis to discover AS variants across different conditions. Mehmood, et al described a classification for differential AS algorithms (Mehmood et al., 2019; Sahraeian et al., 2017):

    a. **Isoform-based:** these methods reconstruct full length transcripts and then quantifies the abundance for comparison among groups (Cufflinks(Trapnell et al., 2012), DiffSplice(Hu et al., 2013), Sleuth(Pimentel et al., 2017)).

    b. **Count-based:** these methods are subclassified in:

        i. Exon-based: which measure reads falling into an exon or junction region (limma(Ritchie et al., 2015), edgeR(Robinson et al., 2010). DESeq2(Love et al., 2014)).

        ii. Event-based: which calculate the fraction of spliced event types (spliced exons, retained introns, etc) and then compare ratios between groups (rMATS(Shen et al., 2014), SUPPA(Alamancos et al., 2015)).

Besides, taking into account if a reference genome or transcriptome is available, another important consideration for selecting a method to study AS is the experimental design because not all methods support more than two groups(Mehmood et al., 2019) (*e.i.* time series experimental designs). There have been different benchmark studies(Mehmood et al., 2019; Merino et al., 2019; Sahraeian et al., 2017; Teng et al., 2016) comparing these methods, an important note is the relevance of understanding all parameters in use by each method and adapt them accordingly. The selection of default values can deeply affect results, as these will not correctly model every experimental design(Baruzzo et al., 2017).

Representing AS events with an efficient visualization plot can be challenging because of the complex nature of this information. Heatmaps have been one of the classic displays used to show gene expression profiles but this is not useful for accurate AS representation. In AS the most common visualization strategy is to use sashimi plots(Garrido-Martín et al., 2018) and box plots with expression per exon. Moreover, Strobelt *et al*. developed a visual analysis tool for AS exploration with different alternative visualization graphs (Strobelt et al., 2016).

With single cell data (discussed below), new methods are under development (Huang and Sanguinetti, 2017) and tools such as Salmon, RSEM, Kallisto, among others have been shown similar performance in scRNA-seq and bulk RNA-seq (Westoby et al., 2018). However, technical obstacles need to be addressed to study alternative isoforms in single cell data as was discussed by (Westoby et al., n.d.).

c. Expression Quantitative Trait Loci Assessment

Genetic variation refers to the loci in a genome that can differ across members of a population. When two humans selected by chance from the population are compared, their genomes will defer approximately 0.1%, the only humans that share the same genome are identical twins. This means that about one in every 1,000 bps will differ between any two individuals (National Institutes of Health (US) and Biological Sciences Curriculum Study,

2007). Genetic variation comprises: single nucleotide polymorphisms (SNPs), changes of one nucleotide between two individual (*i.e.* A>T) ; insertions and deletions of nucleotides (indels) ; and copy number variations (CNVs), which is the repetition in tandem of a chain of nucleotides. Variation can be assessed either by microarrays and by whole genome sequencing, the latter being more expensive. SNPs are widely studied variations in humans as they can be easily assayed by microarrays.

One of the main reasons to study genetic variation is that it can be related to health and disease, this is widely addressed by Genome Wide Associations Studies (GWAS) that establish a relation between genetic variants and human traits(Manolio, 2010). It has been reported that more than 90% of the SNPs that have been associated with traits or diseases in humans through GWAS are located in non-coding regions (MacArthur et al., 2017). These variants have been related to changes in gene expression, probably affecting transcriptional regulation, as they are commonly found in open chromatin regions where regulatory proteins tend to bind (Vierstra et al., 2020).

A SNP in a non-coding region with a potential regulatory function associated with gene expression is known as expression quantitative trait loci (eQTL)(Nica and Dermitzakis, 2013). In order to establish the relation between a SNP and gene expression phenotype, we require a direct association test between genetic markers and gene expression levels, this analysis typically requires tens or hundreds of individuals.

eQTLs lay a connection between the genetic background and cell function, so as gene expression can be tissue specific, also eQTLs can be tissue and cell type specific(GTEx Consortium, 2020). Moreover, eQTLs can also be population specific and ancestry has to be taken into account, for example a study performed in colon samples of Han chinese population found 5,940 eQTLs, from which only 21.4% had been previously reported in another population(Guo et al., 2016).

Currently, the Genotype-Tissue Expression (GTEx) project launched in 2013 (GTEx Consortium, 2020, 2013), is the biggest resource of eQTL mapping in the human genome, with data for 15,201 RNA-sequencing samples from 49 tissues and 838 postmortem donors. GTEx measures tissue specific gene expression and identifies eQTLs in humans, these results have been harnessed in GWAS research to establish links between diseases and tissues (Gamazon et al., 2018). Moreover, eQTLs can be leveraged to establish biological mechanisms related to GWAS associated variants (Marigorta et al., 2017).

In general, an eQTL mapping study requires three things:

1. **Gene RNA levels** (transcriptome data) for tens (Lock et al., 2015), hundreds (Pala et al., 2017) or thousands of participants (Zhernakova et al., 2017). Gene expression analysis can be performed as described above.

2. **Genotypes** (SNPs in a genome) of the same individuals where transcriptome data was obtained from. This data has to undergo its own quality control, methods will differ depending on whether it was obtained from genotyping arrays of whole genome sequencing. This subject is out of the scope of the chapter.

3. **Covariates:** Information from participants such as sex, age, health status, etc.

This information then will be used to determine associations between SNPs and genetic expression, then later interpreted as a molecular trait. This association is possible based on the fact that gene expression, measured as RNA abundance, is heritable(Ouwens et al., 2020).

To establish the relation of a SNP loci to gene expression we have to assess the correlation between the expression of a gene and the alleles present in a group of people. While this sounds easy, it is important to take into account the genetic composition of the population (*i.e.* local ancestry and global ancestry)(Gay et al., 2020), and the established covariates that could bias the analysis. Several tools have been designed to perform this analysis, Matrix eQTL(Shabalin, 2012), QTL Tools(Delaneau et al., 2017) and FastQTL (Ongen et al., 2016) are some of them.

There are some considerations to be taken into account when performing eQTL mapping that could lead to false discoveries, we will mention the most relevant ones:

- Allele frequency in the population can bias results, this means the number of individuals should be higher if the variant is in low frequency.
- Another common issue is the winner's curse, this means that the true genetic effect of an eQTL is lower than the estimation obtained from the sample population, which can lead to replication problems (Huang et al., 2018).

As stated before, eQTL are variants associated with gene expression. Nevertheless, this association does not imply causality, and the identification of the causal variant requires fine mapping. eQTL variants are proxies of variants that are in linkage disequilibrium, this means that variants within a loci are inherited together more times than expected by chance. There are different tools to infer causal variants, some of them are: CaVEMAN(Brown et al., 2017), CAVIAR(Hormozdiari et al., 2014) and dap-g(Wen et al., 2017).

eQTLs are a useful resource to identify the plausible regulatory mechanisms affecting the expression of a gene in a given tissue or cell type. With the lowering cost of genotyping and RNA-sequencing, this tool has become broadly accessible and will enable a better understanding of the interplay between genetics and function.

4. Single cell RNA-seq

Single-cell RNA sequencing (scRNA-seq) is a technology that enables the obtention of transcriptome profiles per cell. We can obtain a better resolution of the molecular events inside a cell than the previous experiments using bulk RNA sequencing. The first experiment was performed in X???, and in a short time period different methods have been developed and standardized.

Single-cell RNA-seq is applied to resolve cell type subpopulations, evaluate the heterogeneity of the cells, or understand the dynamics of biological processes.

In general, they share some steps: 1) isolation of single cells, 2) reverse transcription, 3) cDNA amplification, and 4) sequencing library preparation (Hedlund and Deng 2018).

A crucial step on scRNA-seq is the isolation of the cells because it is essential to maintain RNA integrity. RNA isolation could be done with different experimental methods, the most used are: enzymatic treatment, laser capture microdissection and patch clamping.

Different numbers of cells can be used. Micro-pipetting or patch-seq are used to capture a few cells (range?). Fluorescent activated cell sorting and microfluidic approaches can be used to obtain a higher number of cells (range?) with specific characteristics. It is also possible to perform scRNA-seq of nuclei using FACS, in this case, it is obtained from unprocessed mRNA, which is less compared to the mRNA of the whole cell. The big challenge with this experiment is to remove the cytoplasm.

Once RNA is obtained, then reverse transcription can be used to synthesize cDNA. Usually, the protocol uses oligodT priming to avoid ribosomal RNA and select mRNAs, and some non-coding RNAs. cDNA can be amplified using SMART technology that switches the mechanism at the 5' end of the RNA template. Another option could be ligating the 5'end of cDNA with poly(A) or poly(C) to build common adaptors for PCR amplification. Here, the problem could be the generation of short amplicons shorter and with less G-C content. To solve this, a method CEL-seq or MARS-seq can be used to perform in vitro transcription.

After this step, there are different methods of sequencing: 1) those based on full-length where it is obtained the full coverage, and 2) those based on 5' or 3' tags. scRNA-seq methods based on tags can be combined with unique molecular identifiers (UMIs) to quantify transcript molecules, but read obtained through this method cannot be used for splicing detection. So, depending on the biological question, one method could be better than the other. For projects interested in discovering cell types and knowing cell composition, a method based on tags could be enough and help reduce costs. When comparing methods, Smart-seq2 (REF) is better in sensitivity and reproducibility, but for a large number of cells, a technology like Drop-seq (REF) could detect up to 4000 genes.

When analyzing single-cell data there are some challenges that need to be considered. First, the matrix with the genes in rows, and cells in columns is going to be full of zeros, and high cell-to-cell variability that can be caused by technical and biological noise. Particularly, a cell can be in a specific cell cycle, in the transition to a different cellular state, have a specific size, and stochastic gene expression. Like other experiments, it is possible to have batch effects that can be detected through some methods like principal component analysis, and can be corrected using multiple batch correction methods. It is also possible to add spike-in RNA standards of known abundance to the endogenous samples (e.g., spike-ins of the RNA Control Consortium ERCC), can be used for quality control or normalization.

5. Public databases

a. Databases

Public biological databases are helpful to share data and promote open science by organizing and making data available. The main public databases storing transcriptome data are: 1) the National Center for Biotechnology Information (NCBI), 2) the European Molecular

Biology Laboratory's European Bioinformatics Institute (EMBL-EBI), and 3) the DNA Databank of Japan (DDBJ). These databases often share data between them, and provide links to interconnect information, they store different genetic, transcriptomic, proteomic and other biological information which have led to reorganizing each database in more specialized repositories("Genomic Data Resources: Curation, Databasing, and Browsers," n.d.) that can store raw and pre-processed high-throughput data.

The particular repositories to store functional genomic projects from high-throughput experiments (e.i. microarrays and sequencing data) are Gene Expression Omnibus (GEO)(geo, n.d.) from NCBI, and ArrayExpress(EMBL-EBI, n.d.) from EMBL-EBI. However, with development of NGS technologies the amount of data increased exponentially, and specific databases for efficiently storing raw data and alignment information had to be developed. As a consequence, the Sequence Read Archive (SRA)("Home - SRA - NCBI," n.d.) and its analogous in EMBL-EBI, the European Nucleotide Archive (ENA)(EMBL-EBI, n.d.) were created (Figure 1D).

Both databases, SRA and ENA, contain high throughput DNA and RNA sequencing data of different organisms, metagenome studies and environmental impact surveys. It is essential to highlight the importance of raw data for reproducibility, as results replication highly depend on this. We will further discuss this issue in section 6.

i. NCBI

1. GEO-NCBI

GEO has been a useful free resource to archive genomic high-throughput data, we can find gene expression experiments using microarrays or RNA-seq, but also methylation, genetic variation (SNPs), non-coding microarrays or sequencing, as well as immunoprecipitation studies for analysing chromatin accessibility, protein analysis among others. This repository not only aims to store raw and/or normalized data, but also to be up to date for the scientific community, and to provide an interface that allows to query, detect and download data.

GEO can be overwhelming for new users due to how the data is stored, the amount of data it contains, and the lack of uniformity to annotate the information. We will review each of these points.

GEO organizes the information by using three main records, and each of them has its own ID:

- Samples (GSMxxx): This record corresponds to a unique sample element and it is used to describe the background, conditions, experimental details and specific descriptors of the biological sample, such as disease status, treatment, age, sex, culture media, temperature, etc.. This ID contains preprocessed data and the normalized method should be indicated, optionally, a supplementary file with raw data can be included. The identification number is constructed by GSM plus numbers (*i.e.* GSM4565413) which will be connected to only one platform and can be part of multiple experiments (or Series).

- Series (GSExxx): This record links samples and it is used to give a global description of the experiment, such as the summary, analysis and finding of each study as well as the aim, general methodology, number of samples used in each comparative group, and person in charge, contact information, submission date, location, etc. This ID is unique and can contain raw data, meta information from all samples, or any additional file required for the experiment. Typically, each Series will group samples from the sample platform and later assign another GSE ID, this record is called SuperSerie and it can indicate an experiment that uses different platforms (e.i. when a sample was measured using gene expression arrays and methylation arrays, check GSE56342 as example). However, there are no rules on how to use SuperSeries.

- Platform (GPL): A GPL ID corresponds to a unique identification for a platform used to measure high--throughput data in a sample (*i.e.* an specific Affymetrix array). It contains a description of the technology used, number of measured features, as well as manufacturer information such as company, year, etc. and can have tables with probe information. It is particularly useful for getting back gene names from array probes.

In addition to these records, there is another type which tracks GSE IDs (or series experiments) that have been curated by the NCBI team and corresponds to GDS ID. These types of IDs can be used for tools provided by NCBI such as gene expression profile charts and DataSet clusters, but this advantage for using additional tools is limited to specific curated data sets.

It is important to understand how public databases store high-throughput data but it is also relevant to know how to download all this information. Some useful ways to do this process are:

- Website: The normalized data will be accessible for either microarrays or RNA-seq and raw data for microarrays can be found in as supplementary data either as an experiment using GSE or per sample using GSM. However, be aware that some experiments will not provide that information. Raw data for RNA-seq can be downloaded from SRA (see below).

- E-Utils: a tool for downloading data through a programmatic access (geo, n.d.).

- ftp: Data can be downloaded using a ftp structure such as ftp://ftp.ncbi.nlm.nih.gov/geo/series/GSE1nnn/GSE1000/suppl/GSE1000_RAW.tar, for more information check the tutorial(geo, n.d.).

- GEOquery: a package from Bioconductor that will download GEO data into R(Davis and Meltzer, 2007) environment.

- GEOparse: a python package analogous to GEOquery to download GEO data into python("Welcome to GEOparse's documentation! — GEOparse 1.2.0 documentation," n.d.).

2. Sequence Read Archive (SRA)

NCBI has a special repository for raw high-throughput sequencing data called Sequence Read Archive (SRA). The aim of SRA is to make data available for reproducibility and facilitate collaborative work("Home - SRA - NCBI," n.d.; Leinonen et al., 2011).

For downloading SRA data you can use:

- Website: SRA website contains a browser that facilitated exploration of data. This information is linked to GEO, so it is easy to move across the two databases. It is possible to use this interface to download the desired raw data files and meta information.
- SRA toolkit: Series of commands that enable reading and downloading of sequencing files from SRA.
- SRAdb: It is a Bioconductor package to download SRA data into R(Zhu et al., 2013).
- pysradb: a python package to download SRA and ENA data into python("pysradb," n.d.).

ii. EMBL-EBI

1. ArrayExpress

ArrayExpress contains high-throughput data from genomic experiments in different organisms, is part of the EMBL-EBI service, and is connected to the GEO database. ArrayExpress, as well as GEO, storages description and metadata information, raw data files and pre-processed data, all this information is publicly available. The data contained in ArrayExpress can be either imported from GEO or directly submitted to the repository, ergo it has some GEO experiments and unique submissions(EMBL-EBI, n.d.).

The datasets directly submitted into ArrayExpress are manually curated to verify the minimum requirements to enable reproducibility of the results(EMBL-EBI, n.d.). This process follows specific guidelines which are Minimum Information About a Microarray Experiment (MIAME)("FGED: MIAME," n.d.) and Minimum Information about a high-throughput SEQuencing Experiment (MINSEQE)(Abeygunawardena, 2008). These standards make ArrayExpress a consistent repository.

Data can be downloaded by:

- Website: A searchable web site that allows the usage of keywords to facilitate exploration. Direct links to processed and raw data are available.
- ftp: Raw data is stored in the European Nucleotide Archive database that is directly linked in ArrayExpress. In each Project (*i.e.* experiment) website, a table will be displayed with the available files, direct download options are displayed.ArrayExpress is connected directly to the Galaxy platform (Afgan et al., 2018) (https://usegalaxy.org/) which is a web based analysis resource highly recommended for new NGS users that wish to analyse their own data.

2. ENA

Similar to SRA, the ENA stores raw data, assembly information and functional annotation to facilitate data accessibility.

ENA does not provide specialized tools to download data, as easy to use links are available and can be plugged into wget commands in a Unix terminal.

b. Public data-based resources

Public databases have been used to re-analyze and collect available data such as Recount (Collado-Torres et al., 2017), COLOMBOS (Moretto et al., 2016a), VESPUCCI (Moretto et al., 2016b) and PulmonDB (Villaseñor-Altamirano et al., 2020) with different aims but taking advantage of published stored data. Moreover, different tools allow you to reanalyze public data through specific programming languages (e.i. GEOquery, SRAdb, pysradb), but other tools have created more user-friendly and web based applications such as GREIN (Mahi et al., 2019), ImaGEO (Toro-Domínguez et al., 2019), GEOprofiles (Barrett et al., 2005). However, not all experiments may be available, and it is focused generally in human, rat, mouse and model organisms.

6. Reproducibility across studies

Part of the scientific process is to have a systematic methodology allowing other groups to replicate results , this is the very fundament of how a scientific community creates knowledge. Therefore, a researcher's responsibility is to ensure trustable experiments by minimizing potential errors and noise so another researcher can replicate findings. Particularly to computational experiments, results should be recreated by using the same data and code, this is reproducibility(Plesser, 2017; Rougier et al., 2017).

Reproducibility and replicability are usually confused terms, but generally replicability refers to recreating an entire study from an independent investigator with new data and similar algorithms. While reproducibility is the ability to produce the same results using the same data. Plesser discussed these concepts in detail and reviewed different meanings used in the literature(Plesser, 2017).

Also The Turing Way community has created a great resource that describes these terms and created guidelines for reproducibility. Same data and same analysis: Reproducibility; different data and same analysis: Replicability; same data and different analysis: Robust; and finally, different data and different analysis: Generalisable(The Turing Way Community et al., 2019).

In this section, we will focus on computational reproducibility and replicability in high throughput techniques to assess gene expression, we will also briefly discuss methods that allow us to summarize information from different studies such as batch effect correction and meta-analysis.

When RNA-seq started to be used for measuring gene expression, scientists questioned if microarrays and RNA-seq were generalisable results. (Marioni et al., 2008), proved microarrays can be comparable with RNA-seq by using the same samples in different platforms with technical replicates per technology. The authors found RNA-seq has small changes in technical replicates, and high correlation with microarray results since both platforms identified similar differentially expressed genes.

Moreover, several publications have evaluated the generalisability of high throughput transcriptomic platforms by comparing microarrays and RNA-seq. For example, using data from TCGA, 11,120 transcripts were found to be correlated across the two technologies (Chen et al., 2017), nevertheless, authors reported that transcripts with too high or low expression levels were more discrepant across methods.

Batch effect

In all experiments there will be noise but depending on the experimental design that noise may interfere in the analysis and lead us to wrong conclusions or this unwanted heterogeneity can be identified and corrected to perform our analysis.

As discussed above, it is ideal to have all samples tested at the same time by the same person to avoid differences of non-biological relevance. However, it is not always feasible and is necessary to add this variation into our analysis, but the experiment needs to be planned carefully for being able to identify the technical variation.

When samples can not be processed together, the experiment can be designed in batches but it is crucial to always process the samples with your interested condition (*i.e.* disease, mutation, temperature, drug) together with samples from the control group. It does not matter the used technology or platform, experimental design is key and it will help answer the desired scientific question. Some guidelines with good experimental design points can be found in (Conesa et al., 2016; Leek et al., 2010; McIntyre et al., 2011; Schurch et al., 2016) and for scRNA-seq (Hicks et al., 2018).

Running the analysis in batches can lead to technical noise and exploratory analyses are highly recommended before performing the downstream analyses(Leek et al., 2010). One of the most used exploratory analyses is dimensional reduction algorithms such as principal component analysis (PCA), and t-Distributed Stochastic Neighbor (tSNE). It is a good practice to always visualize your data before differential expression analysis to detect potential batch effects. If the experiment has additional variables (also called covariates such as age, sex, day of processing, library size, etc.) which creates batches, these covariates can be described to fit the differential expression model design(Love et al., 2020, 2014; Mostafavi et al., 2013).

If an experiment has not been designed properly there is slim room for improvement, nevertheless, there are methods for combining different experiments. Adding more samples into the analysis will increase power, however, raw data from different experiments can not be merged directly, even if they have the same hypothesis, because they are run by different laboratories and most of the time using different platforms.

There are some approaches to integrate datasets with similar hypotheses but different origins; some authors refer to these methods as cross-platform(Walsh et al., 2015) when

multiple experiments are integrated into one dataset. These methods can estimate the technical variance and correct the expression before a downstream analysis such as Surrogate Variable Analysis (SVA)(Leek, 2014; Leek et al., 2012). A benchmarking comparison for different batch effect correction methods in microarray data (without considering SVA) concluded that ComBat outperformed over other methods(Chen et al., 2011). However, in RNA-seq experiments, ComBat showed over estimation of batch effect correction and SVA had better performance(Liu and Markatou, 2016).

Meta-analysis

Another approach to summarize different results is a meta-analysis, which helps to identify generalisable results. A meta-analysis is a statistical approach that helps to integrate results from different studies to obtain global conclusions. There are two main models for a meta-analysis, i) Fixed-effect model: which assumes a conditional inference and estimates the effect size only using the studies included in the meta-analysis, and ii) Random-effect model: that assumes studies are a random sample, and calculates a theoretical overall population which provides an unconditional inference(Sweeney et al., 2017; Viechtbauer, 2010).

A meta-analysis is not restricted to high throughput data but it has been used for gene expression(Reinhold et al., 2017; Sweeney et al., 2017; Waldron et al., 2014). Additional to R packages for conducting a meta-analysis(Polanin et al., 2017) (e.i. metafor(Viechtbauer, 2010) as rmeta(Lumley, 2009)), there are web based tools to help researchers to do a meta-analysis using public gene expression data such as ImaGEO(Toro-Domínguez et al., 2019) and ExAtlas(Sharov et al., 2015), for a more complete list of softwares and websites available for microarray meta-analysis please check (Walsh et al., 2015).

7. Conclusion and remarks

High throughput techniques have transformed and improved transcriptomic analysis since they can identify gene expression levels from thousands of elements to the whole genome of an organism. Microarrays were the first technology and still in use nowadays but different RNA-seq methodologies emerged to enable unbiased profiling gene expression, and over the years, RNA-seq has become a popular, cheap, and versatile technology. Most recently, scRNA-seq is used for characterizing gene expression from a unique cell.

In this chapter, we introduced general pipelines for analyzing transcriptomic data depending on the technology used to measure gene expression. These pipelines are key to avoid systematic errors that can drive the analysis to wrong conclusions or misleading results. Therefore, a mandatory practice in transcriptomic analysis is to check the data quality and remove noise coming from platforms. Other standard practices during preprocessing data are i) to annotate the used parameters and additional files such as the reference genome used for alignment in RNA-seq, ii) to visualize raw data and pre-processed data, and iii) to identify potential outliers and/or batch effects.

Multiple analysis can be performed after preprocessing transcriptomic data, a frequent analysis is to compare gene expression profiling which determines differential expression genes that change across different conditions. At the same time, high throughput transcriptomic data (usually RNA-seq) can evaluate splicing events and investigate eQTLs if

SNP information is available. In this chapter, we discussed the most recurrent analysis for gene expression data, however, there are other analyses that can be computed. For example, allele expression in which RNA-seq data can be used to identify allele-specific expressions that together with eQTLs can be integrated to determine the regulatory effect on each allele(Castel et al., 2020).

A key factor for ensuring reproducibility and replicability is sharing data as far as the ethics and proper consideration allow. This chapter summarized two massive public repositories for gene expression, GEO from NCBI and ArrayExpress from EMBL-EBI, we briefly reviewed repositories for sequencing data (*e.i.* SRA and ENA), and offered procedures for downloading data.

As a summary, this chapter detailed the two most common high throughput technologies to evaluate gene expression as well as their general pipeline and discussed different applications for transcriptomic data. We also described two public databases for sharing data and promoting reproducible and replicable results.

8. References

Abeygunawardena, N., 2008. MINSEQE-Workgroups-FGED.

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F., 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252, 1651–1656.

Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Cech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B.A., Guerler, A., Hillman-Jackson, J., Hiltemann, S., Jalili, V., Rasche, H., Soranzo, N., Goecks, J., Taylor, J., Nekrutenko, A., Blankenberg, D., 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res. 46, W537–W544.

Ahmadian, A., Ehn, M., Hober, S., 2006. Pyrosequencing: history, biochemistry and future. Clin. Chim. Acta 363, 83–94.

Alamancos, G.P., Pagès, A., Trincado, J.L., Bellora, N., Eyras, E., 2015. Leveraging transcript quantification for fast computation of alternative splicing profiles. RNA 21, 1521–1531.

Andrews, S., Others, 2010. FastQC: a quality control tool for high throughput sequence data.

Arora, S., Pattwell, S.S., Holland, E.C., Bolouri, H., 2020. Variability in estimated gene expression among commonly used RNA-seq pipelines. Sci. Rep. 10, 2734.

Baralle, F.E., Giudice, J., 2017. Alternative splicing as a regulator of development and tissue identity. Nat. Rev. Mol. Cell Biol. 18, 437–451.

Barbosa-Morais, N.L., Irimia, M., Pan, Q., Xiong, H.Y., 2012. The evolutionary landscape of alternative splicing in vertebrate species.

Barrett, T., Suzek, T.O., Troup, D.B., Wilhite, S.E., Ngau, W.-C., Ledoux, P., Rudnev, D., Lash, A.E., Fujibuchi, W., Edgar, R., 2005. NCBI GEO: mining millions of expression profiles--database and tools. Nucleic Acids Res. 33, D562–6.

Baruzzo, G., Hayer, K.E., Kim, E.J., Di Camillo, B., FitzGerald, G.A., Grant, G.R., 2017. Simulation-based comprehensive benchmarking of RNA-seq aligners. Nat. Methods 14, 135–139.

Blencowe, B.J., 2006. Alternative splicing: new insights from global analyses. Cell 126, 37–47.

Boyd, S.D., 2008. Everything you wanted to know about small RNA but were afraid to ask. Lab. Invest. 88, 569–578.

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., Albert, F.W., Zeller, U., Khaitovich, P., Grützner, F., Bergmann, S., Nielsen, R., Pääbo, S., Kaessmann, H., 2011. The evolution of gene expression levels in mammalian organs. Nature 478, 343–348.

Bray, N., Pimentel, H., Melsted, P., Pachter, L., 2016. Near-optimal RNA-Seq quantification with kallisto.

Brown, A.A., Viñuela, A., Delaneau, O., Spector, T.D., Small, K.S., Dermitzakis, E.T., 2017. Predicting causal variants affecting expression by using whole-genome sequencing and RNA-seq from multiple human tissues. Nat. Genet. 49, 1747–1751.

Byrne, A., Beaudin, A.E., Olsen, H.E., Jain, M., Cole, C., Palmer, T., DuBois, R.M., Forsberg, E.C., Akeson, M., Vollmers, C., 2017. Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. Nat. Commun. 8, 16027.

Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M., 2013. The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. 45, 1113–1120.

Castel, S.E., Aguet, F., Mohammadi, P., GTEx Consortium, Ardlie, K.G., Lappalainen, T., 2020. A vast resource of allelic expression data spanning human tissues. Genome Biol. 21, 234.

Chen, C., Grennan, K., Badner, J., Zhang, D., Gershon, E., Jin, L., Liu, C., 2011. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. PLoS One 6, e17238.

Chen, L., Sun, F., Yang, X., Jin, Y., Shi, M., Wang, L., Shi, Y., Zhan, C., Wang, Q., 2017. Correlation between RNA-Seq and microarrays results using TCGA data. Gene 628, 200–204.

Choudhury, O., Chakrabarty, A., Emrich, S.J., 2018. HECIL: A Hybrid Error Correction Algorithm for Long Reads with Iterative Learning. Sci. Rep. 8, 9936.

Clark, T.A., Schweitzer, A.C., Chen, T.X., Staples, M.K., Lu, G., Wang, H., Williams, A., Blume, J.E., 2007. Discovery of tissue-specific exons using comprehensive human exon microarrays. Genome Biol. 8, R64.

Claussen, C.M., Lee, H., Shah, J.J., Richards, T., Shah, N., Patel, K., Bashir, Q., Parmar, S., Thomas, S., Nieto, Y., Qazilbash, M.H., Davis, R.E., Neelapu, S.S., Weber, D.M., Orlowski, R.Z., Feng, L., Manasanch, E.E., 2016. Gene Expression Profiling Predicts Clinical Outcomes in Newly Diagnosed Multiple Myeloma Patients in a Standard of Care Setting. Blood 128, 5628–5628.

Collado-Torres, L., Nellore, A., Jaffe, A.E., 2017. recount workflow: Accessing over 70,000 human RNA-seq samples with Bioconductor. F1000Res. 6, 1558.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M.W., Gaffney, D.J., Elo, L.L., Zhang, X., Mortazavi, A., 2016. A survey of best practices for RNA-seq data analysis. Genome Biol. 17, 13.

Cornwell, M., Vangala, M., Taing, L., Herbert, Z., Köster, J., Li, B., Sun, H., Li, T., Zhang, J., Qiu, X., Pun, M., Jeselsohn, R., Brown, M., Liu, X.S., Long, H.W., 2018. VIPER: Visualization Pipeline for RNA-seq, a Snakemake workflow for efficient and complete RNA-seq analysis. BMC Bioinformatics 19, 135.

Csuros, M., Rogozin, I.B., Koonin, E.V., 2011. A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes. PLoS Comput. Biol. 7, e1002150.

Davis, S., Meltzer, P.S., 2007. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics.

Delaneau, O., Ongen, H., Brown, A.A., Fort, A., Panousis, N.I., Dermitzakis, E.T., 2017. A complete tool set for molecular QTL discovery and analysis. Nat. Commun. 8, 15452.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21.

EMBL-EBI, n.d. ArrayExpress [WWW Document]. URL https://www.ebi.ac.uk/arrayexpress/ (accessed 9.21.20a).

EMBL-EBI, n.d. ENA Browser [WWW Document]. URL https://www.ebi.ac.uk/ena/browser/home (accessed 9.21.20b).
Ezpeleta, J., Krsticevic, F.J., Bulacio, P., Tapia, E., 2017. Designing robust watermark barcodes for multiplex long-read sequencing. Bioinformatics 33, 807–813.
FGED: MIAME [WWW Document], n.d. URL http://www.fged.org/projects/miame/ (accessed 10.19.20).
Forrest, A.R.R., Kawaji, H., Rehli, M., Baillie, J.K., de Hoon, M.J.L., Lassmann, T., Itoh, M., Summers, K.M., Suzuki, H., Daub, C.O., Kawai, J., Heutink, P., Hide, W., Freeman, T.C., Lenhard, B., Bajic, V.B., Taylor, M.S., Makeev, V.J., Sandelin, A., Hume, D. a., Carninci, P., Hayashizaki, Y., 2014. A promoter-level mammalian expression atlas. Nature 507, 462–470.
Gamazon, E.R., Segrè, A.V., van de Bunt, M., Wen, X., Xi, H.S., Hormozdiari, F., Ongen, H., Konkashbaev, A., Derks, E.M., Aguet, F., Quan, J., GTEx Consortium, Nicolae, D.L., Eskin, E., Kellis, M., Getz, G., McCarthy, M.I., Dermitzakis, E.T., Cox, N.J., Ardlie, K.G., 2018. Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. Nat. Genet. 50, 956–967.
Garrido-Martín, D., Palumbo, E., Guigó, R., Breschi, A., 2018. ggsashimi: Sashimi plot revised for browser- and annotation-independent splicing visualization. PLoS Comput. Biol. 14, e1006360.
Gay, N.R., Gloudemans, M., Antonio, M.L., Abell, N.S., Balliu, B., Park, Y., Martin, A.R., Musharoff, S., Rao, A.S., Aguet, F., Barbeira, A.N., Bonazzola, R., Hormozdiari, F., GTEx Consortium, Ardlie, K.G., Brown, C.D., Im, H.K., Lappalainen, T., Wen, X., Montgomery, S.B., 2020. Impact of admixture and ancestry on eQTL analysis and GWAS colocalization in GTEx. Genome Biol. 21, 233.
GeneChip, A., n.d. Exon Array Whitepaper Collection,“Alternative Transcript Analysis Methods for Exon Arrays,” rev. Oct. 11, 2005, ver. 1.1.
Genomic Data Resources: Curation, Databasing, and Browsers [WWW Document], n.d. URL https://www.nature.com/scitable/topicpage/genomic-data-resources-challenges-and-promises-743721/ (accessed 9.21.20).
Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A.J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J.Y.H., Zhang, J., 2004. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 5, R80.
geo, n.d. Home - GEO - NCBI [WWW Document]. URL https://www.ncbi.nlm.nih.gov/geo/ (accessed 9.21.20a).
geo, n.d. Programmatic access to GEO [WWW Document]. URL https://www.ncbi.nlm.nih.gov/geo/info/geo_paccess.html (accessed 10.21.20b).
Ghosh, S., Chan, C.-K.K., 2016. Analysis of RNA-Seq Data Using TopHat and Cufflinks. Methods Mol. Biol. 1374, 339–361.
Goodwin, S., McPherson, J.D., McCombie, W.R., 2016. Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. 17, 333–351.
Gordon, S.P., Tseng, E., Salamov, A., Zhang, J., Meng, X., Zhao, Z., Kang, D., Underwood, J., Grigoriev, I.V., Figueroa, M., Schilling, J.S., Chen, F., Wang, Z., 2015. Widespread Polycistronic Transcripts in Fungi Revealed by Single-Molecule mRNA Sequencing. PLoS One 10, e0132628.

Govindarajan, R., Duraiyan, J., Kaliyappan, K., Palanisamy, M., 2012. Microarray and its applications. J. Pharm. Bioallied Sci. 4, S310–2.
Grau-Bové, X., Ruiz-Trillo, I., Irimia, M., 2018. Origin of exon skipping-rich transcriptomes in animals driven by evolution of gene architecture. Genome Biol. 19, 135.
GTEx Consortium, 2013. The Genotype-Tissue Expression (GTEx) project. Nat. Genet. 45, 580–585.
GTEx Consortium, 2020. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science 369, 1318–1330.
Guo, C.C., Wei, N., Liang, S.H., Wang, B.L., Sha, S.M., Wu, K.C., 2016. Population-specific genome-wide mapping of expression quantitative trait loci in the colon of Han Chinese. J. Dig. Dis. 17, 600–609.
Hardwick, S.A., Joglekar, A., Flicek, P., Frankish, A., Tilgner, H.U., 2019. Getting the Entire Message: Progress in Isoform Sequencing. Front. Genet. 10, 709.
Harrow, J., Frankish, A., Gonzalez, J., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., Steward, C., Harte, R., Lin, M., Howald, C., Tanzer, A., Derrien, T., Chrast, J., Walters, N., Balasubramanian, S., Pei, B., Tress, M., Rodriguez, J., Ezkurdia, I., Van Baren, J., Brent, M., Haussler, D., Kellis, M., Valencia, A., Reymond, A., Gerstein, M., Guigo, R., Hubbard, T., 2012. GENCODE: The reference human genome annotation for The ENCODE Project. Genome Res. 22, 1760–1774.
Hicks, S.C., Townes, F.W., Teng, M., Irizarry, R.A., 2018. Missing data and technical variability in single-cell RNA-sequencing experiments. Biostatistics 19, 562–578.
Hölzer, M., Marz, M., 2019. De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. Gigascience 8.
Home - SRA - NCBI [WWW Document], n.d. URL https://www.ncbi.nlm.nih.gov/sra (accessed 9.21.20a).
Home - SRA - NCBI [WWW Document], n.d. URL https://www.ncbi.nlm.nih.gov/sra (accessed 10.15.20b).
Hormozdiari, F., Kostem, E., Kang, E.Y., Pasaniuc, B., Eskin, E., 2014. Identifying causal variants at loci with multiple signals of association. Genetics 198, 497–508.
Huang, J., Liang, X., Xuan, Y., Geng, C., Li, Y., Lu, H., Qu, S., Mei, X., Chen, H., Yu, T., Sun, N., Rao, J., Wang, J., Zhang, W., Chen, Y., Liao, S., Jiang, H., Liu, X., Yang, Z., Mu, F., Gao, S., 2017. A reference human genome dataset of the BGISEQ-500 sequencer. Gigascience 6, 1–9.
Huang, Q.Q., Ritchie, S.C., Brozynska, M., Inouye, M., 2018. Power, false discovery rate and Winner's Curse in eQTL studies. Nucleic Acids Res. 46, e133.
Huang, Y., Sanguinetti, G., 2017. BRIE: transcriptome-wide splicing quantification in single cells. Genome Biol. 18, 123.
Huber, W., Carey, V.J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B.S., Bravo, H.C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K.D., Irizarry, R.A., Lawrence, M., Love, M.I., MacDonald, J., Obenchain, V., Oleś, A.K., Pagès, H., Reyes, A., Shannon, P., Smyth, G.K., Tenenbaum, D., Waldron, L., Morgan, M., 2015. Orchestrating high-throughput genomic analysis with Bioconductor. Nat. Methods 12, 115–121.
Hu, Y., Huang, Y., Du, Y., Orellana, C.F., Singh, D., Johnson, A.R., Monroy, A., Kuan, P.-F., Hammond, S.M., Makowski, L., Randell, S.H., Chiang, D.Y., Hayes, D.N., Jones, C., Liu, Y., Prins, J.F., Liu, J., 2013. DiffSplice: the genome-wide

detection of differential splicing events with RNA-seq. Nucleic Acids Res. 41, e39.
Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., Speed, T.P., 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4, 249–264.
Jaksik, R., Iwanaszko, M., Rzeszowska-Wolny, J., Kimmel, M., 2015. Microarray experiments and factors which affect their reliability. Biol. Direct 10, 46.
Johnson, D.S., Mortazavi, A., Myers, R.M., Wold, B., 2007. Genome-wide mapping of in vivo protein-DNA interactions. Science 316, 1497–1502.
Kanitz, A., Gypas, F., Gruber, A.J., Gruber, A.R., Martin, G., Zavolan, M., 2015. Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. Genome Biol. 16, 150.
Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359.
Lapuk, A., Marr, H., Jakkula, L., Pedro, H., Bhattacharya, S., Purdom, E., Hu, Z., Simpson, K., Pachter, L., Durinck, S., Wang, N., Parvin, B., Fontenay, G., Speed, T., Garbe, J., Stampfer, M., Bayandorian, H., Dorton, S., Clark, T.A., Schweitzer, A., Wyrobek, A., Feiler, H., Spellman, P., Conboy, J., Gray, J.W., 2010. Exon-level microarray analyses identify alternative splicing programs in breast cancer. Mol. Cancer Res. 8, 961–974.
Leek, J.T., 2014. svaseq: removing batch effects and other unwanted noise from sequencing data. Nucleic Acids Research.
Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E., Storey, J.D., 2012. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics 28, 882–883.
Leek, J.T., Scharpf, R.B., Bravo, H.C., Simcha, D., Langmead, B., Johnson, W.E., Geman, D., Baggerly, K., Irizarry, R.A., 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat. Rev. Genet. 11, 733–739.
Leinonen, R., Sugawara, H., Shumway, M., International Nucleotide Sequence Database Collaboration, 2011. The sequence read archive. Nucleic Acids Res. 39, D19–21.
Levy, S.E., Myers, R.M., 2016. Advancements in Next-Generation Sequencing. Annu. Rev. Genomics Hum. Genet. 17, 95–115.
Li, H.-D., Menon, R., Omenn, G.S., Guan, Y., 2014. The emerging era of genomic data integration for analyzing splice isoform function. Trends Genet. 30, 340–347.
Liu, Q., Markatou, M., 2016. Evaluation of methods in removing batch effects on RNA-seq data. Infect Dis Transl Med 2, 3–9.
Lock, E.F., Soldano, K.L., Garrett, M.E., Cope, H., Markunas, C.A., Fuchs, H., Grant, G., Dunson, D.B., Gregory, S.G., Ashley-Koch, A.E., 2015. Joint eQTL assessment of whole blood and dura mater tissue from individuals with Chiari type I malformation. BMC Genomics 16, 11.
Love, M.I., Anders, S., Huber, W., 2020. Analyzing RNA-seq data with DESeq2 [WWW Document]. URL https://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html (accessed 12.2.20).
Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550.
Lowe, R., Shirley, N., Bleackley, M., Dolan, S., Shafee, T., 2017. Transcriptomics

technologies. PLoS Comput. Biol. 13, e1005457.
Lumley, T., 2009. rmeta: Meta-analysis. R package version 2.
Lupski, J.R., Reid, J.G., Gonzaga-Jauregui, C., Rio Deiros, D., Chen, D.C.Y., Nazareth, L., Bainbridge, M., Dinh, H., Jing, C., Wheeler, D.A., McGuire, A.L., Zhang, F., Stankiewicz, P., Halperin, J.J., Yang, C., Gehman, C., Guo, D., Irikat, R.K., Tom, W., Fantin, N.J., Muzny, D.M., Gibbs, R.A., 2010. Whole-Genome Sequencing in a Patient with Charcot–Marie–Tooth Neuropathy. N. Engl. J. Med. 362, 1181–1191.
MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., Pendlington, Z.M., Welter, D., Burdett, T., Hindorff, L., Flicek, P., Cunningham, F., Parkinson, H., 2017. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 45, D896–D901.
Mahi, N.A., Najafabadi, M.F., Pilarczyk, M., Kouril, M., Medvedovic, M., 2019. GREIN: An Interactive Web Platform for Re-analyzing GEO RNA-seq Data. Sci. Rep. 9, 7580.
Manolio, T.A., 2010. Genomewide association studies and assessment of the risk of disease. N. Engl. J. Med. 363, 166–176.
Marigorta, U.M., Denson, L.A., Hyams, J.S., Mondal, K., Prince, J., Walters, T.D., Griffiths, A., Noe, J.D., Crandall, W.V., Rosh, J.R., Mack, D.R., Kellermayer, R., Heyman, M.B., Baker, S.S., Stephens, M.C., Baldassano, R.N., Markowitz, J.F., Kim, M.-O., Dubinsky, M.C., Cho, J., Aronow, B.J., Kugathasan, S., Gibson, G., 2017. Transcriptional risk scores link GWAS to eQTLs and predict complications in Crohn's disease. Nat. Genet. 49, 1517–1521.
Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., Gilad, Y., 2008. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genome Research.
McIntyre, L.M., Lopiano, K.K., Morse, A.M., Amin, V., Oberg, A.L., Young, L.J., Nuzhdin, S.V., 2011. RNA-seq: technical variability and sampling. BMC Genomics 12, 293.
Mehmood, A., Laiho, A., Venäläinen, M.S., McGlinchey, A.J., Wang, N., Elo, L.L., 2019. Systematic evaluation of differential splicing tools for RNA-seq studies. Brief. Bioinform.
Merino, G.A., Conesa, A., Fernández, E.A., 2019. A benchmarking of workflows for detecting differential splicing and differential expression at isoform level in human RNA-seq studies. Brief. Bioinform. 20, 471–481.
Modafferi, E.F., Black, D.L., 1999. Combinatorial control of a neuron-specific exon. RNA 5, 687–706.
Moerman, T., Aibar Santos, S., Bravo González-Blas, C., Simm, J., Moreau, Y., Aerts, J., Aerts, S., 2019. GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. Bioinformatics 35, 2159–2161.
Moore, M.J., Silver, P.A., 2008. Global analysis of mRNA splicing. RNA 14, 197–203.
Moretto, M., Sonego, P., Dierckxsens, N., Brilli, M., Bianco, L., Ledezma-Tejeida, D., Gama-Castro, S., Galardini, M., Romualdi, C., Laukens, K., Collado-Vides, J., Meysman, P., Engelen, K., 2016a. COLOMBOS v3.0: leveraging gene expression compendia for cross-species analyses. Nucleic Acids Res. 44, D620–3.
Moretto, M., Sonego, P., Pilati, S., Malacarne, G., Costantini, L., Grzeskowiak, L., Bagagli, G., Grando, M.S., Moser, C., Engelen, K., 2016b. VESPUCCI:

Exploring Patterns of Gene Expression in Grapevine. Front. Plant Sci. 7, 633.
Mostafavi, S., Battle, A., Zhu, X., Urban, A.E., Levinson, D., Montgomery, S.B., Koller, D., 2013. Normalizing RNA-Sequencing Data by Modeling Hidden Covariates with Prior Knowledge.
National Institutes of Health (US), Biological Sciences Curriculum Study, 2007. Understanding Human Genetic Variation. National Institutes of Health (US).
Nica, A.C., Dermitzakis, E.T., 2013. Expression quantitative trait loci: present and future. Philos. Trans. R. Soc. Lond. B Biol. Sci. 368, 20120362–20120362.
Noh, S.-J., Lee, K., Paik, H., Hur, C.-G., 2006. TISA: tissue-specific alternative splicing in human and mouse genes. DNA Res. 13, 229–243.
Ongen, H., Buil, A., Brown, A.A., Dermitzakis, E.T., Delaneau, O., 2016. Fast and efficient QTL mapper for thousands of molecular phenotypes. Bioinformatics 32, 1479–1485.
Ouwens, K.G., Jansen, R., Nivard, M.G., van Dongen, J., Frieser, M.J., Hottenga, J.-J., Arindrarto, W., Claringbould, A., van Iterson, M., Mei, H., Franke, L., Heijmans, B.T., A C 't Hoen, P., van Meurs, J., Brooks, A.I., BIOS Consortium, Penninx, B.W.J.H., Boomsma, D.I., 2020. A characterization of cis- and trans-heritability of RNA-Seq-based gene expression. Eur. J. Hum. Genet. 28, 253–263.
Pala, M., Zappala, Z., Marongiu, M., Li, X., Davis, J.R., Cusano, R., Crobu, F., Kukurba, K.R., Gloudemans, M.J., Reinier, F., Berutti, R., Piras, M.G., Mulas, A., Zoledziewska, M., Marongiu, M., Sorokin, E.P., Hess, G.T., Smith, K.S., Busonero, F., Maschio, A., Steri, M., Sidore, C., Sanna, S., Fiorillo, E., Bassik, M.C., Sawcer, S.J., Battle, A., Novembre, J., Jones, C., Angius, A., Abecasis, G.R., Schlessinger, D., Cucca, F., Montgomery, S.B., 2017. Population- and individual-specific regulatory variation in Sardinia. Nature Genetics.
Patro, R., Duggal, G., Kingsford, C., 2015. Salmon: Accurate, Versatile and Ultrafast Quantification from RNA-seq Data using Lightweight-Alignment.
Pease, J., Sooknanan, R., 2012. A rapid, directional RNA-seq library preparation workflow for Illumina® sequencing. Nat. Methods 9, i–ii.
Pimentel, H., Bray, N.L., Puente, S., Melsted, P., Pachter, L., 2017. Differential analysis of RNA-seq incorporating quantification uncertainty. Nat. Methods 14, 687–690.
Plesser, H.E., 2017. Reproducibility vs. Replicability: A Brief History of a Confused Terminology. Front. Neuroinform. 11, 76.
Polanin, J.R., Hennessy, E.A., Tanner-Smith, E.E., 2017. A Review of Meta-Analysis Packages in R. J. Educ. Behav. Stat. 42, 206–242.
Purdom, E., Simpson, K.M., Robinson, M.D., Conboy, J.G., Lapuk, A.V., Speed, T.P., 2008. FIRMA: a method for detection of alternative splicing from exon array data. Bioinformatics 24, 1707–1714.
pysradb [WWW Document], n.d. URL https://pypi.org/project/pysradb/ (accessed 10.21.20).
Rasche, A., Herwig, R., 2010. ARH: predicting splice variants from genome-wide data with modified entropy. Bioinformatics 26, 84–90.
Reinhold, D., Morrow, J.D., Jacobson, S., Hu, J., Ringel, B., Seibold, M.A., Hersh, C.P., Kechris, K.J., Bowler, R.P., 2017. Meta-analysis of peripheral blood gene expression modules for COPD phenotypes. PLoS One 12, e0185682.
Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., Smyth, G.K., 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43, e47.

Ritchie, M.E., Silver, J., Oshlack, A., Holmes, M., Diyagama, D., Holloway, A., Smyth, G.K., 2007. A comparison of background correction methods for two-colour microarrays. Bioinformatics 23, 2700–2707.
Robinson, M.D., McCarthy, D.J., Smyth, G.K., 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140.
Rougier, N.P., Hinsen, K., Alexandre, F., Arildsen, T., Barba, L.A., Benureau, F.C.Y., Titus Brown, C., de Buyl, P., Caglayan, O., Davison, A.P., Delsuc, M.-A., Detorakis, G., Diem, A.K., Drix, D., Enel, P., Girard, B., Guest, O., Hall, M.G., Henriques, R.N., Hinaut, X., Jaron, K.S., Khamassi, M., Klein, A., Manninen, T., Marchesi, P., McGlinn, D., Metzner, C., Petchey, O., Plesser, H.E., Poisot, T., Ram, K., Ram, Y., Roesch, E., Rossant, C., Rostami, V., Shifman, A., Stachelek, J., Stimberg, M., Stollmeier, F., Vaggi, F., Viejo, G., Vitay, J., Vostinar, A.E., Yurchak, R., Zito, T., 2017. Sustainable computational science: the ReScience initiative. PeerJ Comput. Sci. 3, e142.
Ryan, M.C., Cleland, J., Kim, R., Wong, W.C., Weinstein, J.N., 2012. SpliceSeq: a resource for analysis and visualization of RNA-Seq data on alternative splicing and its functional impacts. Bioinformatics 28, 2385–2387.
Ryan, M., Wong, W.C., Brown, R., Akbani, R., Su, X., Broom, B., Melott, J., Weinstein, J., 2016. TCGASpliceSeq a compendium of alternative mRNA splicing in cancer. Nucleic Acids Res. 44, D1018–22.
Sahraeian, S.M.E., Mohiyuddin, M., Sebra, R., Tilgner, H., Afshar, P.T., Au, K.F., Bani Asadi, N., Gerstein, M.B., Wong, W.H., Snyder, M.P., Schadt, E., Lam, H.Y.K., 2017. Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. Nat. Commun. 8, 59.
Sanger, F., Nicklen, S., Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U. S. A. 74, 5463–5467.
Schena, M., Shalon, D., Davis, R.W., Brown, P.O., 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270, 467–470.
Schurch, N.J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., Wrobel, N., Gharbi, K., Simpson, G.G., Owen-Hughes, T., Blaxter, M., Barton, G.J., 2016. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA 22, 839–851.
Shabalin, A.A., 2012. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics 28, 1353–1358.
Sharov, A.A., Schlessinger, D., Ko, M.S.H., 2015. ExAtlas: An interactive online tool for meta-analysis of gene expression data. J. Bioinform. Comput. Biol. 13, 1550019.
Shen, S., Park, J.W., Lu, Z.-X., Lin, L., Henry, M.D., Wu, Y.N., Zhou, Q., Xing, Y., 2014. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc. Natl. Acad. Sci. U. S. A. 111, E5593–601.
Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., Fukuda, S., Sasaki, D., Podhajska, A., Harbers, M., Kawai, J., Carninci, P., Hayashizaki, Y., 2003. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc. Natl. Acad. Sci. U. S. A. 100, 15776–15781.
Shi, X., Sun, M., Liu, H., Yao, Y., Song, Y., 2013. Long non-coding RNAs: A new

frontier in the study of human diseases. Cancer Lett. 339, 159–166.
Srinivasan, K., Shiue, L., Hayes, J.D., Centers, R., Fitzwater, S., Loewen, R., Edmondson, L.R., Bryant, J., Smith, M., Rommelfanger, C., Welch, V., Clark, T.A., Sugnet, C.W., Howe, K.J., Mandel-Gutfreund, Y., Ares, M., Jr, 2005. Detection and measurement of alternative splicing using splicing-sensitive microarrays. Methods 37, 345–359.
Strobelt, H., Alsallakh, B., Botros, J., Peterson, B., Borowsky, M., Pfister, H., Lex, A., 2016. Vials: Visualizing Alternative Splicing of Genes. IEEE Trans. Vis. Comput. Graph. 22, 399–408.
Subbaram, S., Kuentzel, M., Frank, D., Dipersio, C.M., Chittur, S.V., 2010. Determination of alternate splicing events using the Affymetrix Exon 1.0 ST arrays. Methods Mol. Biol. 632, 63–72.
Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., Collins, R., 2015. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 12, e1001779.
Sweeney, T.E., Haynes, W.A., Vallania, F., Ioannidis, J.P., Khatri, P., 2017. Methods to increase reproducibility in differential gene expression via meta-analysis. Nucleic Acids Res. 45, e1.
Szalat, R., Avet-Loiseau, H., Munshi, N.C., 2016. Gene Expression Profiles in Myeloma: Ready for the Real World? Clin. Cancer Res. 22, 5434–5442.
Taylor, J., Schenck, I., Blankenberg, D., Nekrutenko, A., 2007. Using galaxy to perform large-scale interactive data analyses. Curr. Protoc. Bioinformatics Chapter 10, Unit 10.5.
Teng, M., Love, M.I., Davis, C.A., Djebali, S., Dobin, A., Graveley, B.R., Li, S., Mason, C.E., Olson, S., Pervouchine, D., Sloan, C.A., Wei, X., Zhan, L., Irizarry, R.A., 2016. A benchmark for RNA-seq quantification pipelines. Genome Biol. 17, 74.
The Turing Way Community, Arnold, B., Bowler, L., Gibson, S., Herterich, P., Higman, R., Krystalli, A., Morley, A., O'Reilly, M., Whitaker, K., 2019. The Turing Way: A Handbook for Reproducible Data Science.
Tilgner, H., Raha, D., Habegger, L., Mohiuddin, M., Gerstein, M., Snyder, M., 2013. Accurate identification and analysis of human mRNA isoforms using deep long read sequencing. G3 3, 387–397.
Toro-Domínguez, D., Martorell-Marugán, J., López-Domínguez, R., García-Moreno, A., González-Rumayor, V., Alarcón-Riquelme, M.E., Carmona-Sáez, P., 2019. ImaGEO: integrative gene expression meta-analysis from GEO database. Bioinformatics 35, 880–882.
Torre, D., Lachmann, A., Ma'ayan, A., 2018. BioJupies: Automated Generation of Interactive Notebooks for RNA-Seq Data Analysis in the Cloud. Cell Syst 7, 556–561.e3.
Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., Pachter, L., 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. 7, 562–578.
Travers, K.J., Chin, C.-S., Rank, D.R., Eid, J.S., Turner, S.W., 2010. A flexible and efficient template format for circular consensus sequencing and SNP detection. Nucleic Acids Res. 38, e159.
Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., Zeng, K.,

Malek, J.A., Costa, G., McKernan, K., Sidow, A., Fire, A., Johnson, S.M., 2008. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. Genome Res. 18, 1051–1063.
Velculescu, V.E., Zhang, L., Vogelstein, B., Kinzler, K.W., 1995. Serial analysis of gene expression. Science 270, 484–487.
Viechtbauer, W., 2010. Conducting meta-analyses in R with the metafor package. J. Stat. Softw. 36, 1–48.
Vierstra, J., Lazar, J., Sandstrom, R., Halow, J., Lee, K., Bates, D., Diegel, M., Dunn, D., Neri, F., Haugen, E., Rynes, E., Reynolds, A., Nelson, J., Johnson, A., Frerker, M., Buckley, M., Kaul, R., Meuleman, W., Stamatoyannopoulos, J.A., 2020. Global reference mapping of human transcription factor footprints. Nature 583, 729–736.
Villaseñor-Altamirano, A.B., Moretto, M., Maldonado, M., Zayas-Del Moral, A., Munguía-Reyes, A., Romero, Y., García-Sotelo, J.S., Aguilar, L.A., Aldana-Assad, O., Engelen, K., Selman, M., Collado-Vides, J., Balderas-Martínez, Y.I., Medina-Rivera, A., 2020. PulmonDB: a curated lung disease gene expression database. Sci. Rep. 10, 514.
Villaseñor-Altamirano, A.B., Watson, J.D., Prokopec, S.D., Yao, C.Q., Boutros, P.C., Pohjanvirta, R., Valdés-Flores, J., Elizondo, G., 2019. 2,3,7,8-Tetrachlorodibenzo-p-dioxin modifies alternative splicing in mouse liver. PLoS One 14, e0219747.
Waldron, L., Haibe-Kains, B., Culhane, A.C., Riester, M., Ding, J., Wang, X.V., Ahmadifar, M., Tyekucheva, S., Bernau, C., Risch, T., Ganzfried, B.F., Huttenhower, C., Birrer, M., Parmigiani, G., 2014. Comparative meta-analysis of prognostic gene signatures for late-stage ovarian cancer. J. Natl. Cancer Inst. 106.
Walsh, C.J., Hu, P., Batt, J., Santos, C.C.D., 2015. Microarray Meta-Analysis and Cross-Platform Normalization: Integrative Genomics for Robust Biomarker Discovery. Microarrays (Basel) 4, 389–406.
Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., Burge, C.B., 2008. Alternative isoform regulation in human tissue transcriptomes. Nature 456, 470–476.
Wang, X., Lin, Y., Song, C., Sibille, E., Tseng, G.C., 2012. Detecting disease-associated genes with confounding variable adjustment and the impact on genomic meta-analysis: With application to major depressive disorder. BMC Bioinformatics 13, 52.
Wang, Y., Liu, J., Huang, B.O., Xu, Y.-M., Li, J., Huang, L.-F., Lin, J., Zhang, J., Min, Q.-H., Yang, W.-M., Wang, X.-Z., 2015. Mechanism of alternative splicing and its regulation. Biomed Rep 3, 152–158.
Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10, 57–63.
Weirick, T., Militello, G., Müller, R., John, D., Dimmeler, S., Uchida, S., 2016. The identification and characterization of novel transcripts from RNA-seq data. Brief. Bioinform. 17, 678–685.
Welcome to GEOparse's documentation! — GEOparse 1.2.0 documentation [WWW Document], n.d. URL https://geoparse.readthedocs.io/en/latest/index.html (accessed 10.21.20).
Wen, X., Pique-Regi, R., Luca, F., 2017. Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. PLoS Genet. 13, e1006646.

Westoby, J., Artemov, P., Hemberg, M., Ferguson-Smith, A., n.d. Obstacles to Studying Alternative Splicing Using scRNA-seq.
Westoby, J., Herrera, M.S., Ferguson-Smith, A.C., Hemberg, M., 2018. Simulation-based benchmarking of isoform quantification in single-cell RNA-seq. Genome Biol. 19, 191.
Wyman, D., Balderrama-Gutierrez, G., Reese, F., Jiang, S., Rahmanian, S., Forner, S., Matheos, D., Zeng, W., Williams, B., Trout, D., England, W., Chu, S.-H., Spitale, R.C., Tenner, A.J., Wold, B.J., Mortazavi, A., 2020. A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification.
Wyman, D., Mortazavi, A., 2019. TranscriptClean: variant-aware correction of indels, mismatches and splice junctions in long-read transcripts. Bioinformatics 35, 340–342.
Xing, Y., Stoilov, P., Kapur, K., Han, A., Jiang, H., Shen, S., Black, D.L., Wong, W.H., 2008. MADS: a new and improved method for analysis of differential alternative splicing by exon-tiling microarrays. RNA 14, 1470–1479.
Ye, J., Cheng, S., Zhou, X., Chen, Z., Kim, S.U., Tan, J., Zheng, J., Xu, F., Zhang, W., Liao, Y., Zhu, Y., 2019. A global survey of full-length transcriptome of Ginkgo biloba reveals transcript variants involved in flavonoid biosynthesis. Ind. Crops Prod. 139, 111547.
Zampetaki, A., Albrecht, A., Steinhofel, K., 2018. Long Non-coding RNA Structure and Function: Is There a Link? Front. Physiol. 9, 1201.
Zhang, J., Chiodini, R., Badr, A., Zhang, G., 2011. The impact of next-generation sequencing on genomics. J. Genet. Genomics 38, 95–109.
Zhang, X., Jonassen, I., 2020. RASflow: an RNA-Seq analysis workflow with Snakemake. BMC Bioinformatics 21, 110.
Zhang, Y., Yan, L., Zeng, J., Zhou, H., Liu, H., Yu, G., Yao, W., Chen, K., Ye, Z., Xu, H., 2019. Pan-cancer analysis of clinical relevance of alternative splicing events in 31 human cancers. Oncogene 38, 6678–6695.
Zhao, L., Zhao, H., Yan, H., 2018. Gene expression profiling of 1200 pancreatic ductal adenocarcinoma reveals novel subtypes. BMC Cancer 18, 603.
Zhernakova, D.V., Deelen, P., Vermaat, M., van Iterson, M., van Galen, M., Arindrarto, W., van 't Hof, P., Mei, H., van Dijk, F., Westra, H.-J., Bonder, M.J., van Rooij, J., Verkerk, M., Jhamai, P.M., Moed, M., Kielbasa, S.M., Bot, J., Nooren, I., Pool, R., van Dongen, J., Hottenga, J.J., Stehouwer, C.D.A., van der Kallen, C.J.H., Schalkwijk, C.G., Zhernakova, A., Li, Y., Tigchelaar, E.F., de Klein, N., Beekman, M., Deelen, J., van Heemst, D., van den Berg, L.H., Hofman, A., Uitterlinden, A.G., van Greevenbroek, M.M.J., Veldink, J.H., Boomsma, D.I., van Duijn, C.M., Wijmenga, C., Slagboom, P.E., Swertz, M.A., Isaacs, A., van Meurs, J.B.J., Jansen, R., Heijmans, B.T., 't Hoen, P.A.C., Franke, L., 2017. Identification of context-dependent expression quantitative trait loci in whole blood. Nat. Genet. 49, 139–145.
Zhu, Y., Stephens, R.M., Meltzer, P.S., Davis, S.R., 2013. SRAdb: query and use public next-generation sequencing data from within R. BMC Bioinformatics 14, 19.

Bibliografía

2020 Global Initiative for Chronic Obstructive Lung Disease, I. (2020). Global strategy for the diagnosis, management, and prevention of chronic obstructive lung disease (2020 report).

Aberger, F. and i Altaba, A. R. (2014). Context-dependent signal integration by the gli code: the oncogenic load, pathways, modifiers and implications for cancer therapy. In *Seminars in cell & developmental biology*, volume 33, pages 93–104. Elsevier.

Adams, T. S., Schupp, J. C., Poli, S., Ayaub, E. A., Neumark, N., Ahangari, F., Chu, S. G., Raby, B. A., DeIuliis, G., Januszyk, M., et al. (2020). Single-cell rna-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. *Science advances*, 6(28):eaba1983.

Agusti, A. (2014). The path to personalised medicine in copd. *Thorax*, 69(9):857–864.

Agustí, A. and Hogg, J. C. (2019). Update on the pathogenesis of chronic obstructive pulmonary disease. *New England Journal of Medicine*, 381(13):1248–1256.

Arnold, B., Bowler, L., Gibson, S., Herterich, P., Higman, R., Krystalli, A., Morley, A., O'Reilly, M., Whitaker, K., et al. (2019). The turing way: A handbook for reproducible data science. *Zenodo*.

Barbas-Filho, J., Ferreira, M., Sesso, A., Kairalla, R., Carvalho, C., and Capelozzi, V. (2001). Evidence of type ii pneumocyte apoptosis in the pathogenesis of idiopathic pulmonary fibrosis (ifp)/usual interstitial pneumonia (uip). *Journal of clinical pathology*, 54(2):132–138.

Barnes, P. J., Burney, P. G. J., Silverman, E. K., Celli, B. R., Vestbo, J., Wedzicha, J. A., and Wouters, E. F. M. (2015). Chronic obstructive pulmonary disease. *Nature News*.

Barrett, K. E. (2013). *Ganong fisiología médica (24a.* McGraw Hill Mexico.

Bártholo, T. P., Porto, L. C., Pozzan, R., Nascimento, A., and Da Costa, C. H. (2019). Evaluation of hhip polymorphisms and their relationship with chronic obstructive pulmonary disease phenotypes. *International Journal of Chronic Obstructive Pulmonary Disease*, 14:2267.

Bell, O., Tiwari, V. K., Thomä, N. H., and Schübeler, D. (2011). Determinants and dynamics of genome accessibility. *Nature Reviews Genetics*, 12(8):554–564.

Boron, W. F. and Boulpaep, E. L. (2016). *Medical physiology E-book*. Elsevier Health Sciences.

Briscoe, J. and Thérond, P. P. (2013). The mechanisms of hedgehog signalling and its roles in development and disease. *Nature reviews Molecular cell biology*, 14(7):416–429.

Camelo, A., Dunmore, R., Sleeman, M. A., and Clarke, D. L. (2014). The epithelium in idiopathic pulmonary fibrosis: breaking the barrier. *Frontiers in pharmacology*, 4:173.

Chaitow, L., Bradley, D., and Gilbert, C. (2002). The structure and function of breathing. *Multidisciplinary Approaches to Breathing Pattern Disorders*, pages 1–41.

Chang, W.-A., Tsai, M.-J., Jian, S.-F., Sheu, C.-C., and Kuo, P.-L. (2018). Systematic analysis of transcriptomic profiles of copd airway epithelium using next-generation sequencing and bioinformatics. *International journal of chronic obstructive pulmonary disease*, 13:2387.

Churg, A., Cosio, M., and Wright, J. L. (2008). Mechanisms of cigarette smoke-induced copd: insights from animal models. *American Journal of Physiology-Lung Cellular and Molecular Physiology*, 294(4):L612–L631.

Cobos, F. A., Alquicira-Hernandez, J., Powell, J. E., Mestdagh, P., and De Preter, K. (2020). Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nature Communications*, 11(1):1–14.

Collado-Torres, L., Nellore, A., Kammers, K., Ellis, S. E., Taub, M. A., Hansen, K. D., Jaffe, A. E., Langmead, B., and Leek, J. T. (2017). Reproducible rna-seq analysis using recount2. *Nature biotechnology*, 35(4):319–321.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., et al. (2016). A survey of best practices for rna-seq data analysis. *Genome biology*, 17(1):1–19.

Cortes, A., Albers, P. K., Dendrou, C. A., Fugger, L., and McVean, G. (2020). Identifying cross-disease components of genetic risk across hospital data in the uk biobank. *Nature genetics*, 52(1):126–134.

Costa, V., Aprile, M., Esposito, R., and Ciccodicola, A. (2013). Rna-seq and human complex diseases: recent accomplishments and future perspectives. *European Journal of Human Genetics*, 21(2):134–142.

Cramer, P. (2019). Organization and regulation of gene transcription. *Nature*, 573(7772):45–54.

De Rose, V., Molloy, K., Gohy, S., Pilette, C., and Greene, C. M. (2018). Airway epithelium dysfunction in cystic fibrosis and copd. *Mediators of inflammation*, 2018.

DeMeo, D. and Silverman, E. (2004). α1-antitrypsin deficiency · 2: Genetic aspects of α1-antitrypsin deficiency: phenotypes and genetic modifiers of emphysema risk. *Thorax*, 59(3):259–264.

di Magliano, M. P. and Hebrok, M. (2003). Hedgehog signalling in cancer formation and maintenance. *Nature reviews cancer*, 3(12):903–911.

Dransfield, M. T., Voelker, H., Bhatt, S. P., Brenner, K., Casaburi, R., Come, C. E., Cooper, J. A. D., Criner, G. J., Curtis, J. L., Han, M. K., Hatipoğlu, U., Helgeson, E. S., Jain, V. V., Kalhan, R., Kaminsky, D., Kaner, R., Kunisaki, K. M., Lambert, A. A., Lammi, M. R., Lindberg, S., Make, B. J., Martinez, F. J., McEvoy, C., Panos, R. J., Reed, R. M., Scanlon, P. D., Sciurba, F. C., Smith, A., Sriram, P. S., Stringer, W. W., Weingarten, J. A., Wells, J. M., Westfall, E., Lazarus, S. C., Connett, J. E., and BLOCK COPD Trial Group

(2019). Metoprolol for the prevention of acute exacerbations of COPD. *N. Engl. J. Med.*, 381(24):2304–2314.

Engelen, K., Fu, Q., Meysman, P., Sánchez-Rodríguez, A., De Smet, R., Lemmens, K., Fierro, A. C., and Marchal, K. (2011). COLOMBOS: access port for cross-platform bacterial expression compendia. *PLoS One*, 6(7):e20938.

Feghali-Bostwick, C. A., Gadgil, A. S., Otterbein, L. E., Pilewski, J. M., Stoner, M. W., Csizmadia, E., Zhang, Y., Sciurba, F. C., and Duncan, S. R. (2008). Autoantibodies in patients with chronic obstructive pulmonary disease. *American journal of respiratory and critical care medicine*, 177(2):156–163.

Franks, T. J., Colby, T. V., Travis, W. D., Tuder, R. M., Reynolds, H. Y., Brody, A. R., Cardoso, W. V., Crystal, R. G., Drake, C. J., Engelhardt, J., et al. (2008). Resident cellular components of the human lung: current knowledge and goals for research on cell phenotyping and function. *Proceedings of the American Thoracic Society*, 5(7):763–766.

Ganser, L. R., Kelly, M. L., Herschlag, D., and Al-Hashimi, H. M. (2019). The roles of structural dynamics in the cellular functions of rnas. *Nature Reviews Molecular Cell Biology*, 20(8):474–489.

Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). affy—analysis of affymetrix genechip data at the probe level. *Bioinformatics*, 20(3):307–315.

Gerstein, M. B., Bruce, C., Rozowsky, J. S., Zheng, D., Du, J., Korbel, J. O., Emanuelsson, O., Zhang, Z. D., Weissman, S., and Snyder, M. (2007). What is a gene, post-encode? history and updated definition. *Genome research*, 17(6):669–681.

Gimond, M. (consulta 2021).

Golpon, H. A., Coldren, C. D., Zamora, M. R., Cosgrove, G. P., Moore, M. D., Tuder, R. M., Geraci, M. W., and Voelkel, N. F. (2004). Emphysema lung tissue gene expression profiling. *American journal of respiratory cell and molecular biology*, 31(6):595–600.

Gordon, G. J., Jensen, R. V., Hsiao, L.-L., Gullans, S. R., Blumenstock, J. E., Ramaswamy, S., Richards, W. G., Sugarbaker, D. J., and Bueno, R. (2002). Translation of microarray

data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer research*, 62(17):4963–4967.

Graham, B. L., Steenbruggen, I., Miller, M. R., Barjaktarevic, I. Z., Cooper, B. G., Hall, G. L., Hallstrand, T. S., Kaminsky, D. A., McCarthy, K., McCormack, M. C., et al. (2019). Standardization of spirometry 2019 update. an official american thoracic society and european respiratory society technical statement. *American journal of respiratory and critical care medicine*, 200(8):e70–e88.

Green, C. E. and Turner, A. M. (2017). The role of the endothelium in asthma and chronic obstructive pulmonary disease (copd). *Respiratory research*, 18(1):20.

Grieco, D. L., Chen, L., and Brochard, L. (2017). Transpulmonary pressure: importance and limits. *Annals of translational medicine*, 5(14).

Halbert, R., Natoli, J., Gano, A., Badamgarav, E., Buist, A. S., and Mannino, D. (2006). Global burden of copd: systematic review and meta-analysis. *European Respiratory Journal*, 28(3):523–532.

Ham, S., Oh, Y.-M., and Roh, T.-Y. (2019). Evaluation and interpretation of transcriptome data underlying heterogeneous chronic obstructive pulmonary disease. *Genomics Inform.*, 17(1):e2.

Hammer, G. D., McPhee, S. J., Bari, S. M. O., and Muñoz, B. R. (2015). *Fisiopatología de la enfermedad: una introducción a la medicina clínica*. McGraw-Hill Education.

Heid, C. A., Stevens, J., Livak, K. J., and Williams, P. M. (1996). Real time quantitative pcr. *Genome research*, 6(10):986–994.

Horowitz, J. C., Martinez, F. J., and Thannickal, V. J. (2009). Mesenchymal cell fate and phenotypes in the pathogenesis of emphysema. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, 6(3):201–210.

Hughes, J. and Bates, D. (2003). Historical review: The carbon monoxide diffusing capacity (dlco) and its membrane (dm) and red cell ($\theta \cdot$ vc) components. *Respiratory physiology & neurobiology*, 138(2-3):115–142.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264.

Ito, K. and Barnes, P. J. (2009). Copd as a disease of accelerated lung aging. *Chest*, 135(1):173–180.

Jacob, L., Gagnon-Bartsch, J. A., and Speed, T. P. (2016). Correcting gene expression data when neither the unwanted variation nor the factor of interest are observed. *Biostatistics*, 17(1):16–28.

Jensen, F. B. (2004). Red blood cell ph, the bohr effect, and other oxygenation-linked phenomena in blood o2 and co2 transport. *Acta Physiologica Scandinavica*, 182(3):215–227.

Kanehisa, M. and Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30.

Kassambara, A. (2018). ggpubr:'ggplot2'based publication ready plots. r package version 0.2. *Retrieved September*, 12:2020.

Kim, W. J., Lim, J. H., Lee, J. S., Lee, S.-D., Kim, J. H., and Oh, Y.-M. (2015). Comprehensive analysis of transcriptome sequencing data in the lung tissues of COPD subjects. *Int. J. Genomics Proteomics*, 2015:206937.

Kim, W. J., Oh, Y.-M., Lee, J. H., Park, C.-S., Park, S. W., Park, J. S., and Lee, S. D. (2013). Genetic variants in hhip are associated with fev1 in subjects with chronic obstructive pulmonary disease. *Respirology*, 18(8):1202–1209.

Kirkham, P. A. and Barnes, P. J. (2013). Oxidative stress in copd. *Chest*, 144(1):266–273.

Kitamura, Y. and Ito, A. (2005). Mast cell-committed progenitors. *Proceedings of the National Academy of Sciences*, 102(32):11129–11130.

Kori, M. and Yalcin Arga, K. (2018). Potential biomarkers and therapeutic targets in cervical cancer: Insights from the meta-analysis of transcriptomics data within network biomedicine perspective. *PloS one*, 13(7):e0200717.

Krogh, M. (1915). The diffusion of gases through the lungs of man. *The Journal of physiology*, 49(4):271–300.

Kulkarni, T., O'Reilly, P., Antony, V. B., Gaggar, A., and Thannickal, V. J. (2016). Matrix remodeling in pulmonary fibrosis and emphysema. *American journal of respiratory cell and molecular biology*, 54(6):751–760.

Lao, T., Jiang, Z., Yun, J., Qiu, W., Guo, F., Huang, C., Mancini, J. D., Gupta, K., Laucho-Contreras, M. E., Naing, Z. Z. C., et al. (2016). Hhip haploinsufficiency sensitizes mice to age-related emphysema. *Proceedings of the National Academy of Sciences*, 113(32):E4681–E4687.

Liang, K.-H. (2013). *Bioinformatics for biomedical science and clinical applications*. Elsevier.

Lowe, R., Shirley, N., Bleackley, M., Dolan, S., and Shafee, T. (2017). Transcriptomics technologies. *PLoS computational biology*, 13(5):e1005457.

Lundbäck, B., Lindberg, A., Lindström, M., Rönmark, E., Jonsson, A.-C., Jönsson, E., Larsson, L.-G., Andersson, S., Sandström, T., and Larsson, K. (2003). Not 15 but 50% of smokers develop copd?—report from the obstructive lung disease in northern sweden studies. *Respiratory medicine*, 97(2):115–122.

Lutz, S. M., Cho, M. H., Young, K., Hersh, C. P., Castaldi, P. J., McDonald, M.-L., Regan, E., Mattheisen, M., DeMeo, D. L., Parker, M., et al. (2015). A genome-wide association study identifies risk loci for spirometric measures among smokers of european and african ancestry. *BMC genetics*, 16(1):1–11.

MacNee, W. (2019). Beta-Blockers in COPD - a controversy resolved? *N. Engl. J. Med.*, 381(24):2367–2368.

Malacarne, G., Pilati, S., Valentini, S., Asnicar, F., Moretto, M., Sonego, P., Masera, L., Cavecchia, V., Blanzieri, E., and Moser, C. (2018). Discovering causal relationships in grapevine expression data to expand gene networks. a case study: Four networks related to climate change. *Front. Plant Sci.*, 9:1385.

Martinez, F. J., Collard, H. R., Pardo, A., Raghu, G., Richeldi, L., Selman, M., Swigris, J. J.,

Taniguchi, H., and Wells, A. U. (2017). Idiopathic pulmonary fibrosis. *Nature reviews Disease primers*, 3(1):1–19.

Mathai, S. K., Newton, C. A., Schwartz, D. A., and Garcia, C. K. (2016). Pulmonary fibrosis in the era of stratified medicine. *Thorax*, 71(12):1154–1160.

McGettigan, P. A. (2013). Transcriptomics in the rna-seq era. *Current opinion in chemical biology*, 17(1):4–11.

Meiners, S., Eickelberg, O., and Königshoff, M. (2015). Hallmarks of the ageing lung. *European Respiratory Journal*, 45(3):807–827.

Meysman, P., Sonego, P., Bianco, L., Fu, Q., Ledezma-Tejeida, D., Gama-Castro, S., Liebens, V., Michiels, J., Laukens, K., Marchal, K., Collado-Vides, J., and Engelen, K. (2014). COLOMBOS v2.0: an ever expanding collection of bacterial expression compendia. *Nucleic Acids Res.*, 42(Database issue):D649–53.

Milward, E., Shahandeh, A., Heidari, M., Johnstone, D., Daneshi, N., and Hondermarck, H. (2016). Transcriptomics. *Introduction to Bioinformatics in Microbiology*, pages 160–165.

Mora, A. L., Rojas, M., Pardo, A., and Selman, M. (2017). Emerging therapies for idiopathic pulmonary fibrosis, a progressive age-related disease. *Nature reviews Drug discovery*, 16(11):755.

Moretto, M., Sonego, P., Dierckxsens, N., Brilli, M., Bianco, L., Ledezma-Tejeida, D., Gama-Castro, S., Galardini, M., Romualdi, C., Laukens, K., Collado-Vides, J., Meysman, P., and Engelen, K. (2016a). COLOMBOS v3.0: leveraging gene expression compendia for cross-species analyses. *Nucleic Acids Res.*, 44(D1):D620–3.

Moretto, M., Sonego, P., Pilati, S., Malacarne, G., Costantini, L., Grzeskowiak, L., Bagagli, G., Grando, M. S., Moser, C., and Engelen, K. (2016b). VESPUCCI: Exploring patterns of gene expression in grapevine. *Front. Plant Sci.*, 7:633.

Moretto, M., Sonego, P., Villaseñor-Altamirano, A. B., and Engelen, K. (2019). First step toward gene expression data integration: transcriptomic data acquisition with COMMAND> _. *BMC Bioinformatics*, 20(1):54.

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., Hoang, C. D., Diehn, M., and Alizadeh, A. A. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nature methods*, 12(5):453–457.

Olajuyin, A. M., Zhang, X., and Ji, H.-L. (2019). Alveolar type 2 progenitor cells for lung injury repair. *Cell death discovery*, 5(1):1–11.

Ong, C.-T. and Corces, V. G. (2014). Ctcf: an architectural protein bridging genome topology and function. *Nature Reviews Genetics*, 15(4):234–246.

Orozco-Levi, M., Garcia-Aymerich, J., Villar, J., Ramirez-Sarmiento, A., Anto, J., and Gea, J. (2006). Wood smoke exposure and risk of chronic obstructive pulmonary disease. *European Respiratory Journal*, 27(3):542–546.

Paraskeva, M. A., Borg, B. M., Naughton, M. T., et al. (2011). Spirometry. *Australian family physician*, 40(4):216.

Pardo, A., Cabrera, S., Maldonado, M., and Selman, M. (2016). Role of matrix metalloproteinases in the pathogenesis of idiopathic pulmonary fibrosis. *Respiratory research*, 17(1):23.

Pearson, H. (2006). What is a gene?

Petty, T. L. (2006). The history of copd. *International journal of chronic obstructive pulmonary disease*, 1(1):3.

Pierce, R. et al. (2005a). Spirometry: an essential clinical measurement. *Australian family physician*, 34(7):535.

Pierce, R. J., Hillman, D., Young, I. H., O'donoghue, F., Zimmerman, P. V., West, S., and Burdon, J. G. (2005b). Respiratory function tests and their application. *Respirology*, 10:S1–S19.

Pilati, S., Bagagli, G., Sonego, P., Moretto, M., Brazzale, D., Castorina, G., Simoni, L., Tonelli, C., Guella, G., Engelen, K., Galbiati, M., and Moser, C. (2017). Abscisic acid is a major regulator of grape berry ripening onset: New insights into ABA signaling network. *Front. Plant Sci.*, 8:1093.

Portin, P. and Wilkins, A. (2017). The evolving definition of the term "gene". *Genetics*, 205(4):1353–1364.

Raghu, G., Remy-Jardin, M., Myers, J. L., Richeldi, L., Ryerson, C. J., Lederer, D. J., Behr, J., Cottin, V., Danoff, S. K., Morell, F., et al. (2018). Diagnosis of idiopathic pulmonary fibrosis. an official ats/ers/jrs/alat clinical practice guideline. *American journal of respiratory and critical care medicine*, 198(5):e44–e68.

Ramos, C., Montaño, M., García-Alvarez, J., Ruiz, V., Uhal, B. D., Selman, M., and Pardo, A. (2001). Fibroblasts from idiopathic pulmonary fibrosis and normal lungs differ in growth rate, apoptosis, and tissue inhibitor of metalloproteinases expression. *American journal of respiratory cell and molecular biology*, 24(5):591–598.

Ramos, F. L., Krahnke, J. S., and Kim, V. (2014). Clinical issues of mucus accumulation in copd. *International journal of chronic obstructive pulmonary disease*, 9:139.

Rennard, S. I. and Drummond, M. B. (2015). Early chronic obstructive pulmonary disease: definition, assessment, and prevention. *The Lancet*, 385(9979):1778–1788.

Reyfman, P. A., Walter, J. M., Joshi, N., Anekalla, K. R., McQuattie-Pimentel, A. C., Chiu, S., Fernandez, R., Akbarpour, M., Chen, C.-I., Ren, Z., et al. (2019). Single-cell transcriptomic analysis of human lung provides insights into the pathobiology of pulmonary fibrosis. *American journal of respiratory and critical care medicine*, 199(12):1517–1536.

Risso, D., Ngai, J., Speed, T. P., and Dudoit, S. (2014). Normalization of rna-seq data using factor analysis of control genes or samples. *Nature biotechnology*, 32(9):896–902.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47.

Ritchie, M. E., Silver, J., Oshlack, A., Holmes, M., Diyagama, D., Holloway, A., and Smyth, G. K. (2007). A comparison of background correction methods for two-colour microarrays. *Bioinformatics*, 23(20):2700–2707.

Rutgers, S. R., Postma, D. S., ten Hacken, N. H., Kauffman, H. F., van der Mark, T. W.,

Koëter, G. H., and Timens, W. (2000). Ongoing airway inflammation in patients with copd who do not currently smoke. *Thorax*, 55(1):12–18.

Santana-Garcia, W., Rocha-Acevedo, M., Ramirez-Navarro, L., Mbouamboua, Y., Thieffry, D., Thomas-Chollier, M., Contreras-Moreira, B., van Helden, J., and Medina-Rivera, A. (2019). Rsat variation-tools: An accessible and flexible framework to predict the impact of regulatory variants on transcription factor binding. *Computational and structural biotechnology journal*, 17:1415–1428.

Sauler, M., McDonough, J. E., Adams, T. S., Kothapalli, N., Schupp, J. S., Nouws, J., Chioccioli, M., Omote, N., Cosme, C., Poli, S., et al. (2020). Single-cell rna sequencing identifies aberrant transcriptional profiles of cellular populations and altered alveolar niche signalling networks in chronic obstructive pulmonary disease (copd). *medRxiv*.

Scano, G., Innocenti-Bruni, G., and Stendardi, L. (2010). Do obstructive and restrictive lung diseases share common underlying mechanisms of breathlessness? *Respiratory medicine*, 104(7):925–933.

Selman, M., Martinez, F. J., and Pardo, A. (2019). Why does an aging smoker's lung develop idiopathic pulmonary fibrosis and not chronic obstructive pulmonary disease? *American Journal of Respiratory and Critical Care Medicine*, 199(3):279–285.

Selman, M. and Pardo, A. (2014). Revealing the pathogenic and aging-related mechanisms of the enigmatic idiopathic pulmonary fibrosis. an integral model. *American journal of respiratory and critical care medicine*, 189(10):1161–1172.

Sheila-10x (2017). Single-cell rna-seq: An introductory overview and tools for getting started - 10x genomics. https://www.10xgenomics.com/blog/single-cell-rna-seq-an-introductory-overview-and-tools-for-getting-started. Accessed: 2020-12-7.

Silverman, E. K. (2020). Genetics of copd. *Annual review of physiology*, 82:413–431.

SILVERMAN, E. K., WEISS, S. T., DRAZEN, J. M., CHAPMAN, H. A., CAREY, V., CAMPBELL, E. J., DENISH, P., SILVERMAN, R. A., CELEDON, J. C., REILLY, J. J., et al. (2000). Gender-related differences in severe, early-onset chronic obstructive

pulmonary disease. *American journal of respiratory and critical care medicine*, 162(6):2152–2158.

Spira, A., Beane, J., Pinto-Plata, V., Kadar, A., Liu, G., Shah, V., Celli, B., and Brody, J. S. (2004a). Gene expression profiling of human lung tissue from smokers with severe emphysema. *American journal of respiratory cell and molecular biology*, 31(6):601–610.

Spira, A., Beane, J., Shah, V., Liu, G., Schembri, F., Yang, X., Palma, J., and Brody, J. S. (2004b). Effects of cigarette smoke on the human airway epithelial cell transcriptome. *Proceedings of the National Academy of Sciences*, 101(27):10143–10148.

Tam, A., Hughes, M., McNagny, K., Obeidat, M., Hackett, T., Leung, J., Shaipanich, T., Dorscheid, D., Singhera, G., Yang, C., et al. (2019). Hedgehog signaling in the airway epithelium of patients with chronic obstructive pulmonary disease. *Scientific reports*, 9(1):1–13.

Taylor, R. (1990). Interpretation of the correlation coefficient: a basic review. *Journal of diagnostic medical sonography*, 6(1):35–39.

Togo, S., Holz, O., Liu, X., Sugiura, H., Kamio, K., Wang, X., Kawasaki, S., Ahn, Y., Fredriksson, K., Skold, C. M., et al. (2008). Lung fibroblast repair functions in patients with chronic obstructive pulmonary disease are altered by multiple mechanisms. *American journal of respiratory and critical care medicine*, 178(3):248–260.

Travaglini, K. J., Nabhan, A. N., Penland, L., Sinha, R., Gillich, A., Sit, R. V., Chang, S., Conley, S. D., Mori, Y., Seita, J., et al. (2020). A molecular cell atlas of the human lung from single-cell rna sequencing. *Nature*, 587(7835):619–625.

Uhal, B. (2008). The role of apoptosis in pulmonary fibrosis. *European Respiratory Review*, 17(109):138–144.

Van de Sande, B., Flerin, C., Davie, K., De Waegeneer, M., Hulselmans, G., Aibar, S., Seurinck, R., Saelens, W., Cannoodt, R., Rouchon, Q., et al. (2020). A scalable scenic workflow for single-cell gene regulatory network analysis. *Nature Protocols*, 15(7):2247–2276.

Villaseñor-Altamirano, A. B., Moretto, M., Maldonado, M., Zayas-Del Moral, A., Munguía-Reyes, A., Romero, Y., García-Sotelo, J. S., Aguilar, L. A., Aldana-Assad, O., Engelen, K., et al. (2020). Pulmondb: a curated lung disease gene expression database. *Scientific reports*, 10(1):1–9.

Wang, C., de Mochel, N. S. R., Christenson, S. A., Cassandras, M., Moon, R., Brumwell, A. N., Byrnes, L. E., Li, A., Yokosaki, Y., Shan, P., et al. (2018). Expansion of hedgehog disrupts mesenchymal identity and induces emphysema phenotype. *The Journal of clinical investigation*, 128(10):4343–4358.

Wang, X., Park, J., Susztak, K., Zhang, N. R., and Li, M. (2019). Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature communications*, 10(1):1–9.

Wang, Z., Gerstein, M., and Snyder, M. (2009). Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63.

Wedzicha, J. A. (2000). The heterogeneity of chronic obstructive pulmonary disease. *Thorax*, 55(8):631–632.

Wen, L., Krauss-Etschmann, S., Petersen, F., and Yu, X. (2018). Autoantibodies in chronic obstructive pulmonary disease. *Frontiers in immunology*, 9:66.

Wright, J. L. and Churg, A. (2002). Animal models of cigarette smoke-induced copd. *Chest*, 122(6):301S–306S.

Xie, T., Wang, Y., Deng, N., Huang, G., Taghavifar, F., Geng, Y., Liu, N., Kulur, V., Yao, C., Chen, P., et al. (2018). Single-cell deconvolution of fibroblast heterogeneity in mouse pulmonary fibrosis. *Cell reports*, 22(13):3625–3640.

Yadav, D., Tanveer, A., Malviya, N., and Yadav, S. (2018). Overview and principles of bioengineering: the drivers of omics technologies. In *Omics Technologies and Bio-Engineering*, pages 3–23. Elsevier.

Zepp, J. A. and Morrisey, E. E. (2019). Cellular crosstalk in the development and regeneration of the respiratory system. *Nature Reviews Molecular Cell Biology*, 20(9):551–566.

Zhang, Z., Wang, J., Zheng, Z., Chen, X., Zeng, X., Zhang, Y., Li, D., Shu, J., Yang, K., Lai, N., et al. (2017). Genetic variants in the hedgehog interacting protein gene are associated with the fev1/fvc ratio in southern han chinese subjects with chronic obstructive pulmonary disease. *BioMed research international*, 2017.

Zhou, X., Baron, R. M., Hardin, M., Cho, M. H., Zielinski, J., Hawrylkiewicz, I., Sliwinski, P., Hersh, C. P., Mancini, J. D., Lu, K., et al. (2012). Identification of a chronic obstructive pulmonary disease genetic determinant that regulates hhip. *Human molecular genetics*, 21(6):1325–1335.