Universidad Nacional Autónoma de México

ENES Juriquilla

Identificación y caracterización de redes de regulación que modulan la respuesta transcripcional a infección por SARS-CoV-2 en humano

T E S I S

QUE PARA OPTAR POR EL GRADO DE:

Licenciada en Ciencias Genómicas

PRESENTA:

Mónica Padilla Gálvez

TUTOR:

Dra. Alejandra E. Medina Rivera

Juriquilla, Querétaro, 2023

A mis padres

Reconocimientos

Quisiera reconocer a mi tutora, la Dra. Alejandra Medina Rivera, quién con paciencia y diligencia discutió conmigo los detalles no sólo de la tesis presente, sino de mi camino en la ciencia, así como me ha dado tantas oportunidades para formarlo. Por ser un gran ejemplo a seguir, gracias. Asimismo, reconocer a mis sinodales, la Dra. Ana Beatriz Villaseñor Altamirano, quien co-asesoró esta tesis y me ha ayudado cuando lo he necesitado; a la Dra. Yalbi I. Balderas, al Dr. Daniel Blanco Melo y al Dr. Julio Collado Vides, cuyas detalladas revisiones y preguntas mejoraron ampliamente este trabajo. También quiero agradecerle al Dr. Javier De Las Rivas, cuya asesoría mejoró la metodología para el análisis de resultados y me recibió con gusto en su laboratorio. Al Dr. Leonardo Collado Torres por sugerir un experimento que mejoró este trabajo. Quiero agradecer también a los miembros presentes y anteriores del RegGenoLab, cuya retroalimentación y apoyo han servido mucho a lo largo de este trayecto. A Karen, por darme la oportunidad de aprender de ella cuando iba iniciando mi camino; a Paula, Ana Lau y Sofia, mis amigas, a Evelia, Walter, Lucía, Leo, Domingo y Olga.

Agradezco a Margareta, tu apoyo ha sido indispensable en todo el camino.

Agradezco sumamente al personal técnico del Laboratorio Internacional de Investigación sobre el Genoma Humano (LIIGH): Jair García Sotelo, por salvar mi computadora múltiples veces (y por lo tanto, este trabajo), y del Laboratorio Nacional de Visualización Científica Avanzada (LAVIS) a Luis Aguilar, por apoyarme tanto a lo largo de este proyecto. Asimismo, agradezco el respaldo económico del Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT) de la Universidad Nacional Autónoma de México (UNAM) con el grant número IA203021, a la ENES Juriquilla y sus administrativos, y a la UNAM en general.

Agradezco a mi familia y amigos. A mis padres, por siempre estar, creer en mí y darme su apoyo que hicieron de mi formación una realidad. A mi hermana y hermanos. A mis abuelas, de inquebrantable espíritu y fuerza. A Pinky, mi compañera peluda. A mis amigos, por su invaluable compañía y apoyo, en especial a Soph, mi mejor amiga, a Victor, a Fernando y a Xoch.

Gracias a quienes me acompañaron, y a quienes me compartieron un poco de su conocimiento y tiempo. Soy lo que pude rescatar en mí de ustedes.

A todos, de corazón, gracias.

Declaración de autenticidad

Por la presente declaro que, salvo cuando se haga referencia específica al trabajo de otras personas, el contenido de esta tesis es original y no se ha presentado total o parcialmente para su consideración para cualquier otro título o grado en esta o cualquier otra Universidad. Esta tesis es resultado de mi propio trabajo y no incluye nada que sea el resultado de algún trabajo realizado en colaboración, salvo que se indique específicamente en el texto.

Mónica Padilla Gálvez. Juriquilla, Querétaro, 2023

Resumen

Es necesario tener un mejor entendimiento sobre la fisiopatología que da lugar a la enfermedad del Coronavirus (COVID-19). Se han identificado los componentes principales que llevan de la infección por SARS-CoV-2 a las complicaciones dentro de las células y tejidos humanos que culminan en esta enfermedad (explícitamente, el receptor ACE2 y cofactores como TMPRSS2, interferón y la cascada de citoquinas (Delorey et al. (15), Tanaka et al. (59), Yang et al. (70))). Sin embargo, no se ha llegado a un consenso sobre cómo se ensamblan estos elementos en organizaciones celulares conocidas como redes de regulación genéticas (GRNs, por sus siglas en inglés). Las GRNs son responsables de aspectos fundamentales en sistemas biológicos al organizar la respuesta celular a señales del ambiente y de la progresión de enfermedades (Singh et al. (57)). Adicionalmente, no se ha visto extensamente cómo estas redes encendidas por la infección por SARS-CoV-2 cambian entre los tejidos afectados (Wanner et al. (64)). Más aún, generar GRNs nos permitirá evaluar su reproducibilidad entre estudios y casos. En el presente trabajo de tesis, se realizó un análisis de redes de regulación genéticas integrativo y robusto, donde de manera simultánea se explora la infección por SARS-CoV-2 en comparación con la infección de otros virus que afectan el sistema respiratorio (el virus respiratorio sincicial, la influenza y la parainfluenza), y COVID-19 a múltiples tejidos u órganos. Se integraron datos de distintas fuentes y tratando de llevar a cabo un procesamiento homogéneo. Para cada condición de infección o de tejido enfermo, se buscó la activación diferencial y se evaluó la especificidad de los regulones descubiertos, elucidando así los conjuntos específicos de factores transcripcionales (TFs) críticos con sus respectivos genes diana. Como se esperaba, al comparar con la respuesta a otros virusm se encontró que los regulones sobre-activados y específicamente encendidos en la infección de SARS-CoV-2, dirigen una respuesta pro-inflamatoria. De manera interesante, entre ellos se encontraron regulones novedosos guiados por los Factores Transcripcionales: ZNF595 y FOXP4. Por otra parte, el análisis de tejidos mostró que hay 195 regulones compartidos en al menos 3 tejidos, donde aquellos con una mayor intersección son pulmón, corazón e hígado; mostrando que los procesos regulatorios compartidos podrían estar dando lugar a las manifestaciones extra pulmonares del COVID-19. En conjunto, el estudio presente reproduce descubrimientos anteriores y provee mayor entendimiento a los mecanismos regulatorios que instigan el COVID-19 que podrían ser relevantes para el desarrollo de tratamientos alternativos para pacien-

tes. Además, compartimos dos herramientas bioinformáticas para el análisis de redes de regulación: (1) la *pipeline* de procesamiento de datos de secuenciación del ácido ribonucleico (ARN) para la generación y el análisis de GRNs, y (2) $network - interactions$ para la generación y comparación de GRNs.

Índice general

Capítulo 1

Introducción

1.1. Presentación

La pandemia de COVID-19 es causada por SARS-CoV-2. Una mayor comprensión del mecanismo de infección y la fisiopatología que conlleva esta enfermedad dentro de las células y entre tejidos es necesaria para la elaboración de tratamientos alternativos. Actualmente se han identificado los principales agentes en este proceso: el receptor ACE2, cofactores como TMPRSS2, interferón y la subsecuente cascada de citocinas que provocan la exacerbada respuesta inflamatoria vista en pacientes. Sin embargo, no se ha llegado a un consenso sobre cómo estos elementos se ensamblan en redes de regulación génicas y varian según el tejido afectado.

1.2. Objetivo

El objetivo general de este trabajo es la elaboración de una *pipeline* de procesamiento de datos de RNA-seq para el análisis de redes de regulación de genes (GRNs, por sus siglas en inglés) y el subsecuente análisis de las redes generadas. Para probar la *pipeline* bioinformática desarrollada, se exploran las GRNs que aparecen ante la infección de SARS-CoV-2.

Los objetivos particulares son:

- Recuperar datos de RNA-seq de distintas fuentes y estandarizar un flujo de trabajo para procesar y analizar estos datos.
- Descubrir redes de regulación que modulan la expresión de los genes humanos que participan durante la respuesta inmune a SARS-CoV-2.
- Conocer cuáles son las GRNs que se presentan de manera específica en comparación con otros virus para comprender a mayor nivel el proceso fisiopatológico de la enfermedad derivada de la infección.

- Discernir las GRNs que se presentan en distintos tejidos para comprender tanto las distintas afectaciones que pueda haber como los procesos biológicos en común con pulmón.

1.3. Motivación

A partir del inicio de la pandemia de COVID-19, el número de estudios relacionados con SARS-CoV-2 creció exponencialmente. Esto ante la gran emergencia mundial para tratar la enfermedad, parar las infecciones y contender la pandemia. Dentro de estos, surgieron un número de estudios que poco a poco señalaron los principales factores responsables de la enfermedad, las proteínas que interactúan con el virus, la cascada de citocinas y el proceso inmunológico (Blanco-Melo et al. (4), Delorey et al. (15), Desai et al. (16), Liao et al. (42), y otros), estudios que proveyeron datos públicos con potencial de uso para preguntas biológicas sin responder.

¿Cómo se conectan estos procesos en la célula?, y de manera particular, ¿qué redes de regulación genéticas colaboran y engloban estos procesos durante la infección?

Como se describe en el Marco Teórico de la tesis presente (véanse secciones 2.2 y 2.4.1), el perfil transcripcional de las células es descriptivo de su identidad y estado. Al ser las redes de regulación génicas capaces de subordinar procesos fisiológicos, el conocer cuáles redes de regulación se activan y cuáles son los Factores Transcripcionales (TFs, por sus siglas en inglés) clave, nos permitirá entender la fisipatología de la infección.

Algunos estudios exploraron esta pregunta con distintos enfoques, datos y métodos de construcción de redes (Liao et al. (42), Wanner et al. (64), Tanaka et al. (59), Jung et al. (32), Chua et al. (10)). Dados estos distintos factores notamos que las intersecciones de resultados entre las redes generadas eran pequeñas, por lo que en el estudio presente tuvimos la motivación de (1) aprovechar los datos de RNA-seq provenientes de distintos estudios generados por la comunidad científica, (2) generar redes de regulación bajo un procesamiento uniforme para llegar a resultados consistentes y comparables, y (3) desarrollar una *pipeline* reproducible para este fin y compartirla con la comunidad científica.

1.4. Planteamiento del problema

Como se mencionó en la sección anterior, se tuvo el interés de responder a la pregunta: ¿qué redes de regulación genéticas, formadas por los TFs y sus genes diana, colaboran y engloban los distintos procesos celulares que se activan tras la infección de SARS-CoV-2?

Adicionalmente, se buscó encontrar cuáles redes se activan exclusivamente por infección de SARS-CoV-2 en comparación con otros virus y cuáles se activan en comparación con tejido pulmonar, en otros órganos afectados cuyos datos transcriptómicos estuvieran disponibles (se encontraron datos de corazón, el hígado, los riñones y el intestino)

lo cuál nos puede dar un panorama general de la diferencia de subprocesos celulares que están ocurriendo en los distintos contextos.

1.5. Metodología

Se utilizaron datos transcriptómicos generados por la tecnología de secuenciación de RNA (conocido comúnmente como RNA-seq) procedentes de experimentos de infección en líneas celulares por SARS-CoV-2 y otros virus (como punto de comparación) y paralelamente, se recuperaron datos de RNA-seq procedentes de biopsias de distintos órganos de pacientes con COVID-19 o sanos: pulmón, corazón, hígado, intestino y riñón.

Se llevó a cabo el procesamiento de datos de RNA-seq de la manera más homogénea posible: revisión de calidad, cortado de secuencias con mala calidad, pseudoalineamiento y conteo de transcritos usando *kallisto* (algoritmo similar al usado en datos descargados como matrices de cuentas de genes), se hizo la corrección de efectos de *batch* por set de datos y se llevó a cabo una segunda corrección al combinar *datasets* y se hizo el filtrado de genes (descrito en la sección 3.2)

Posteriormente se dividió el análisis en tres partes paralelas: (1) Análisis de líneas celulares de epitelio pulmonar infectado con virus afectando vías respiratorias, (2) Análisis de COVID-19 distintos tejidos y (3) Análisis de Pulmón con COVID-19 en comparación con sano.

Se revisó la corrección de *batches* y se visualizó el perfil transcriptómico de los datos procesados con PCA.

Posteriormente, se generaron las redes de regulación utilizando la *pipeline* de pySCENIC cien veces por análisis. De los regulones resultantes, se seleccionaron aquellos que estaban diferencialmente activados en los casos de infección y que estaban altamente presentes en esas muestras.

Se visualizaron y compararon los regulones presentes en cada condición y para finalizar, se realizó el análisis de enriquecimiento de términos biológicos en esos regulones que resultaron de mayor interés.

1.6. Contribuciones

Las principales contribuciones de este trabajo se dividen en dos. Por una parte son los resultados directos del análisis del caso biológico estudiado aquí; la identificación de los regulones más relevantes en la infección de SARS-CoV-2 y la identificación de los regulones compartidos en los órganos más afectados por el COVID-19: pulmón, corazón e hígado. Y por otra parte, la generación de herramientas bioinformáticas para analizar redes de regulación: (1) la *pipeline* de procesamiento de datos de RNA-seq para el análisis de redes de regulación de genes y (2) la herramienta $network - interactions$ en RSAT.

Capítulo 2

Marco teórico

2.1. La pandemia causada por el SARS-CoV-2

En Diciembre del 2019 en Wuhan, China se identificó el virus conocido como SARS-CoV-2 (llamado así por causar el síndrome respiratorio agudo severo (SARS) y pertenecer a la familia de virus Coronavirus (CoV), teniendo además un gran parecido genético con el SARS-CoV), que a inicios del 2020 ocasionó una pandemia. A la enfermedad que causa se le conoce como COVID-19. Hasta Febrero del 2023, la Organización Mundial de la Salud ha confirmado más de 757 millones de casos y más de 6 millones 850 mil muertes (who (2)), indicando que la tasa de mortalidad de casos es mayor al 0.9 %. Al rededor del 3 al 20 % de los pacientes con COVID-19 requieren hospitalización y de ellos, alrededor del 10 al 30 % requieren cuidado intensivo (Lamers and Haagmans (41)). Un mejor entendimiento de la progresión patogénica de esta enfermedad es necesaria para el desarrollo de terapias efectivas, para lo cuál creemos que entender la regulación genética de las células infectadas, como se elabora a lo largo de este trabajo, será de ayuda.

La sintomatología desencadenada a partir de la infección por SARS-CoV-2 ha comprendido el rango de asintomática a letal; provocando casos leves, el desarrollo del síndrome de dificultad respiratoria aguda (ARDS, por sus siglas en inglés), sepsis o fallo multiorgánico (MOF, por sus siglas en inglés). Las afectaciones pueden variar por sexo, edad y comorbilidades detectadas (Yuki et al. (73)), así como la variante del SARS-CoV-2 (Hu et al. (27)). Aquí nos enfocamos en el aspecto fisiopatológico de la enfermedad, es decir, cómo es que la infección desencadena mecanismos moleculares que dan lugar a la misma. A continuación se menciona brevemente nuestro entendimiento de la progresión de la enfermedad.

2.1.1. Fisiopatología de la enfermedad COVID-19

La proteína Spike del SARS-CoV-2 reconoce y se une al receptor, la enzima convertidora de la angiotensina 2 (ACE2) en las células hospedera a lo que procede, junto con

la proteasa transmembranal de serina 2 (TMPRSS2), la internalización del virus. Los receptores ACE2 están altamente expresados en vías respiratorias, especialmente en las células epiteliales alveolares tipo 2 (AT2) (Khan et al. (36)), que recubren los alveolos; pero también se expresa en células del miocardio, renales, enterocitos y endoteliales en distintos órganos, lo cual puede explicar que estos también se vean afectados. Se ha visto que una mayor expresión de ACE2 correlaciona con una mayor susceptibilidad de infección (Chua et al. (10)).

Una vez dentro de la célula hospedera, el virus utiliza la maquinaria celular para replicarse y al ser citopático, induce a la muerte celular. Las moléculas exógenas son reconocidas por patrones moleculares asociados al daño (DAMPs, por sus siglas en inglés) y a los patógenos (PAMPs, por sus siglas en inglés); particularmente se reconoce el RNA de doble cadena por MDA5 que da lugar al reclutamiento de ubiquitín ligasas y quinasas de serina/treonina para coordinar la activación de los factores transcripcionales: factores reguladores de interferón (IRFs)(en especial IRF3) y del factor nuclear-kB (NF-kB), los cuales al entrar al núcleo conducen a la transcripción de interferón I o III (IFN-I ó -III) y genes que se activan en respuesta a IFN (ISGs) (Minkoff and tenOever (48)). A su vez, esto resulta en su reconocimiento por parte de los receptores *toll-like* (TLRs, por sus siglas en inglés) de macrófagos y células dendríticas circundantes que inducen la respuesta pro-inflamatoria por parte de citocinas y quimicinas Siddiqi and Mehra (56). La presencia de la citocina IL-1β en pacientes COVID-19, es un indicador de una muerte celular programada altamente inflamatoria conocida como pyroptosis (Tay et al. (60)).

Normalmente, la respuesta inmune innata y adaptativa descrita hasta aquí sería suficiente para eliminar virus, reducir la respuesta inmune y recuperarse; pero lo que se ha visto es que la respuesta inmune se ve descontrolada a partir de este punto en casos de COVID-19 severo caracterizándose por una respuesta sistémica hiperinflamatoria iniciando por una tormenta de citocinas (Bohn et al. (5), Tay et al. (60), Lamers and Haagmans (41)).

En particular en células pulmonares, se ha visto que las células AT2 tienen una alta expresión de *STAT1* el cual parece ser un regulador transcripcional de ACE2 promovido por la interacción intercelular con células T citotóxicas (CTCs) ante la infección (Chua et al. (10)); asimismo, se ha visto una alta expresión genes asociados a la respuesta viral (SFTPC, SFTPA1). La elevada respuesta inflamatoria en estas células resulta en la muerte celular programada y por lo tanto, en la reducción de este tipo celular. Se ha sugerido que las células alveolares tipo 1 (AT1) intentan, pero fallan en reprogramarse de vuelta a AT2, resultando en el daño pulmonar que explica las dificultades respiratorias que son consecuencia de la limitación del intercambio gaseoso en los alveolos y disminución de los niveles de oxígeno en sangre (Delorey et al. (15), Lamers and Haagmans (41)). Al tipo de daño pulmonar provocado por este mecanismo se le conoce en la literatura médica como daño alveolar difuso (DAD).

Diferencias en la expresión de IFN-I, IFN-III, ISGs y los patrones de reconocimiento durante la infección por SARS-CoV-2 se han asociado con diferencias en la severidad de la enfermedad en pacientes con COVID-19 (Minkoff and tenOever (48)).

En concordancia con ello, se ha visto que el antagonismo de la respuesta a interferón por proteínas virales (Nsp10 y Nsp14) ayuda a la replicación viral dando lugar a un incremento en la liberación de productos de la pyroptosis que suman a la respuesta inflamatoria aberrante (Tay et al. (60), Minkoff and tenOever (48)).

Más aún, se ha descrito que este fenómeno juega un rol en el desequilibrio en la respuesta inmune coordinada por la regulación de IRFs y NF-kB, que por una parte, compromete la respuesta celular antiviral por medio de la inducción de IFN-I y IFN-III, y por otra, recluta leucocitos a través de la secreción de citocinas y quimiocinas (Blanco-Melo et al. (4)). En detalle, la primera se ve obstruida (como se describió anteriormente) pero la segunda se mantiene ya que es únicamente controlada por NF-kB, al cuál SARS-CoV-2 no solo es incapaz de inhibir sino que parece estimular a través de la mímica de IL-17A por la proteína viral ORF8. Este mecanismo parece contribuir a la alta respuesta inflamatoria vista en COVID-19 y por lo tanto es central en la patogénesis provocada por este virus.

En casos de COVID-19 severo, se han observado niveles elevados de interleucinas 6,2,7,10; del factor estimulante de granulocitos (G-CSF), MCP1, IFN-γ , la proteína inflamatoria 1-α de macrófagos (MIP1-α) y el factor de necrosis tumoral (TNF) están elevados (Yang et al. (70), Yuki et al. (73), Bohn et al. (5)). Particularmente, la interleucina 6 se ha observado más en pacientes que no sobreviven (Tay et al. (60)).

La liberación de citocinas da lugar al desarrollo de células B y T adaptativas (Lamers and Haagmans (41)). Se ha visto que las células T, macrófagos, monocitos y células dendríticas derivadas de monocitos también pueden ser infectadas por SARS-CoV-2, lo cuál también puede cooperar en la producción aberrante de citocinas (Tay et al. (60)). Lo anterior provoca un llamado excesivo de células inmunológicas al tracto respiratorio inferior, que para lograr la infiltración promueven la extravasación, que puede llevar al daño epitelial y endotelial del pulmón. Sumando a ese efecto, los macrófagos y CTCs se ven enriquecidos en la expresión de marcadores de muerte celular como caspasas y TNF (Chua et al. (10)).

Esta tormenta de citocinas no solo afecta a los pulmones sino también a otros órganos al causar choque séptico y MOF, como se ha visto en el caso de daño al miocardio, en tractos gastrointestinales y en el daño renal y hepático (Tay et al. (60), Lamers and Haagmans (41)). Asimismo, la infiltración de células inmunológicas aunado a los bajos niveles de oxígeno y la reducción de la función normal de ACE2 tras la infección, parecen ocasionar la activación de la coagulación en los tejidos afectados. De manera concordante, en COVID-19 severo, se han observado bajos niveles de plaquetas, probablemente por su uso durante la coagulación. A este perfil protrombótico en pacientes con COVID-19 se le conoce como inmunotrombosis (Lamers and Haagmans (41)).

En la Figura 2.1 se encuentra un resumen de los eventos clave de la progresión patogénica del COVID-19 que fueron descritos anteriormente.

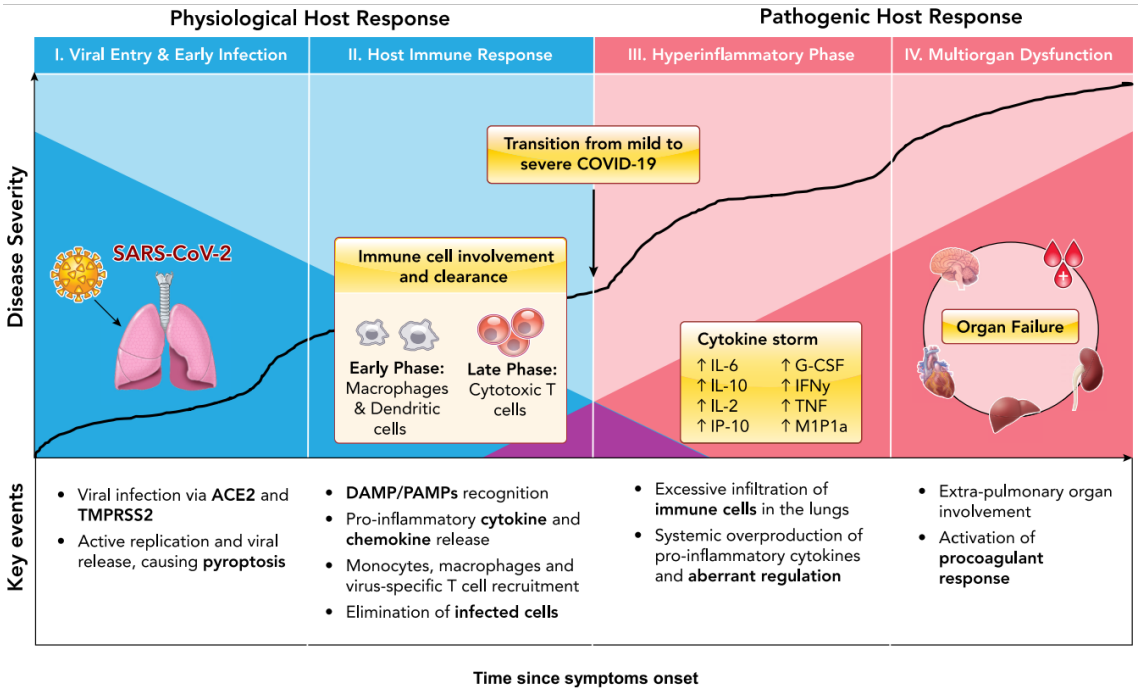


Figura 2.1: Resumen de los eventos clave durante la progresión fisiopatológica del COVID-19. El sombreado azul oscuro indica la respuesta fisiológica al virus del huésped en el tiempo y el sombreado rojo indica la respuesta patogénica hiperinflamatoria en el tiempo. Imagen tomada de Bohn et al. (5).

Una mayor comprensión sobre cómo se ensamblan los genes involucrados en las respuestas inmunológicas innatas y adaptativas para coordinar procesos biológicos en respuesta al ataque de SARS-CoV-2 en redes de regulación génicas (como se verá a continuación) nos ayudará a extender nuestro entendimiento de la enfermedad asociada.

2.2. Regulación Transcripcional de Genes

2.2.1. Principios

En términos generales, la regulación genética es el proceso por el cual una célula es capaz de controlar qué genes se están expresando y de qué forma lo están haciendo para ser capaces de realizar distintos procesos fisiológicos como el mantenimiento celular básico o responder a alguna señal externa. Para entender por qué es necesario para la célula tener control sobre los genes expresados consideremos el siguiente ejemplo en bacterias.

A pesar de que una bacteria es relativamente sencilla en comparación con organismos eucarióticos, si esta estuviera sintetizando todas las proteínas que codifican todos sus genes todo el tiempo, esto consumiría una gran cantidad de recursos energéticos y celulares, lo que no sería viable. En cambio, es favorable que solo invierta recursos para producir las proteínas y enzimas necesarias bajo cierto estímulo o condición. Por ejemplo, si en el ambiente donde cierta bacteria se encuentra no hay glucosa libre pero hay lactosa, entonces regularía sus genes para activar la expresión de genes necesarios para obtener glucosa a partir de la lactosa, a la vez que reprimiría los genes para consumir nutrientes que no se encuentren (Griffiths et al. (22)).

Podemos entender entonces a la regulación genética como necesaria para optimizar recursos en base a las condiciones que se encuentre la célula. En el resto de la tesis presente, nos enfocamos en un organismo multicelular (humano), donde la regulación genética también es necesaria en términos de función celular. Dado que podemos definir a una célula de un organismo eucariótico en términos de su función, y su función está dada por las proteínas realizando procesos fisiológicos, podemos decir que una célula está definida por los genes que expresa, o mas específicamente, por el perfil de expresión génica que muestra.

En el humano por ejemplo, por cada tejido, por cada tipo celular, dadas las condiciones cambiantes a las que se exponga cada célula, habrá un perfil de expresión génica particular, dando lugar a miles de estos perfiles transcripcionales que no serían posibles sin la capacidad de la célula de regular sus genes (Griffiths et al. (22)).

Ahora la pregunta es: ¿Cómo funciona la regulación de genes? ó ¿cómo se generan estos perfiles de expresión? La regulación de genes para dar lugar a una proteína ocurre en los distintos niveles: durante la transcripción (regulación transcripcional), durante las modificaciones post transcripcionales de RNA o durante la traducción a proteína (regulación post-transcripcional). Aquí nos enfocamos en la regulación transcripcional.

2.2.2. Factores Transcripcionales, secuencias regulatorias y mecanismos de regulación

Los principales elementos actuando durante la regulación transcripcional son (1) las proteínas regulatorias, que pueden ser Factores Transcripcionales que promueven la transcripción (conocidos y referidos aquí como TFs, por sus siglas en inglés) o TFs que reprimen la transcripción, también conocidos como Represores; y (2) los sitios en el genoma a los que se unen estas proteínas, *i.e.* las secuencias regulatorias, que canónicamente se dividen en tres tipos: (2.1) los promotores: regiones *cis*-regulatorias en el ADN que se encuentran río arriba de cada gene, donde se une la RNA Polimerasa (RNA Pol) para iniciar la transcripción; y regiones *trans*-regulatorias: (2.2) los *enhancers*: regiones distantes a los promotores (hasta un millón de bases nitrogenadas de distancia (Pennacchio et al. (50))) que interactúan de manera tridimensional para ayudar en la activación de la transcripción a múltiples genes; y (2.3) los *insulator elements*: regiones que como los *enhancers*, se encuentran distantes de los promotores y tienen una fun-

ción antagónica, que es bloquear los enhancers para bloquear la transcripción (Griffiths et al. (22)).

Los TFs y/o represores tienen dos funciones principales: (1) se unen a promotores, *enhancers* o *insulator elements* al reconocer sitios específicos en términos de la secuencia nucleotídica conocidos como Sitios de Unión de Factores Transcripcionales (TFBSs, por sus siglas en inglés) o motivos (motifs en inglés); y (2) interactúan directa o indirectamete con la RNA Polimerasa (los co-factores pueden funcionar como intermediarios entre TFs y la RNA Pol) para promover o reprimir su unión para dar inicio a la transcripción de manera cooperativa.

¿Cómo es que los TFs logran unirse a sus TFBSs? Recordemos que los TFs son proteínas, y al ser proteínas tienen dominios proteicos especializados en alguna función. La función especializada en cuestión aquí es precisamente, el reconocimiento de patrones de secuencia en el ADN. Por consiguiente, a estos dominios se les conoce como Dominios de Unión al ADN (DBD, por sus siglas en inglés) y existen varios tipos determinados por el o los motivos proteicos estructurales que lo conformen, por ejemplo: *helix-turn-helix, zinc-finger, leucine-zipper*, etc (Griffiths et al. (23)). De manera interesante, esto proviene de una razón evolutiva, es decir, aquellos TFs con un mismo tipo de DBD muy probablemente serán pertenecientes a la misma familia de TFs o dicho de otra manera, son parálogos de un gen ancestral en común.

Qué motivos o TFBSs reconoce cada TF y cómo estos patrones de secuencias y su posicionamiento en el genoma varían bajo distintas condiciones sigue siendo una pregunta de interés en el campo de la Genómica de la Regulación (Castro-Mondragon et al. (8), Lambert et al. (40)). Véase sección 2.3.3.

Ahondado en la organización de la regulación genética, cada gen eucariótico tiene un conjunto de TFs que lo regulan bajo distintas condiciones; y encontrar el TFBSs de un TF en particular en las regiones regulatorias de un gen es indicativo de que putativamente ese TF pertenece a ese conjunto. Asimismo, cada TF regula un conjunto de genes, mismo conjunto que puede cambiar a través de tipos celulares, tejidos (Lambert et al. (40)) y de distintas condiciones. Qué conjuntos de genes es regulado por cada TF (a lo que más adelante nos referimos como regulon) sigue siendo una pregunta abierta (Garcia-Alonso et al. (21)). Se ahonda sobre este tipo de estructuración de la regulación transcripcional en forma de redes en la sección 2.4.1.

2.2.3. La topología del ADN influencia la regulación genética

Al estar en el escenario eucariótico existe un tercer elemento relevante en la regulación genética: el estado de la cromatina. Se asume generalmente que el estado transcripcional basal de una célula eucariótica está "apagado'; esto se debe a que los nucleosomas (conjuntos proteicos donde se empaca el ADN de manera tridimensional, especialmente durante la condensación de la cromatina) ocupan el lugar de los TFs y de la RNA Pol inhibiendo la transcripción con mayor estabilidad que los represores, mecanismo que además es heredable entre células y se cree que cada tipo celular tendrá un perfil de cromatina abierta o cerrada particular que estará relacionado con el perfil

de expresión de una célula dada (Griffiths et al. (22), Marstrand and Storey (45)).

En años recientes se ha encontrado que la cromatina está organizada en subcompartimentos funcionales denominados Dominios Topológicamente Asociados (TADs, por sus siglas en inglés). Se encontraron al observar que distintas regiones de hasta 1 Mega base (diez millones de bases nucleotídicas de distancia) a lo largo de los cromosomas se encontraban en alto contacto físico y que esto se correlacionaba con la co-expresión de los genes contenidos en esas mismas regiones o TADs. El modelo aceptado para la formación de los TADs es conocido como *loop extrusion*, donde un largo segmento de ADN se dobla formando un bucle y aquello uniendo los extremos es el complejo proteico de la cohesina interactuando con el factor transcripcional CTCF. En concordancia con ello, se ha visto que quitar los TFBSs de CTCF rompe la estructura de los TADs (Ibrahim and Mundlos (29)).

La co-regulación de los genes embebidos en un mismo TAD puede darse por la proximidad tridimensional de los genes y sus promotores a un *enhancer* en común contenido en el mismo TAD.

2.3. Bioinformática

La bioinformática es un área de campo interdisciplinaria entre informática, estadística y ciencias biológicas, donde en general se desarrollan y aplican programas computacionales para analizar, organizar, comprender, visualizar y guardar información asociada con moléculas biológicas (Luscombe et al. (44)).

2.3.1. Bases de Datos

En la tesis presente se hace uso de algunas de las Bases de Datos (DBs) genómicas más relevantes actualmente. A continuación se da una breve descripción de cada una de ellas.

- El Centro Nacional para la Información Biotecnológica (**NCBI**, por sus siglas en inglés) de Estados Unidos es una base de datos que contiene bases de datos más pequeñas dedicadas a cierto tipo de información biológica particular. Los datos se pueden consultar en la página web o en línea de comandos con las herramientas de $e-utilities$. Gran parte de los datos públicos en investigaciones científicas se depositan en una de sus bases de datos lo cual añade a su gran valor dentro de la comunidad científica como recurso de información. Algunos ejemplos de información biológica depositada aquí son: datos de secuenciación, datos de estructuras de proteínas, publicaciones científicas, anotaciones de genes, entre otros.

- El *Gene Expression Omnibus* (**GEO**, por sus siglas en inglés) es una de las DBs contenidas dentro de NCBI. Es un repositorio especializado en datos de expresión

génica de alto rendimiento. Los autores de un estudio suben, acompañado de identificadores únicos del estudio y de los datos, las matrices de cuentas génicas o incluso los archivos de secuenciación resultantes de su trabajo a esta plataforma.

- El proyecto de Genotipos y Expresión de Tejidos (**GTEx**, por sus siglas en inglés) es un esfuerzo entre muchos grupos de investigación para construir una DB especializada en el estudio de expresión génica y regulación tejido-específica. Contiene datos de RNA-seq (matrices de cuentas), datos de secuenciación del genoma completo, datos del exoma provenientes de biopsias de distintos tejidos de casi mil personas, entre otros.

- El Navegador de Genomas del UCSC (**UCSC Genome Browser** en inglés) es eso, un repositorio y navegador visual de datos y anotaciones de genomas completos de organismos modelo generalmente "grandes", es decir, multicelulares, con la excepción de la levadura y algunos virus. Las anotaciones vienen de otras DBs y son de distintos tipos: variantes genéticas, información de regulación genética, genómica comparativa, entre otros. También se pueden descargar tablas de datos o archivos de secuencias en formatos FASTA, BED u otros de algún genoma completo con anotaciones particulares.

- Bases de datos de matrices descriptivas de TFBSs (mencionadas a mayor extensión en la sección 2.3.3): JASPAR, HOCOMOCO, RSAT (más que una base de datos, contiene una suite de herramientas mencionadas en la sección 2.3.3.2), entre otras.

2.3.2. Transcriptómica: RNA-seq

La transcriptómica se define como el estudio del transcriptoma (el conjunto completo de transcritos de RNA producidos por el genoma bajo condiciones específicas o en una célula específica) usando métodos de alto rendimiento (Wang et al. (63)). La comparación de transcriptomas entre condiciones o poblaciones celulares diferentes, puede dar lugar al descubrimiento de genes diferencialmente expresados.

Existen diferentes técnicas para el estudio del transcriptoma pero aquella de interés en la tesis presente es la de secuenciación del ARN o RNA-seq, que como su nombre lo indica, hace uso de las tecnología de *next-generation sequencing*. En años recientes se le conoce también como *bulk RNA-seq* que se refiere a que las muestras procesadas y cuantificadas provienen de una población celular, ya sea de un cultivo celular o de un tejido "en bulto", a diferencia de la tecnología más novedosa: secuenciación de ARN en célula única, *single-cell RNA-seq* ó *scRNA-seq*, que como su nombre lo indica, las "muestras" procesadas y cuantificadas corresponden a, en teoría, una sola célula.

El procedimiento para llevar a cabo un experimento de RNA-seq se divide en dos partes generales: la preparación de las muestras para secuenciación y el procesamiento de los datos de secuenciación. Existen muchas variaciones a lo largo del procedimiento de ambas partes, sin embargo, aquí nos enfocamos en el experimento de RNA-seq con

el objetivo de cuantificar la expresión de genes conocidos. A continuación se da una breve descripción de cada parte (Hrdlickova et al. (26), Conesa et al. (11)).

Pre-secuenciación

- Diseño experimental. Se debe elegir (1) el tipo de librería, *i.e.* si la secuenciación será *single-end* (secuenciación de una sola lectura por fragmento de ADN) o *paired-end* (secuenciación de dos lecturas localizadas a ambos extremos por fragmento de ADN); (2) profundidad de secuenciación, *i.e.* el número de secuenciación de lecturas para una muestra dada; y (3) el número de réplicas biológicas (la replica pertenece a distintas muestras biológicas) y de réplicas técnicas (donde la muestra biológica es la misma pero hay un procesamiento técnico separado, que sería útil para ver efectos de lote).

- Diseño de secuenciación. Este paso se debe llevar a cabo para evitar producir sesgos técnicos o factores de confusión cuando hay un gran número de muestras. Se debe tomar en cuenta la aleatorización de muestras en las rondas de secuenciación y en las líneas de la celda de flujo (donde se introducen las muestras para secuenciar) para contender con los efectos de lote.

- Pasos experimentales. (1) Extracción de ARN mensajero (abreviado como mRNA en inglés) a partir de su enriquecimiento usando selección de poly(A), (2) síntesis de ADN complementario (abreviado como cDNA en inglés) de doble cadena a partir del mRNA, (3) fragmentación del cDNA, (4) inserción de adaptadores y (5) amplificación.

Para el estudio presente, todas las muestras se secuenciaron con la tecnología de Illumina.

Post-secuenciación

- Revisión de calidad de lecturas de secuenciación: presencia de adaptadores, calidad de llamado de base, ver si hay efecto de lote por líneas de celda, distribución de porcentaje de pares de base de Citosina-Guanina, tamaño de lecturas, etc.

- Cortado de secuencias de mala calidad y adaptadores.

- Revisión de calidad de lecturas después del cortado para ver si mejoró.

- Alineamiento o pseudo-alineamiento de lecturas a genoma o transcriptoma de referencia. Paso necesario para identificar la identidad de cada transcrito secuenciado.

- Cuantificación de transcritos a partir del número de lecturas alineadas al gen o transcrito.

- Normalización de cuentas génicas o de transcritos. Necesario dado que el número de lecturas asociadas a una molécula están sujetas a sesgos técnicos: profundidad de secuenciación de muestras, longitud del gen y/o composición del ARN.
- Corrección de efectos de lote en caso de ser necesario, *i.e.*, si las muestras fueron procesadas distintamente en algún paso previo hasta la secuenciación.
- Se puede realizar un Análisis de Componentes Principales (PCA, por sus siglas en inglés) el cual proyecta los datos en términos de la combinación lineal de las mediciones rescatando la mayor variabilidad de los mismos. En general, se puede ver qué tan similares son las muestras entre sí de acuerdo al perfil transcripcional y también se puede usar para revisar si desaparece el efecto de lote después del paso de corrección.
- Análisis posterior dependiendo del objetivo del estudio. Un análisis común es el de expresión diferencial, donde se buscan los genes que estén diferencialmente expresados al comparar una condición de interés con su respectivo control.

2.3.3. Genómica de la regulación

Como se mencionó en la sección 2.2, la regulación genética puede dar lugar a miles patrones de expresión distintos, difiriendo en identidad y cantidad de moléculas a través del tiempo y de los tipos celulares; explicar este fenómeno es el reto del área de la Genómica de la Regulación (Kellis (35)).

2.3.3.1. pattern-matching

Recapitulando, los factores transcripcionales se unen a sitios específicos en el ADN conocidos como TFBSs o motivos para cooperar con otros TFs y la RNA Pol para regular la transcripción de genes específicos. Una opción bioinformática para conocer qué genes está regulando un TF dado es buscar su TFBSs en regiones genómicas, por ejemplo, el genoma completo o de manera más puntual, en regiones regulatorias no codificantes. A este escaneo del genoma para buscar TFBSs se le conocer como *pattern-matching* (Turatsinze et al. (62)) pues precisamente, se está buscando el patrón de secuencia que describe el sitio al que el TF se pega en el ADN.

De manera relevante en el campo de la Genómica de la Regulación, el patrón de secuencia que subyace un TFBS, formado por la interacción tridimensional fisicoquímica del dominio de unión proteico del TF al ADN, raramente será una secuencia de bases nitrogenadas en concreto. Lo que se ha visto de manera más general es que un TF dado puede unirse a muchos sitios en el genoma que sean similares pero difieran entre sí, de forma que un TFBS tendrá "preferencias" por ciertas bases nitrogenadas en posiciones puntuales de interacción entre la proteína y el ADN.

Una manera popular en la que se han representado los TFBSs con la información de preferencias por ciertas bases en ciertas posiciones es usando matrices tamaño p

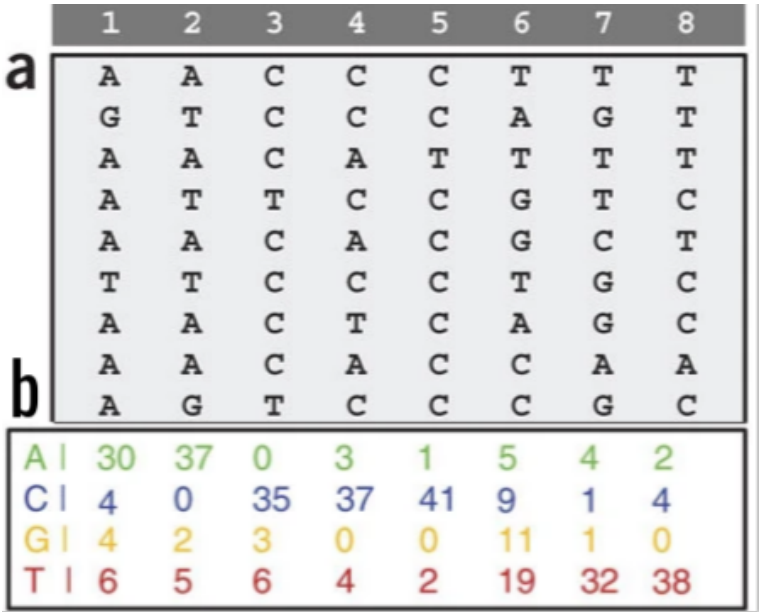


Figura 2.2: Representación del TFBS de Krüppel como PSSM. A) Alineamiento de secuencias por posición de interacción TF-ADN. B) PSSM del TFBS encontrado para Krüppel en *drosophila melanogaster*. (Imagen modificada de Turatsinze et al. (62))

posiciones x cada base nitrogenada (A,T,C y G), donde la posición es tan larga como la longitud de interacción proteína-ADN. Básicamente, aquellas bases que sean preferenciales en ciertas posiciones tendrán un número más grande en esa posición que el resto de las bases. A estas matrices se les conoce como Matrices de Peso de Posiciones (PWMs, por sus siglas en inglés) ó Matrices de *Scoring* Posición-específico (PSSMs, por sus siglas en inglés).

Como ejemplo consideremos la construcción de una PSSM para el TFBSs del factor transcripcional Krüppel, encontrado en al menos 44 sitios distintos en el genoma de Drosophila melanogaster. Como primer paso, se alinean las secuencias de los sitios de interacción entre el TF y ADN y se enumeran (Figura 2.2 A). Como segundo paso, se cuentan las veces que apareció cada base (A,T,C o G) en cada posición (aquí sumando a 40 por posición) y se muestra en formato de PSSM (Figura 2.2 B).

Existen distintos formatos para representar las PSSMs o incluso pueden tener transformaciones de las cuentas de bases (e.g. pesos) o información adicional. Un formato popular es el de "transfac". Como ejemplo se muestra la PWM de Krüppel descargada de la base de datos de JASPAR en la Figura 2.3.

Otro formato de interés en la tesis presente es el usado por el programa *cluster – buster*, ejemplificado en la Figura 2.4 para el TFBS de Krüppel. Se habla de este programa más adelante.

Por último en esta sección, cabe mencionar a JASPAR, que es la base de datos

```
AC MA0452.2
XX
ID Kr
XX
DE MA0452.2 Kr ; From JASPAR
PO       A       C       G       T
01       179.0   422.0   238.0   457.0
02       205.0   189.0   218.0   684.0
03       151.0   386.0   73.0    686.0
04       1248.0  0.0     0.0     48.0
05       1080.0  216.0   0.0     0.0
06       11.0    1170.0  0.0     115.0
07       0.0     1177.0  0.0     119.0
08       0.0     1296.0  0.0     0.0
09       0.0     202.0   0.0     1094.0
10       0.0     0.0     0.0     1296.0
11       0.0     113.0   0.0     1183.0
12       113.0   399.0   231.0   553.0
13       318.0   322.0   285.0   371.0
14       211.0   434.0   219.0   432.0
XX
CC tax_group:insects
CC tf_family:More than 3 adjacent zinc fingers
CC tf_class:C2H2 zinc finger factors
CC pubmed_ids:18332042
CC uniprot_ids:P07247
CC data_type:ChIP-chip
XX
//
```

Figura 2.3: PSSM del TFBS de Krüppel en formato transfac.

```
>MA0452.2 /name=Kr /info=9.787 /gc_content=0.418 /consensus=YTYAACCCTTTYTY /size=14
179     422     238     457
205     189     218     684
151     386     73      686
1248    0       0       48
1080    216     0       0
11      1170    0       115
0       1177    0       119
0       1296    0       0
0       202     0       1094
0       0       0       1296
0       113     0       1183
113     399     231     553
318     322     285     371
211     434     219     432
```

Figura 2.4: PSSM del TFBS de Krüppel en formato cluster-buster.

de matrices de TFBSs más grande y actualizada en la actualidad (Castro-Mondragon et al. (8)).

2.3.3.2. RSAT

Existe un gran número de recursos bioinformáticos para el campo de genómica de la regulación, ya sean bases de datos o paquetes de herramientas. Uno popular que contiene ambas es la plataforma de RSAT, cuyas siglas en inglés corresponden a "Herramientas para el Análisis de Secuencias Regulatorias" que, como su nombre lo indica, es una plataforma web y de línea de comandos que contiene un conjunto de herramientas bioinformáticas para facilitar el análisis del genoma regulatorio y de experimentos de regulación genética, además, contiene un gran número de bases de datos de PSSMs y el genoma completo de distintos organismos (Santana-Garcia et al. (55)).

2.4. Ingeniería reversa de redes de regulación

2.4.1. Idea general

La decodificación de qué interacciones regulatorias pueden existir en el genoma humano en general y en distintas condiciones biológicas es uno de los retos actuales de la biología moderna (Garcia-Alonso et al. (21), Mercatelli et al. (47)). Si bien se pueden añadir muchos niveles de complejidad a este problema, como al preguntarnos cuáles podrían ser los elementos de las interacciones regulatorias (por ejemplo: genes, factores transcripcionales y/o de represión, micro RNAs, long non-coding RNAs), por simplicidad aquí definiremos las interacciones regulatorias de nuestro interés como aquellas entre genes (regulados) y TFs (reguladores, que también pueden estar regulados), es decir, tipo TF-gene.

Las redes de regulación génicas (GRNs, por sus siglas en inglés) surgen como una forma de interpretar cómo es que un conjunto de TFs es capaz de orquestar distintos patrones de expresión génica y sus derivados procesos fisiológicos al encapsular de manera jerárquica subunidades de regulación conformadas por un regulador y sus genes regulados, que además suelen dar lugar a una función conjunta, de manera que las redes de regulación pueden entenderse como subprocesos biológicos.

A la inferencia de las redes de regulación génicas "prendidas" en una condición biológica dada a partir de un experimento se le conoce como ingeniería reversa de redes de regulación.

2.4.2. Tipos de métodos

Los programas desarrollados para la inferencia de GRNs pueden ser clasificados de acuerdo al tipo de método que utilizan. En el estudio presente nos interesaron en

particular aquellos que toman como datos de entrada los de experimentos de *bulk* RNA-seq o *single-cell* RNA-seq. La Tabla 2.1 describe las metodogías principales en las que se pueden clasificar estos programas.

Método	Descripción	Programas
Co-expresión génica	Se asume que dos genes están coexpresados cuando su expresión génica es significativamente dependiente entre sí. Esta dependencia puede estar dada de manera lineal (por ejemplo, usando correlación de Pearson) o no lineal (por ejemplo, usando Random Forests). Dada una lista de TFs, se busca el conjunto de genes que estén coexpresados con cada TF.	GENIE3, GRNBoost2, ARACNe
Escaneo de Motivos	Se buscan los motivos de ADN reconocidos por los TFs *in silico* en las regiones regulatorias de sus genes diana para construir o refinar redes de regulación.	i-Regulon, cistarget, networkinteractions, TF2Network
ChIP	Similar al método de escaneo de motivos, este añade una capa de información al tener evidencia de qué TFs están en contacto directo con las secuencias regulatorias mediante datos de Inmunoprecipitación de la Cromatina (ChIP, por abreviación en inglés).	ChIPpeakAnno
Literatura	Se realiza una curación literaria para establecer interacciones TF-gen corroboradas por experimentos de perturbación.	dorothea (base de datos)

Tabla 2.1: Métodos de inferencia de redes de regulación.

En la práctica, decidir qué método y programa usar no es trivial, ya que cada uno tiene sus ventajas y desventajas, algunas de las cuales se resumen en la Tabla 2.2.

Método	Ventajas	Desventajas
Co-expresión génica	(1) Las redes generadas son relativas a una condición biológica, (2) Las interacciones TF-gen fueron calculados a partir de la evidencia de un experimento	(1) Puede no ser capaz de rescatar interacciones que no ocurran a nivel transcripcional, (2) No distingue interacciones directas de indirectas, (3) Puede tener falsos positivos

Escaneo de motivos	(1) Discierne interacciones TF-gen directas de indirectas, (2) A pesar de tener falsos positivos, tiene un puntaje asociado que indica la seguridad que se tiene sobre la interacción putativa	(1) Se deben conocer los motivos de ADN reconocidos por TFs, asímismo (2) está sujeto a la calidad de las matrices PWMs representado los motivos usados, (3) genera falsos positivos, (4) dependiendo de los datos de entrada, estos no necesariamente representan redes específicas a condiciones biológicas
ChIP	Añade evidencia sobre la interacción ADN-proteína	(1) La mayoría de los eventos de interacción proteína-ADN pueden ser espurios y (2) dificultad para aunar experimentalmente esta técnica a otras.
Literatura	Las interacciones TF-gen están validadas experimentalmente para las condiciones biológicas dadas	(1) Dada la diferencia en protocolos de experimentos, puede existir poco sobrelape entre las interacciones TF-gen encontradas, (2) La cantidad de interacciones TF-genes avalada por experimentos es muy poca y se tiene un sesgo a TFs ampliamete estudiados

Tabla 2.2: Ventajas y desventajas de los métodos de inferencia de redes de regulación.

Lo recomendado por la comunidad científica ha sido utilizar una combinación de programas bioinformáticos que integren distintas metodologías, ya que suman las ventajas y las desventajas pueden cancelarse o aminorarse entre sí, como se describe con el ejemplo de SCENIC a continuación.

2.4.3. SCENIC

En el estudio presente se decidió utilizar un programa llamado SCENIC que incorpora dos de las metodologías mencionadas en la Tabla 2.1: co-expresión génica y escaneo de motivos (de Sande et al. (14)).

Para dar mayor detalle, SCENIC o pyscenic (su implementación en R o en python, respectivamente; que se utilizarán intercambiablemente aquí) es una *pipeline* para la inferencia de regulones a partir de datos provenientes de *single-cell* RNA-seq (o como veremos más adelante, de RNA-seq en general).

Para enfatizar, porque será relevante en las secciones siguientes de la tesis presente, los **regulones** son subredes de regulación génicas constituidas por un factor transcripcional y su conjunto de genes diana directamente regulados.

compuesta de tres pasos principales que se describen a continuación:

Paso 1: Inferencia de redes de regulación por co-expresión a partir de una matriz de cuentas génicas generada con datos de RNA-seq. Para esto, SCENIC integra el programa GENIE3 o GRNBoost2, que son métodos basados en árboles capaces de detectar dependencias de expresión combinatorias. El resultado de usar cualquiera de estos programas en la *pipeline* es una tabla de las interacciones TF-gen encontradas con una métrica de importancia (referido a continuación como puntaje) asociada que indica la fuerza de la dependencia.

- Como nota importante, los algoritmos de GENIE3 y GRNBoost2 tienen un componente aleatorio, por lo que las redes que se generan cada vez que se utiliza el programa son distintas entre sí.
- Dado que los datos de entrada para estos programas son una matriz de cuentas génicas, esta matriz puede tanto provenir de un experimento de *single-cell* RNA-seq como de *bulk* RNA-seq. De hecho, el programa GENIE3 fue diseñado inicialmente para datos de *bulk* RNA-seq (Huynh-Thu et al. (28)).
- GRNBoost2 se basa en el algoritmo de GENIE3 pero el tiempo de ejecución es considerablemente más rápido, alcanzando resultados similares. La diferencia entre GENIE3 y GRNBoost2 es que el primero calcula más arboles para estimar el parámetro de selección del "mejor" árbol, mientras que le segundo calcula menos pero utiliza un método heurístico de decisión para este parámetro y de esta manera sólo calcula los árboles "suficientes" (Moerman et al. (49)).

Paso 2: Construcción de regulones a partir del refinamiento de las redes de regulación construidas en el paso anterior. A *grosso modo*, en este paso se combinan los resultados de dos métodos: para cada par de interacciones TF-gen encontradas por co-expresión, se realiza la búsqueda del motivo de unión a ADN del TF (o TFBS representado por una matriz) en la región regulatoria del gen indicado. En detalle, si no se encuentra que dicho motivo está enriquecido en la región regulatoria, la interacción se descarta, lo cual añade peso sobre las interacciones conservadas y además permite discernir entre interacciones directas de indirectas.

- Este paso ocurre utilizando un programa previamente realizado por el mismo grupo llamado cistarget (más sobre esto se describe en la sección 3.5.1).
- La *pipeline* de pyscenic además realiza un filtrado de interacciones tomando en cuenta el puntaje establecido en el paso de co-expresión.
- Un detalle importante es que predeterminadamente en este paso se descartan las interacciones TF-gen cuya expresión génica este negativamente correlacionada utilizando un coeficiente de Pearson $<= -0.03$. Esto, porque como los autores indican, los regulones guiados por represores suelen ser pocos, tener un score bajo y tener menor enriquecimiento de su TFBS.

Paso 3: Cuantificación de la actividad de los regulones encontrados por muestra o célula. A *grosso modo*, para cada regulon, se calcula el "enriquecimiento" de la expresión de sus respectivos genes diana por célula.

- Para este paso se utiliza un programa interno de la *pipeline* conocido como AUCell.
- A mayor detalle, se utiliza una técnica de "rankeo y recuperación" donde para cada r regulón, se grafica la expresión génica teniendo en el eje x los genes ordenados de mayor a menor expresión y en el eje y el número de genes recuperados pertenecientes al regulón r dado. Una vez obtenida esta distribución se calcula el Área debajo del Curva (AUC, por sus siglas en inglés), que representa qué tan mayormente expresados se encuentran los genes del r regulón respecto al resto de los genes en la muestra. Es importante aquí notar entonces que la métrica de actividad de regulon por célula es relativo a la expresión génica en cada muestra.
- El *output* de este paso (y de SCENIC en sí), es la matriz de métricas AUC tamaño n regulones x m muestras.

scenic multi-runs y robustez de resultados Como se mencionó en el Paso 1 de la *pipeline* de pyscenic, los algoritmos de inferencia de redes de regulación utilizan un método basado en árboles que tiene un componente aleatorio lo cual provoca que las redes generadas sean distintas cada vez que se utilice este programa. Esto representa un problema por dos razones: (1) como biólogxs, nos interesa recuperar la red de regulación que es real y que está ocurriendo dentro de las células, y (2) esto reduce la reproducibilidad de los resultados.

Para contender con este problema, lxs autores integraron SCENIC en la estructura de vsn-pipelines y generaron el recurso de scenic multiruns pipeline, el cual permite correr la *pipeline* de pyscenic múltiples veces e integrar los resultados en cada iteración. Al usar scenic multi-runs, se puede filtrar tanto para quedarse sólo con aquellas redes de regulación *per se* (y subsecuentemente los regulones) que se encontraron una mayor cantidad de veces, como para quedarse sólo con el respectivo conjunto de genes diana que se encontraron como regulados por el TF dado (guiando al regulon) una mayor cantidad de veces. Toda esta información, en conjunto con la matriz de AUC, se guarda en el archivo formato *loom* (descrito en la sección 3.5.2) de salida.

Como se puede ver, es mejor usar scenic multi-runs ya que se tendrá un mayor confianza sobre los regulones inferidos del experimento, pues al haberse encontrado repetidamente, es más probable que estos regulones hayan sido de hecho reconstruidos por la biología detrás del experimento y no por un artefacto técnico.

Capítulo 3

Metodología

3.1. Recuperación de datos públicos

Para el objetivo principal de este estudio, desarrollar una *pipeline* de procesamiento de datos de RNA-seq para el análisis de redes de regulación de genes teniendo como caso de estudio la infección por SARS-CoV-2 en células humanas, justamente fueron necesarios datos transcriptómicos provenientes de condiciones biológicas relevantes a la infección por SARS-CoV-2. Para ello se recopilaron datos públicos generados por la comunidad científica ante la emergencia de la pandemia ocasionada por el mismo virus. Se buscó utilizar datos de RNA-seq, ya que, en gran parte y como se mencionó anteriormente en el Marco Teórico, los métodos de ingeniería inversa de redes de regulación han sido generados para estos. A continuación se describe la búsqueda, selección y se da una breve descripción de los conjuntos de datos por estudios (desde aquí referidos como *datasets*) elegidos para usar en el estudio presente.

3.1.1. Dataset: SARS-CoV-2 y otros virus en líneas celulares de epitelio pulmonar

Primero, se usaron los datos generados por Blanco-Melo et al. (4) por tres razones: (1) fue uno de los primeros estudios sobre la respuesta transcripcional a la infección por SARS-CoV-2 que hizo disponibles sus datos, (2) al haber sido generado por uno de nuestros colaboradores, se tenía el interés particular sobre el tema de investigación aquí abordado y (3) al contener muestras de infección a otros virus, permite una mejor inspección de los mecanismos celulares particulares a infección por SARS-CoV-2.

En detalle, estos datos corresponden a experimentos realizados en tres líneas celulares procedentes de epitelio pulmonar: A549, Calu-3 y NHBE; las primeras dos de adenocarcinoma pulmonar y la tercera proveniente de tejido bronquial primario sano, mismas que fueron infectadas en experimentos paralelos con distintos virus: Coronavirus 2 del Síndrome Respiratorio Agudo Severo (SARS-CoV-2), Virus tipo 3 de la Parainfluenza Humana (HPIV3), Virus de la Influeza A (IAV), Virus de la Influeza A

sin la proteína NS1 (IAVdNS1) (NS1 es la proteína patogénica que inhibe la respuesta innata de la célula inmune (LU et al. (43))) ó Virus Sincitial Respiratorio (RSV); y/o tuvieron distintos tratamientos: transducción de vector expresando ACE2, tratamiento con Ruxilitinib (que inhibe la señalización de IFN-I a través de JAK1/2) ó tratamiento con IFN-β. Los experimentos y la cantidad de muestras correspondientes están detalladas en la Tabla 3.1. Una descripción más detallada de los datos se puede consultar en el siguiente link: SRA Run Selector: PRJNA631753 .

Experimento	Procedencia	Tratamiento	Núm. muestras
A549 infección HPIV3	Adenocarcinoma pulmonar	HPIV3 infección	3
A549 infección IAV	Adenocarcinoma pulmonar	IAV infección	2
NHBE infección IAV	Células epiteliales bronquiales humanas primarias	IAV infección	4
NHBE infección IAVdNS1	Células epiteliales bronquiales humanas primarias	IAVdNS1 infección	4
Biopsia pulmonar para Control negativo sano	Biopsia pulmonar	Control	2
Biopsia pulmonar de paciente COVID-19 postmortem	Biopsia pulmonar	Paciente COVID-19	2
A549 control transducido con vector expresando ACE2 humano	Adenocarcinoma pulmonar	Control	6
A549 control	Adenocarcinoma pulmonar	Control	13
Calu-3 control	Adenocarcinoma pulmonar	Control	3
NHBE control	Células epiteliales bronquiales humanas primarias	Control	7
A549 infección RSV	Adenocarcinoma pulmonar	RSV infección	5
A549 infeccción SARS-CoV-2 transducido con vector expresando ACE2 humano (pretratamiento Ruxolitinib)	Adenocarcinoma pulmonar	SARS-CoV-2 infección + ACE2 + Ruxolitinib	3
A549 infeccción SARS-CoV-2 transducido con vector expresando ACE2 humano	Adenocarcinoma pulmonar	SARS-CoV-2 infección + ACE2	6
A549 infección SARS-CoV-2	Adenocarcinoma pulmonar	SARS-CoV-2 infección	6
Calu-3 infección SARS-CoV-2	Adenocarcinoma pulmonar	SARS-CoV-2 infección	3
NHBE infección SARS-CoV-2	Células epiteliales bronquiales humanas primarias	SARS-CoV-2 infección	3

NHBE tratamiento IFNB humano	Células epiteliales bronquiales humanas primarias	IFNB humano tratamiento	6

Tabla 3.1: Resumen de los experimentos en líneas celulares o muestras de pulmón generados por Blanco-Melo et al. (4)

3.1.2. Datasets: Biopsias de tejidos en pacientes de COVID-19 o sanos

Segundo, para explorar las redes de regulación en distintos tejidos, se buscaron datos transcriptómicos accesibles en la base de datos del National Center for Biotechnology Information (NCBI, por sus siglas en inglés), con principal interés en recopilar más datos de pulmón que complementaran las muestras de Blanco-Melo et al. (4).

Para ello, se utilizaron las herramientas $e-utilities$ desarrolladas por el NCBI como interfaz para acceder a su base de datos a través de la línea de comandos. Específicamente, se buscaron los estudios en la base de datos de GEO (Gene Expression Omnibus) pues todos los datos que almacena son transcriptómicos, con la siguiente línea de código lógico:

```
"rna-seq[ALL] AND ("lung"[MESH] OR lung[ALL])
AND (covid[ALL]
OR "Severe acute respiratory syndrome coronavirus 2"[ALL] OR
sars-cov-2[ALL])) AND "Homo sapiens"[ORGN]"
```

Este indica la búsqueda de *datasets* que sean de RNA-seq, que contengan los términos 'pulmón' y 'COVID', 'Severe acute respiratory syndrome coronavirus 2' ó 'SARS-CoV-2' y que sean de *Homo sapiens*. Esta búsqueda arrojó 121 resultados. Para la selección de estudios fue necesaria una revisión detallada del tipo de dato generado, los tipos de experimentos realizados, la cantidad de muestras, si había otros tejidos incluidos, etc. A partir de esa revisión se seleccionaron dos *datasets* que cumplían con las características deseadas: (1) SRP261138, de Desai et al. (16) y GSE171668 de Delorey et al. (15) . Del primer dataset, los datos crudos (las lecturas de secuenciación) estaban disponibles y del segundo las cuentas por gen. Aunque ambos estudios proveían muestras de distintos órganos, se descartaron algunos pues no tenían suficiente número de réplicas biológicas (o como se describe más adelante, no se encontraron controles) como puede ser consultado en las siguientes ligas: SRA Run Selector: SRP261138 y GSE171668 (en el archivo GSE171668_bulk_metadata.csv.gz); por lo tanto, solo se seleccionaron las muestras indicadas en la Tabla 3.2.

Dataset	Tejido	Estado	Núm. muestras
Delorey	Pulmón	COVID-19	18
Desai	Intestino	COVID-19	4
Desai	Corazón	COVID-19	7

Desai	Riñón	COVID-19	3
Desai	Hígado	COVID-19	6
Desai	Pulmón	COVID-19	52
Desai	Pulmón	Sano	5

Tabla 3.2: Resumen de las muestras seleccionadas para usar generadas por Desai y Delorey.

Para descargar los datos de Desai et al. (16) se usaron las herramientas de $e-utilities$: $e-search$ para buscar específicamente el proyecto SRP261138, $e-fetch$ para obtener los identificadores (o IDs, como se usará en este escrito intercambiablemente) de cada muestra y finalmente la herramienta $fastq-dump$ para descargar los archivos fastq de cada muestra.

La descarga de los datos de Delorey et al. (15) fue más sencilla, simplemente se descargó la tabla de cuentas por gen de las muestras de RNA-seq *bulk* de la página de GEO del estudio (con el identificador mencionado anteriormente) con la herramienta de la terminal unix *wget*.

3.1.3. Datasets: Biopsias de tejidos sanos de GTEx

Como se puede notar en la Tabla 3.2, los datasets de tejidos no cuentan con las muestras control necesarias para el análisis, por lo que fue necesario obtener muestras de otra fuente. La fuente elegida fue la base de datos de GTEx (véase sección 2.3.1), ya que cuenta con datos transcriptómicos de múltiples muestras de 54 tejidos hasta su octava actualización (consultada aquí).

Los criterios iniciales de búsqueda y selección de datos de muestras para usar en el estudio presente fueron los siguientes: (1) datos de RNA-seq; (2) similitudes en preparación de muestras, específicamente: extracción de RNA por medio de lisado derivado de PAXgene de QIAGEN y preparación de librería para secuenciación con el kit TruSeq de Illumina; (3) muestras provenientes de los tejidos de interés enlistados en la subsección anterior (pulmón, corazón, hígado, intestino y riñón); (4) para poder corregir por efectos técnicos, que los lotes de secuenciación o técnica (también se referirán aquí como *batches*) se repitieran entre muestras del mismo y distintos tejidos, ocupando el mínimo número de lotes posible (ya que a mayor número de variación técnica, más difícil es corregirla); y finalmente, (5) una buena calidad de secuenciación reportada.

Para seleccionar las muestras, se descargó el archivo de anotación de muestras de la versión 8 de GTEx y se utilizó el lenguaje de programación R con las funciones del paquete *dplyr* para filtrar por los primeros 3 criterios y encontrar el mínimo número de batches compartidos. Además, se pretendió encontrar un número de muestras control cercano al número de muestras de COVID del dataset de Desai et al. (16), es decir, alrededor de 10 por tejido. Como nota, las únicas muestras de RNA-seq disponibles de riñon provenían de la corteza, para el intestino solo había intestino delgado y en el caso

del corazón, había tanto del ventrículo izquierdo como del apéndice atrial. El resultado de la selección de muestras de GTEx se muestra en la Tabla 3.3.

NxT	Núm. muestras	Tejido	Tejido Específico	Lote 1	Lote 2
12	1	Corazón	Apéndice Auricular	BP-44261	LCSET-4796
	2	Corazón	Apéndice Auricular	BP-47696	LCSET-4907
	2	Corazón	Apéndice Auricular	BP-49102	LCSET-5305
	2	Corazón	Ventrículo Izquierdo	BP-47696	LCSET-4907
	1	Corazón	Ventrículo Izquierdo	BP-48576	LCSET-4955
	1	Corazón	Ventrículo Izquierdo	BP-49102	LCSET-5304
	1	Corazón	Ventrículo Izquierdo	BP-49102	LCSET-5305
	1	Corazón	Ventrículo Izquierdo	BP-67833	LCSET-8132
	1	Corazón	Ventrículo Izquierdo	BP-75925	LCSET-9767
8	1	Riñón	Corteza	BP-44261	LCSET-4796
	1	Riñón	Corteza	BP-47696	LCSET-4907
	2	Riñón	Corteza	BP-49102	LCSET-5304
	2	Riñón	Corteza	BP-67833	LCSET-8132
	2	Riñón	Corteza	BP-75925	LCSET-9767
10	3	Hígado	Hígado	BP-47696	LCSET-4907
	2	Hígado	Hígado	BP-48576	LCSET-4955
	2	Hígado	Hígado	BP-56446	LCSET-6445
	3	Hígado	Hígado	BP-75925	LCSET-9767
12	2	Pulmón	Pulmón	BP-44261	LCSET-4796
	2	Pulmón	Pulmón	BP-47696	LCSET-4907
	2	Pulmón	Pulmón	BP-48576	LCSET-4955
	2	Pulmón	Pulmón	BP-49102	LCSET-5304
	2	Pulmón	Pulmón	BP-67833	LCSET-8132
	2	Pulmón	Pulmón	BP-75925	LCSET-9767
8	2	Intestino	Delgado - Terminal Ileum	BP-44261	LCSET-4796
	1	Intestino	Delgado - Terminal Ileum	BP-47696	LCSET-4907

	2	Intestino	Delgado - Terminal Ileum	BP-48576	LCSET-4955
	2	Intestino	Delgado - Terminal Ileum	BP-56446	LCSET-6445
	1	Intestino	Delgado - Terminal Ileum	BP-75925	LCSET-9767

Tabla 3.3: Resumen de las muestras seleccionadas proveninetes de GTEx. NxT: Número de muestras por tejido. Lote 1 corresponde a la extracción de ADN y el Lote 2 al experimento de expresión

Tomando en cuenta ambos *batches* (Lote 1 y 2), hay 8 combinaciones de ellos, y todos están compartidos en al menos dos tejidos específicos distintos como se resume en la Tabla 3.4. Esto permite que al realizar la correción de efectos de *batch* en pasos subsecuentes esta no esté sesgada al tipo de tejido de proveniencia.

Lote 1	Lote 2	Tejidos
BP-48576	LCSET-4955	4: Corazón VI, Hígado, Pulmón e Intestino
BP-75925	LCSET-9767	5: Corazón VI, Riñón, Hígado, Pulmón e Intestino
BP-44261	LCSET-4796	4: Corazón AA, Riñón, Pulmón e Intestino
BP-47696	LCSET-4907	5: Corazón VI y , Riñón, Hígado, Pulmón e Intestino
BP-67833	LCSET-8132	3: Corazón VI, Riñón y Pulmón
BP-56446	LCSET-6445	2: Hígado e Intestino
BP-49102	LCSET-5305	2: Corazón VI y AA
BP-49102	LCSET-5304	2: Corazón VI, Riñón y Pulmón

Tabla 3.4: Batches y tejidos que los comparten. VI: Ventrículo Izquierdo. AA: Apéndice Auricular.

Finalmente, las matrices de cuentas de estas muestras fueron descargadas directamente del GTEx Portal (archivo: GTEx_Analysis_2017-06-05_v8_RSEMv1.3.0_transcript_tpm.gct.gz).

3.1.4. Diseño del Análisis

Dado que se recopilaron datos de distintas procedencias experimentales, con distinto número de repeticiones biológicas y se tenía una falta de muestras control para las muestras de tejido infectado, se decidió dividir el análisis en tres partes pero con el mismo objetivo en cada uno: encontrar las redes de regulación relevantes en el progreso de la infección; esto, con los objetivos de realizar un análisis ordenado sobre los objetivos mencionados y evadir tener que lidiar con más efectos de lote. El análisis de datos se dividió en tres partes:

1. Infección de SARS-CoV-2 en comparación a otros virus en líneas celulares

   Datos: Blanco-Melo et al. (4) (Solo experimentos con líneas celulares)

2. COVID-19 en distintos tejidos

   Datos: biopsias COVID-19 Desai et al. (16) y biopsias tejido sano Aguet et al. (3).

3. Pulmón con COVID-19 en comparación con sano

   Datos: muestras de biopsias de pulmón COVID-19 o sano de Blanco-Melo et al. (4), Desai et al. (16) y Delorey et al. (15).

3.2. *Pipeline* de procesamiento de datos de RNA-seq para el análisis de redes de regulación de genes

Para el estudio presente, se desarrolló y estandarizó la estructura del procesamiento de datos provenientes de la tecnología *bulk RNA-seq*, incluyendo datos provenientes de distintas fuentes, para la construcción y análisis de redes de regulación. El panorama general de esta puede verse en el diagrama de la Figura 3.1, mencionando los datos y análisis específicos realizados aquí.

Esta *pipeline* integró estrategias específicas para sus diferentes objetivos, las cuales se describen a continuación y se ahonda en los detalles y metodología específica en las siguientes secciones.

1. Con el objetivo de obtener la cuantificación de genes conocidos se utilizó la herramienta de *kallisto*, cuyo algoritmo realiza un pseudo-alinemamiento a un transcriptoma de referencia.

2. Para integrar datos provenientes de distintos estudios se incorporó el uso de la herramienta ComBat_seq, particularmente aplicando un segundo paso de corrección de efectos de lote para tomar en cuenta las diferencias por *dataset* de procedencia.

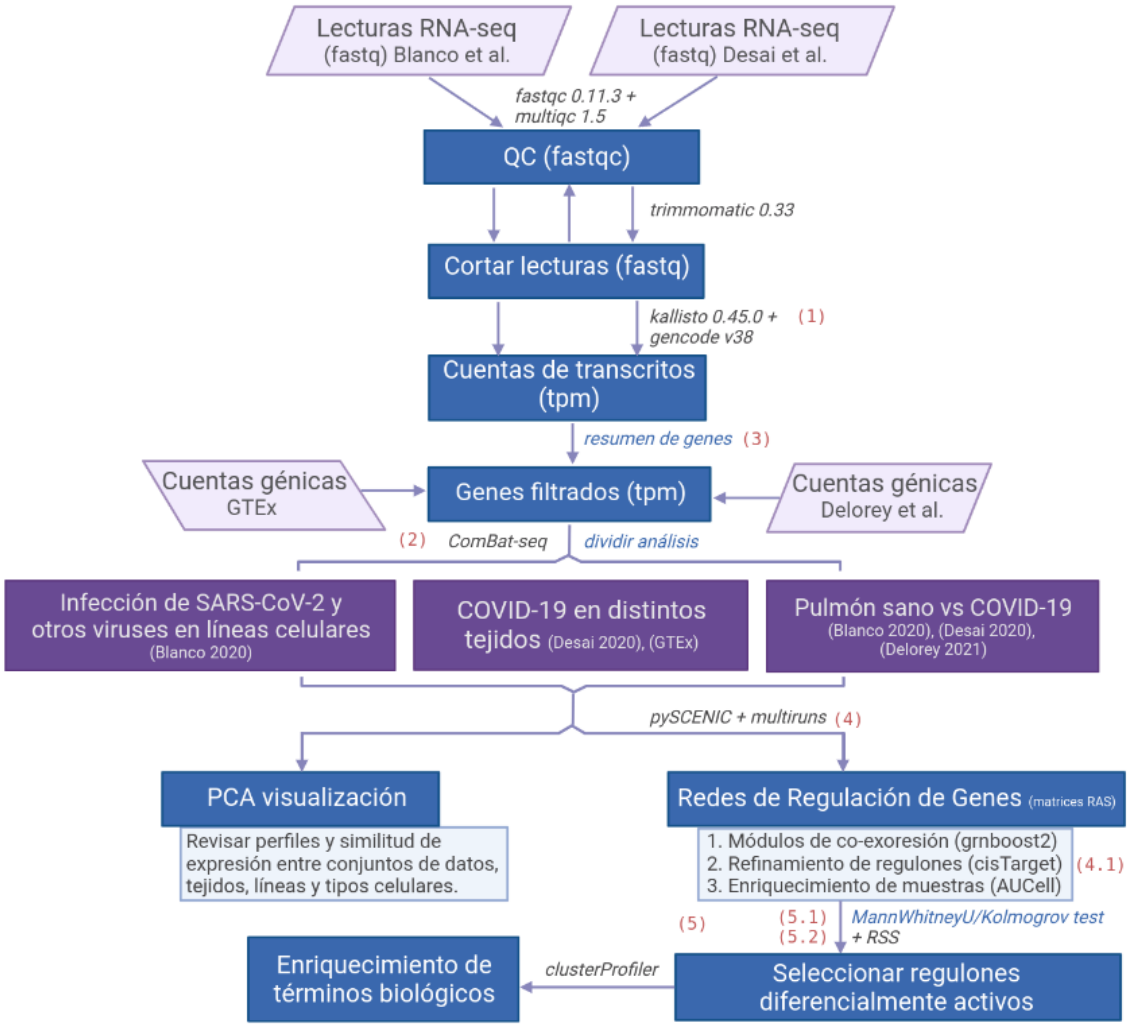


Figura 3.1: Esquema general del procesamiento de datos de RNA-seq para el análisis de redes de regulación de genes. Los datos fuente para el estudio presente se indican con un trapecio color lila claro, los rectángulos azules indican las etapas de la *pipeline* mientras que los métodos se indican a un lado de las flechas con las herramientas (gris) o el procesamiento elaborado aquí (azul). Los tres análisis o datasets analizados en conjunto principales se indican en un cuadro morado. Los números en rojo corresponden a estrategias específicas mencionadas en la sección 3.2.

3. Para asegurarnos de tomar un gen por factor transcripcional y evitar la redundancia en los regulones construidos por la *pipeline* de pyscenic, se aplicó un paso de elección del "mejor" transcrito por gen (referido en el diagrama como *resumen de genes*).

4. Para la construcción de redes de regulación confiables y reproducibles que integraran un mecanismo putativo de regulación (esto es, de regulación directa de TF a su gen diana) se utilizó la *pipeline* de SCENIC integrado en la *multiruns pipeline*.

    4.1. Adicionalmente, para el segundo paso de pyscenic, donde se realiza el refinamiento de las redes de regulación con base al enriquecimiento de motivos de factores transcripcionales en las regiones regulatorias de los genes diana, se actualizó la base de datos de cistarget (descrito en la sección 3.5.1). A pesar de que no se utilizó en el estudio presente, esta se puede utilizar en la pipeline simplemente reemplazando tres archivos (véase 3.5.2.3).

5. Finalmente, para la selección de regulones relevantes:

    5.1. Se buscaron aquellos que se encontraban diferencialmente activados en casos en comparación con controles, de acuerdo a dos pruebas de rankeo (Mann Whitney U y Kolmogrov) con un FDR < 0.0295 (en los análisis de tejidos), FDR < 0.01 (en el análisis de pulmón), o un FDR $<= 0.14$ (en el análisis de líneas celulares menos en dos experimentos en los que se utilizó un FDR $<= 0.19$ y 0.24). Véase sección 3.6.4.

    5.2. Se tomaron aquellos regulones que estaban más presentes en la condición dada que en el resto del análisis en turno utilizando la métrica de *Regulon Specificity Score* (RSS) y un LFC $>= 0.001$ en comparación al control.

El código específico, los componentes y detalles de la *pipeline* presentada pueden consultarse de manera ordenada en el repositorio del github (https://github.com/amedina-liigh/PulmonDB_COVID) creado para el estudio presente.

3.3. Pre-procesamiento de datos

3.3.1. Nota Tipo y Tratamiento de Datos

Como se describió en el apartado de Transcriptómica en el Marco Teórico del documento presente, hay un esquema general sobre el tratamiento de datos de RNA-seq. Esta *pipeline* está sujeta a cambios dependiendo de varios factores: la pregunta biológica de interés en cada estudio, de los datos disponibles, la tecnología ocupada, etc. Para realizar el trabajo presente hubo un problema adicional importante: la introducción de distintos conjuntos de datos provenientes de distintos estudios. Esto es relevante ya que

su procesamiento anterior y posterior a la secuenciación no fue uniforme. Se tomaron distintas medidas para integrar estos *datasets* de la manera más homogénea posible.

De manera más detallada, los *datasets* recopilados se descargaron en la forma de distintos tipo de datos: como secuencias directo de secuenciación (archivos *fastq*) o como matrices de cuentas por gen o transcrito. Dada esta consideración y otras (muestras seleccionadas, *batches*, genoma o transcriptoma de referencia, etc.), el procesamiento de datos de cada *dataset* fue distinto, como se verá caso por caso en las siguientes subsecciones.

Particularmente, se abordaron distintas estrategias para integrar estos datos:

- Se recuperaron los datos de secuenciación crudos (*fastq*s) cuando estos estaban disponibles para que todo el pre-procesamiento de datos fuera homogéneo.
- Dependiendo de si los datos son *fastq*s o matrices de cuentas, se introdujeron los datos a la *pipeline* en distintas etapas.
- De acuerdo al tipo de referencia genómica utilizada para definir los genes en los *datasets* que solo tenían matrices de cuentas génicas disponibles, se decidió qué referencia utilizar (gencode protein coding).
- Asimismo, de acuerdo a el tipo de algoritmo de conteo utilizado, se decidió qué programa utilizar que en este caso el tipo de algoritmo fue un pseudoalineador: *kallisto*.
- Se utilizó $ComBat_seq$ para corregir los efectos de *batch* por y entre *datasets*.
- Para la corrección de *batches*, dado que es necesario introducir las matrices de cuentas completas pero los *datasets* diferían en su definición de genes (particularmente los datos de Blanco-Melo y Desai tenían cuentas por transcritos y GTEx y Delorey cuentas por gene), se convirtieron los nombres de genes a identificadores de transcritos (como en el archivo de salida de *kallisto*) usando el "mejor transcrito" (véase sección 3.3.2.4).
- Una vez hecha la corrección por *batches*, se filtraron las matrices de cuentas para quedarnos solamente con un transcrito por gene, en particular quedándonos con el "mejor" transcrito.
- Para optimizar tanto el procesamiento técnico como el análisis de los resultados, y dado que la corrección de *batches* no es perfecta (es decir, no siempre logra eliminar por completo la variación técnica de los datos), se decidió dividir el análisis de los datos en tres grupos, como se resume en la sección 3.3.6. Esto permitió (1) equilibrar los controles en el análisis de tejido y usar los controles provenientes de los mismos *datasets* en el análisis de pulmón, (2) no tener que corregir los efectos de lote de más de tres *datasets* y (3) realizar tres análisis en paralelo con la misma definición de genes y metodología cuyos resultados fueran altamente comparables.

Algunas diferencias iniciales de los *datasets* recopilados, relevantes para el procesamiento de datos, se resumen en la Tabla 3.5

Dataset	Tipo de Dato	Referencia para conteo	Tipo de algoritmo de alineamiento y conteo	Batches reportados	Otro
Blanco-Melo et al. (4)	secuencias fastq	NA	NA	Rondas de secuenciación	lecturas *single-end* (SE, por sus siglas en inglés) y varios fastqs por muestra
Desai et al. (16)	secuencias fastq	NA	NA	Rondas de secuenciación y plataforma de secuenciación	lecturas *paired-end* (PE, por sus siglas en inglés)
Delorey et al. (15)	matriz de cuentas génicas	transcriptoma	RSEM: pseudoalineador	Ninguno, reportado como muestras procesadas uniformemente.	Excluir muestras de no interés
GTEx Aguet et al. (3)	matriz de cuentas genicas	genoma: gencode v26	RSEM: pseudoalineador	extracción de ADN y fecha de secuenciación	NA

Tabla 3.5: Diferencias entre datasets recopilados relevantes para el tratamiento de los datos.

En las secciones siguientes se describe el pre-procesamiento de datos de cada *dataset* hasta sus respectivas uniones.

3.3.2. Datos Blanco-Melo

3.3.2.1. Revisión de calidad de secuenciación (QC)

Se utilizó el programa fastqc para generar los reportes de calidad de secuenciación por archivo $fastq$. Se analizaron las gráficas de 331 reportes generados. En general, se observó lo siguiente en las secuencias provenientes de experimentos en líneas celulares:

- Buena calidad de secuencia, donde el *phred score* de 30 se conservaba en longitudes de 130-150 pares de bases (pb)

- Las gráficas del contenido de secuencia por base ó *per base sequence content* mostró un sesgo general en la proporción de pares AT y CG, pero esto es algo que se suele ver en librerías de RNA-seq.
- Algunas secuencias duplicadas o sobrerepresentadas
- Algunos picos en la distribución de GC en algunas muestras

Para las muestras de biopsias de pulmón con COVID-19 o sanas, la calidad en general empeoró pero se mantuvo un *phred score* >= 30. Se observó la presencia de secuencias duplicadas, sobrerepresentadas y adaptadores. Adicionalmente, las muestras de pacientes con COVID-19 tenían un fuerte sesgo en la métrica de *per base sequence content* hacia la base G.

3.3.2.2. Corte de secuencia o *trimming*

Se utilizó el programa *trimmomatic* versión 0.33 (Williams et al. (67)) para:

- Cortar secuencias de adaptadores (correspondientes a las usadas en el paquete de preparación de librería TruSeq2-SE de Illumina, tomadas del github de trimmomatic)
- Conservar solo bases de alta calidad
- Descartar secuencias de menos de 40 pb

como se muestra en el siguiente código de ejemplo:

```
trimmomatic SE -phred33 file.fastq file_trmd.fastq
ILLUMINACLIP:trimmomatic/adapters/TruSeq-SE.fa:2:30:10
SLIDINGWINDOW:5:30 MINLEN:40
```

3.3.2.3. Pseudo-alineamiento y conteo

Se eligió utilizar el programa *kallisto quant* versión 0.45, que es un programa para cuantificar la abundancia de transcritos de datos de RNA-seq (Bray et al. (6)), principalmente por dos razones: (1) al utilizar un algoritmo basado en el pseudo-alineamineto de secuencias lo cual además facilita el paso de conteo de trascritos, este es sumamente rápido computacionalmente; y (2) como se vio en la Tabla 3.5, este es similar al programa utilizado para generar las matrices de cuentas génicas de otros datasets.

La utilización de este programa constó de tres pasos descritos a continuación:

1. Generar el transcriptoma de referencia.

- Se utilizó el transcriptoma de referencia de gencode versión 38 con el motivo de homogeneizar la definición de genes entre los datasets utilizados en este estudio, particularmente por las siguientes razones: (1) la definición de transcritos se centró en los genes codificantes de proteína, y (2) como se consultó para otros datasets usados en este estudio, estos fueron generados usando un genoma de referencia de gencode; lo cuál permitió que la definición de genes entre distintos conjuntos de datos fuera más similar.
- Se utilizó el programa *kallisto index* para generar el índice del transcriptoma.

2. Contar el promedio y la desviación estándar de la longitud de secuencias de lecturas por muestra.

   - Se elaboraron programas en bash y perl para acceder a cada archivo .fa de cada corrida de muestra para que calculara ambas métricas. En detalle, este programa (1) accede a cada archivo fasta de cada corrida para contar y guardar las longitudes de cada lectura, así como contar cuántas lecturas son; una vez hecho esto para cada archivo fasta de cada muestra biológica, (2) calcula el promedio de longitud de secuencia y la desviación estándar (*i.e.* la raíz cuadrada de la sumatoria de la diferencia entre las longitudes individuales y el promedio al cuadrado entre la cantidad de lecturas); (3) hace lo anterior para cada muestra de cada experimento y despliega los resultados como una tabla indicando: experimento/muestra promedio_de_longitud desviación_estándar.

3. Cuantificar transcritos por muestra.

   - Se utilizó el programa *kallisto quant* con el transcriptoma de referencia realizado previamente y, al haber sido las secuencias generadas con el modo de lectura *single-end*, *kallisto* requirió parámetros adicionales indicando el promedio y la desviación estándar de la longitud de las lecturas seguidos por todos los archivos de corridas de secuenciación $fastq$ de cada muestra.

Se utilizaron las cuentas de transcritos por millón (TPM, por sus siglas) que calcula kallisto para análisis posteriores.

3.3.2.4. Generación de matrices de cuentas

Se generó un código en R que toma el tipo de cuentas elegidas de un conjunto de archivos salida tsv de *kallisto* (es, decir los archivos de salida por muestra) y los organiza en una matriz de cuentas tamaño m muestras x t transcritos. Este puede ser consultado en github: MakeExperimentCountMatrices.R.

Este mismo código tiene la función adicional de filtrar transcritos para elegir el "mejor" de ellos usando la opción $filter\_transcripts = TRUE$. El motivo detrás de esta función fue homogeneizar la definición de genes y reducir la complejidad del análisis

posterior, en específico, se buscó tomar un solo transcrito como representante de cada gen. A continuación se describe la metodología de esta función añadida: Basado en marcadores bioinformáticos de transcritos conocidos como *transcript flags* (los cuales categorizan la confianza que se tiene sobre que un transcrito sea codificante a proteína o biológicamente relevante basado en componentes de su secuencia, conservación y expresión), este elige el "mejor" transcrito para un gen dado. Brevemente, el código (1) consulta a la base de datos de *ensembl* a través de la herramienta de biomartR versión 2.46.2 (Drost and Paszkowski (18)) las *transcript flags* provenientes de distintos programas o bases de datos (gencode, appris, tsl y ensembl), (2) busca el transcrito que tenga una mejor combinación de marcadores de alta confianza y se elige un transcrito dado. A su vez, en este paso se puede cambiar el identificador del transcrito por el nombre del gen. Como resultado final se tiene una matriz de cuentas de genes por muestra.

3.3.2.5. Corrección por efectos de *batch*

Se eligió utilizar el programa *ComBat_seq* (la versión nueva de ComBat) del paquete sva versión 3.38.0 (Zhang et al. (74)) en R para realizar la corrección de *batches* por las siguientes razones: (1) su versión anterior, ComBat, ha sido ampliamente utilizada en la literatura, (2) es un programa elaborado para datos de RNA-seq que a diferencia de otros algoritmos, no asume una distribución normal de las cuentas de los genes sino una binomial negativa, que es más acertado y conserva las cuentas por enteros, y (3) tiene la opción de añadir covariables para que tome en cuenta condiciones biológicas en el modelo, lo cual previene que se pierda la señal de genes diferencialmente expresados.

Como se indica en la Tabla 3.5, los *batches* a tomar en cuenta en este *dataset* son las fechas de secuenciación, y las covariables de condiciones biológicas son los experimentos especificados en la Tabla 3.2.

Una nota importante durante este procedimiento, la corrección de efectos de *batch*, es que lo óptimo es **no** realizar un filtrado a la matriz de cuentas introducida como *input* al programa dado, ya que la variabilidad técnica suele provenir del experimento de secuenciación, y al excluir muestras o cuentas (como se verá que fue el caso más adelante), la estimación del efecto de esta variabilidad sobre las cuentas puede verse afectada. Tomando ello en cuenta, cada vez que se requirió de una corrección de efectos de lote, se utilizó la matriz de cuentas completa en este estudio. Excepto con los datos de GTEx (véase a continuación), ya que la cantidad de datos provenientes de esta fuente consta de experimentos masivos, por lo que el proceso de corrección habría tardado mucho e impedido el progreso del estudio presente.

3.3.3. Datos Desai

3.3.3.1. Revisión de calidad de secuenciación (QC)

Se generaron los reportes gráficos de calidad de secuenciación como se describió anteriormente. Adicionalmente, se utilizó el programa *multiqc* (Ewels et al. (19)), programa que resume una colección de reportes de $fastqc$ en uno solo, para crear un reporte para todas las muestras.

En general, se observó lo siguiente:

- Buena calidad de secuencia, donde el *phred score* de 30 se conservaba en longitudes de 130-150 pb (pares de bases)
- Largo de secuencia de alrededor de 75 pares de bases.
- En general buena calidad de secuencia, donde el *phred score* $>=$ 29 se conserva a lo largo de la secuencia, excepto en las primeras 7 bases. Lo cual suele ocurrir con datos de RNA-seq.
- Buenas proporciones de pares de bases AT y CG.
- La distribución de contenido de GC para la mayoría de las muestras está centrada en 50 %, con forma de campana. Algunas muestras tienen una distribución con picos en 55 y 70 %.
- Altos niveles de duplicación y de sobrerepresentación de secuencia.
- Contenido de adaptador casi nulo.

3.3.3.2. Conteo de transcritos

Se utilizó el programa *kallisto quant* para hacer e conteo de transcritos por muestra como describió anteriormente excepto a un cambio: el modo de lectura durante la secuenciación de los datos de Delorey fue *paired-end*, por lo que no fue necesario computar el promedio y la desviación estándar de las longitudes de las lecturas.

3.3.3.3. Generación de matrices de cuentas

Se generó la matriz de cuentas de los datos de Delorey de la misma manera que se mencionó anteriormente.

3.3.3.4. Corrección por efectos de *batch*

Se utilizó ComBat_seq tomando en cuenta la combinación de dos batches: fecha de rondas de secuenciación y la plataforma de secuenciación, y como covariables biológicas: el tejido de procedencia conjunto al estado indicado en la Tabla 3.2.

3.3.4. Datos Delorey

3.3.4.1. Conversión de cuentas a CPMs

Usando R se transformó la matriz de cuentas a Cuentas por Millón (CPMs) en los siguientes pasos por muestra:

1. Sumar cuentas de todos los genes
2. Dividir las cuentas por la suma anterior
3. Multiplicar cuentas de genes por $10\hat{6}$

3.3.4.2. Excluir muestras en dataset

El código reportado en github por los autores de este dataset indica que algunas muestras no provenían de la tecnología de RNA-seq en *bulk*. Usando en R un procedimiento similar al reportado, los datos provenientes de esas muestras fueron excluidos.

3.3.4.3. Cambiar nombres de genes por su sinónimo en transcrito

Este paso fue necesario para ejecutar la corrección de *batches* entre *datasets*, ya que la anotación de genes debe ser la misma para todas las muestras en la matriz de cuentas dada como *input* a $ComBat\_seq$.

Para esto se tomaron los identificadores de los transcritos por gen elegidos como mejores usando la opción $filter\_transcripts = TRUE$ en el código MakeExperimentCountMatrixes.R detallado anteriormente.

Esto dió como resultado la matriz de cuentas de datos de Delorey.

3.3.5. Datos GTEx

3.3.5.1. Conversión de cuentas a CPMs

Se realizó como de describió anteriormente (véase subsección 3.3.4.1).

3.3.5.2. Corrección por efectos de *batch*

Se utilizó ComBat_seq tomando en cuenta los *batches* resumidos en la Tabla 3.4.

3.3.5.3. Cambiar nombres de genes por su sinónimo

Al igual que con los datos de Delorey, este paso fue necesario para la procedente mezcla de *datasets*.

Esto dió como resultado la matriz de cuentas de datos de GTEx.

3.3.6. Unión de datasets para 3 análisis

A continuación se describe cómo se realizó la unión de *datasets* para realizar el subsecuente análisis dividido en tres enfoques descritos en la sección 3.1.4.

3.3.6.1. Análisis Líneas Celulares: Datos Blanco-Melo

Este análisis no requirió de la unión con otros datasets, en cambio, se requirió remover las cuatro muestras provenientes de biopsias pulmonares (mencionadas en 3.1) de la matriz de cuentas generada (véase sección 3.3.2).

3.3.6.2. Análisis COVID-19 en distintos tejidos: Datos Desai y GTEx

1. De la matriz de cuentas generada con los datos de Desai (véase subsección 3.3.3), se removieron las muestras pertenecientes a tejido sano de pulmón usando R, esto, (1) dado que el *dataset* carecía de muestras control para otros tejidos; (2) de esta manera el grupo control pertenecería a un solo *batch* (el de GTEx), cuyo efecto se corrigió entre muestras del mismo (véase 3.3.5.2); y (3) la cantidad de muestras control y de enfermedad están equilibrados.

2. Se unieron las matrices de cuentas de Desai (después del paso anterior) y la de GTEx (véase subsección 3.3.5) usando R y el paquete de dplyr (Wickham et al. (66)).

3. Se ejecutó la corrección de *batches* usando el programa *ComBat_seq* para corregir por *datasets* (es decir, los *batches* correspondían al *dataset* de procedencia). En este caso, no se usó una covariable que tuviera en cuenta la variabilidad de condiciones biológicas ya que este factor se confunde con la variabilidad técnica. Bajo este escenario, priorizamos la reducción de ruido técnico a la detección de señal biológica, lo cual es razonable dado que es un mayor problema a tratar en experimentos bioinformáticos.

4. Para reducir la complejidad del análisis posterior, se redujo el *dataset* conjunto para abarcar un transcrito por gen. Esto se realizó usando el argumento $filter\_transcripts = TRUE$ en el código MakeExperimentCountMatrixes.R y usando como referencia gencode versión 38.

Esto dio como resultado la matriz de cuentas conjunta para el análisis de distintos tejidos con COVID-19.

3.3.6.3. Análisis Pulmón con COVID-19 en comparación con sano: Datos Blanco-Melo, Desai y Delorey

1. Se unieron las matrices de cuentas generadas con los datos de (1) Blanco-Melo (véase sección 3.3.2), (2) de Desai (véase sección 3.3.3) y (3) de Delorey (véase sección 3.3.4) usando R y el paquete de dplyr (Wickham et al. (66)).

2. Se ejecutó la corrección de *batches* usando el programa *ComBat_seq* para corregir por *datasets*, tomando en cuenta la covariable perteneciente a la condición biológica de interés, es decir, muestras sanas y enfermas.

3. Se removieron los datos de muestras que no eran provenientes de tejido pulmonar.

4. Para reducir la complejidad del análisis posterior, se redujo el *dataset* conjunto para abarcar un transcrito por gen como se mencionó anteriormente.

Esto dio como resultado la matriz de cuentas conjunta para el análisis de pulmón en COVID-19.

3.4. Exploración de datos con PCA

Para conocer el perfil transcripcional de los datos de cada análisis, se recurrió a usar el método de Análisis de Componentes Principales (PCA), el cual nos permite visualizar datos multi-dimensionales (cada muestra varía en la expresión de 20,299 genes) en una dimensión baja (dos en este caso) localizados por los componentes de mayor variabilidad. Estos componentes de variabilidad son combinaciones lineales, específicamente *eigenvectores*, que capturan la variabilidad de la expresión génica a través de todas las muestras. A mayor cercanía entre puntos de datos (muestras en nuestro caso) mayor similitud del perfil transcriptómico de las muestras.

Se tuvieron distintos objetivos en este análisis:

1. Evaluar que la corrección de *batches* haya funcionado, y de ser necesario, aplicar una segunda corrección para eliminar el efecto de lote a su mayor extensión.

   - Para clarificar, la primera corrección corresponde a la realizada en un mismo *dataset* y la segunda, a la realizada entre *datasets*. Véase sección 3.3.
   - Lo que se esperaría ver si la corrección de batches fue exitosa, es que (1) los datos de cada *dataset* no estén agrupados por batch dado que se espera heterogeneidad entre las condiciones y entre las muestras *per se* dado que son datos tipo *bulk*; y de manera óptima, que (2) los datos entre conjuntos de datos se mezcle.

2. Corroborar que las condiciones biológicas entre *datasets* (una vez aplicado el paso de corrección) se parecieran entre sí.

3. Hacer una descripción general del perfil transcriptómico de las muestras.

4. Evaluar la similitud transcripcional de las células epiteliales alveolares tipo II (AT2) con las muestras de líneas celulares provenientes de epitelio pulmonar.

Para este análisis se utilizó la función *prcomp* de la paquetería básica de R. Como nota, esta función, al ejecutar transformaciones matemáticas, no admite ausencias de cuentas (simbolizadas como "NA" en R) en la matriz introducida; por lo que fue necesario convertir estas cuentas a ceros. Adicionalmente, se descartaron los genes cuya expresión fue nula ya que su contribución a la variabilidad transcriptómica era a su vez nula. Las gráficas se hicieron utilizando los paquetes *ggplot*2 y *dplyr* (para manejar datos) de *tidyverse* (Wickham et al. (66)), y *ggrepel*, que permite colocar etiquetas a puntos de datos en gráficas sin que estas se sobrepongan.

El código para el análisis del perfil transcripcional de los tres análisis principales puede ser consultado en el reporte de rmarkdown correspondiente en github.

Para el cuarto objetivo fue necesario usar datos nuevos y transformarlos para su comparación con los datos de líneas celulares. Esta metodología se describe a continuación:

1. Se tomó la matriz de cuentas génicas provenientes de la tecnología RNA-seq de célula única (abreviado en inglés como *scRNA-seq*) de Liao et al. (42), correspondientes a células inmunológicas (células B,T, Natural Killer (NK), células dendríticas mieloides (mDC abreviado en inglés), mastocitos (Mast), macrófagos (Macrophages), células dendríticas plasmocitoides (pDC abreviado en inglés), plasma y neutrofilos (Neutrophil en inglés)) y epitelial tomadas de lavado bronco alveolar (BAL, por sus siglas en inglés) de pacientes con COVID-19.

2. Las cuentas por célula se transformaron a cuentas *pseudo-bulk* (llamado así al proceso de combinar cuentas de célula única para asemejar cuentas de una muestra en *bulk*) usando un código compartido por Leo Arteaga, miembro entonces del laboratorio de la Dra. Alejandra Medina y colaborador del proyecto, ocupando la función *aggregateAcrossCells* del paquete *scuttle* McCarthy et al. (46).

3. Posteriormente, estas cuentas se transformaron a TPM con un código en R.

4. Se unieron las matrices de cuentas de *pseudo-bulk* con la de líneas celulares y se realizó la corrección de *batches* por la combinación de dos variables técnicas: rondas de secuenciación y tecnología implementada (*i.e. bulk* o *single-cell*).

5. Se llevó a cabo el análisis de PCA como se describió en esta misma sección anteriormente.

El código de esta última parte puede ser consultado en github aquí y aquí.
Véase la sección 4.2 para consultar los resultados de este análisis.

3.5. Construcción de redes de regulación

Se eligió utilizar la *pipeline* de SCENIC (véase sección 2.4.3) para realizar la reconstrucción de las redes de regulación (en este caso en subredes o módulos conocidos como regulones) principalmente por dos razones: (1) Los algoritmos de co-expresión utilizados en el primer paso de la *pipeline* (GENIE3 y GRNBoost2) se han encontrado como los mejores en comparación con otros métodos de inferencia de GRNs con respecto a la reproducibilidad (esto es, la capacidad de recuperar la biología desde estudios independientes para una misma condición biológica) y precisión de resultados para modelos curados y datos experimentales (Kang et al. (34) y Pratapa et al. (51)) y (2) el segundo paso de la *pipeline* de SCENIC, que consiste en el refinamiento de las redes de regulación construidas por co-expresión en base al enriquecimiento de motivos del factor transcripcional (TF, por sus siglas en inglés) en las regiones *cis* regulatorias de sus genes diana, ayuda a eliminar falsos positivos y de interés particular, de hecho integra un mecanismo de regulación asociado a la red.

3.5.1. Actualización de base de datos cistarget

El segundo paso *pipeline* de SCENIC, mencionado previamente, está facilitado por la computación *a priori* del enriquecimiento de los motivos de todos los TFs humanos en las regiones *cis* regulatorias de todos los genes del genoma humano conocidos usando el programa $cluster - buster$ (Frith (20)), cuyos resultados se guardan actualmente en la base de datos de cistarget que puede ser consultada aquí.

Es importante recalcar aquí que al momento de correr la *pipeline* de SCENIC y de consultar dicha base de datos, esta se encontraba desactualizada a partir de los datos con los que esta se había generado, es decir, los motivos de TFs y las regiones regulatorias, cuyas bases de datos eran previas al 2017 y mismas que se han estado actualizando desde entonces y hasta el año de la realización del trabajo de esta tesis. Adicionalmente, específicamente las regiones regulatorias usadas para la *pipeline* de SCENIC, correspondían a regiones 10 kilo bases (kb) río arriba y río abajo de sus genes diana, donde una definición más precisa era posible.

A continuación se describe la metodología que se siguió para actualización de la base de datos de cistarget a partir de la búsqueda, selección y procesamiento bioinformático de motivos y regiones regulatorias de manera *genome-wide* y, a diferencia de la base de datos original, específica para humano (o en su falla en la búsqueda de motivos, vertebrados), es decir, abarcando los motivos para todos los TFs humanos conocidos, todos los genes humanos conocidos y a su vez, todas sus correspondientes regiones regulatorias.

3.5.1.1. Motivos

Bases de datos Los criterios de selección para bases de datos o colecciones de matrices PWMs (definido en 2.3.3) para motivos de TFs fueron los siguientes:

- Que las matrices provinieran de experimentos realizados en muestras biológicas humanas o representaran motivos de TFs compartidos en vertebrados.
- Evitar usar matrices generadas por algoritmos de predicción como cisbp (en contraste a las generadas por experimentos como ChIP-seq), que son menos confiables.
- No incluir matrices representando motivos de dímeros a menos que provinieran de la colección JASPAR (descrita más adelante).

  Razón : Evitar un posible sesgo. Por ejemplo, haber usado dímeros habría dado lugar al caso en que, en pasos posteriores, se generaran regulones (definido en la sección 2.4.3) para el TF A, el TF B y el TF dímero AB, cada uno con su propio *score* de actividad. En etapas subsecuentes del análisis posterior a la generación de redes, se filtran regulones con respecto a ese *score*, que habría dado lugar al riesgo de descartar TFs importantes (TF A y TF B en este ejemplo) por el hecho de tener su *score* repartido con un tercero, el TF dímero (TF AB). Sin embargo, generar una versión con dímeros, para poder generar regulones guiados por TFs dímeros y compararlo con la versión sin dímeros, sería interesante para el presente y futuros estudios.

Como nota, las colecciones de matrices de distintas fuentes incluyen a conjuntos distintos de TFs, por lo que tomar en cuenta una mayor cantidad de bases de datos fue necesario para poder abarcar todos los TFs humanos.

En la Tabla 3.6 se describen las bases de datos de motivos seleccionadas prioritarias bajo estos criterios. En conjunto contemplan 7400 PMWs y provienen de los estudios de: Castro-Mondragon et al. (8), Xie et al. (69), Santana-Garcia et al. (55), Kulakovskiy et al. (39), Dogan et al. (17), Jolma et al. (31) y Kheradpour and Kellis (37).

Base de Datos	Num. Motivos	Año	Experimento	Fuente	Artículo
JASPAR 2022 vertebrates nonredundant	2410	2022	ChIP-seq; SELEX; ChIP-exo	link	JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles

hPDI	437	2010	FAIRE-seq; protein microarray assays	link	hPDI: a database of experimental human protein–DNA interactions
HOCOMOCO	771	2017	ChIP-seq	link	HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis
Zinc Fingers Dogan 2020	109; 126	2020	ChIP-seq; ChIP-exo	link	A domain-resolution map of in vivo DNA binding reveals the regulatory consequences of somatic mutations in zinc finger transcription factors
Jolma 2013	818	2013	SELEX; ChIP-seq	link	DNA-Binding Specificities of Human Transcription Factors
ENCODE Kheradpour 2013	2065	2013	ChIP-seq; ChIP-chip	link	Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments

Tabla 3.6: Bases de datos de motivos prioritariamente seleccionadas para la actualización de la base de datos cistarget feather.

De manera complementaria, se utilizaron colecciones de matrices provenientes de experimentos bioinformáticos por algoritmos de predicción, desactualizadas o de algún estudio que se publicó durante este análisis cuando no fue posible recuperar alguna matriz representativa de algún TF humano de las bases de datos anteriores. Estas bases de datos se mencionan en la Tabla 3.7 y provienen de los estudios de Hernandez-Corchado and Najafabadi (25), Contreras-Moreira and Sebastian (13), Weirauch et al. (65), Zhu et al. (75) y Yin et al. (71).

Base de Datos	Año	Fuente
C2H2 zinc finger proteins	2022	link
footprintDB	2020	link
footprintDB-metazoa	2020	link
cisBP_Homo_sapiens	2018	link
NCAP-SELEX	2018	link

CAP-SELEX	2018	link
HT_Selex	2018	link
Homer	2016	link

Tabla 3.7: Bases de datos de motivos complementarias para la actualización de la base de datos cistarget feather. Excepto las de footprintDB, todas son de humano.

Selección por similitud y calidad Se optó por seleccionar matrices individuales bajo un criterio de especificidad, es decir, seleccionar una matriz por motivo (correspondiente a un TF). Esto por las siguientes razones:

Razón 1: Evitar un posible sesgo. Para elaborar, el programa $cluster - buster$ escencialmente asigna un TF a una sección de una región regulatoria (denominada como *'cluster'*) si encuentra que la matriz correspondiente está más enriquecida que otras a través de un *score*. En el hipotético caso de usar dos matrices para representar un mismo TF A, se generan dos *scores* (menores que el score de haber usado una sola matriz) que podrían competir por un mismo *cluster* con otro TF B cuyo *score* podría solo ser más alto porque el score del TF A fue repartido en dos.

Razón 2: Reducir el tiempo computo. El programa para generar la base de datos de cistarget (create_cisTarget_databases), que a su vez corre por dentro $cluster - buster$, no es rápido. Por ejemplo, para este análisis tardó una semana en correr. Además, muchas de las matrices existentes para un mismo TF en distintas bases de datos son muy similares, por lo que en la mayoría de los casos no es necesario representar un TF con dos matrices.

Se razonó que una manera sencilla de quedarse con una matriz por TF era descartando aquellas que fueran muy similares entre sí. Para este objetivo se utilizó el programa $compare - matrices$ de RSAT (Santana-Garcia et al. (55)), que compara matrices a través de varias métricas: distancia euclidiana, coeficiente de correlación de Pearson(cor), correlación de Pearson normalizada ($Ncor$), etc. Este programa tiene la ventaja de comparar colecciones de matrices completas, lo cual redujo el tiempo de cómputo.

Para establecer un umbral de las métricas de similitud de matrices se consultó el artículo del programa $matrix-clustering$ (Castro-Mondragon et al. (7)), cuyo objetivo es reducir la redundancia entre matrices. Ellos determinaron en su estudio, a través de una evaluación comparativa entre las métricas y qué tan bien estas eran congruentes con matrices cuya similitud se sabía al pertenecer a las mismas familias de TFs, que dos matrices cuyas métricas de comparación sean $cor >= 0.6$ y $Ncor >= 0.4$ son lo suficientemente similares como para resumirse (al aplicar un paso de clusterización) en una sola matriz. Por lo tanto, para descartar matrices muy similares entre sí se utilizó un umbral de $cor >= 0.7$ y $Ncor >= 0.5$ al comparar dos matrices dadas

(donde la segunda dada como *input* era la descartada). Este umbral es más flexible que el establecido por $matrix-clustering$ porque el objetivo no fue descartar TFs que pertenecieran a una misma familia de motivos. Adicionalmente, se aplicó un criterio de prioridad en la selección de matrices respecto a la base de datos utilizada, el orden de prioridad es conforme a está enlistado en las tablas anteriores.

La *pipeline* y los códigos desarrollados para hacer esta selección de matrices y el manejo de datos puede ser consultado en github. Dicha pipeline fue generada para correrla en línea de comandos y está compuesta de los programas $compare-matrices$ y $convert-matrix$ de RSAT, de comandos básicos de unix, bash scripting, perl y awk. Asimismo, los códigos para manejar matrices fueron escritos en perl.

También se le dió mantenimiento al programa $convert-matrix$ de RSAT, cuya función es convertir formatos y tipos de métricas de matrices PWMs, para hacer funcional la opción de convertir matrices del formato "transfac" (definido en 2.3.3) al formato "cluster-buster" (que era el necesario para la generación de la base de datos de cistarget). Esto puede ser consultado en el commit correspondiente en github.

Datos finales Después de seleccionar matrices de la colección prioritaria (Tabla 3.6) se tuvieron 2291 matrices representando 2098 TFs. De los cuales, 1926 TFs tenían una matriz representativa, 153 TFs tenían dos matrices, 17 TFs tenían tres matrices y 2 TFs tenían cuatro matrices.

3.5.1.2. Regiones regulatorias

Bases de datos Se buscaron las versiones actualizadas de las bases de datos de regiones regulatorias en el genoma humano mencionadas en Imrichová et al. (30). Como nota, todas las bases de datos consultadas tenían una versión más actual que la mencionada en dicho artículo, con la excepción de UCNE (descrita a continuación). Así mismo se consultaron otras bases de datos de este tipo que pudieran ser complementarias. Las bases de datos recuperadas para este estudio se especifican en la Tabla 3.8.

Se generaron dos colecciones de regiones regulatorias: "essential" y "meta", donde la primera considera las regiones genómicas o bases de datos consideradas como esenciales para el genoma regulatorio y la segunda engloba el resto excepto ReMap (porque se demostró que sus regiones eran menos específicas para la captura de motivos que otras bases de datos (Garcia-Alonso et al. (21))) y dbCNS (porque otrass base de datos, ANCORA y VISTA, ya contenían las secuencias correspondientes).

Tipo de secuencia	Base de Datos	Año	Fuente	Formato	Uso
representative DNase hypersensitive sites for GRCh38 (DHS)	ENCODEv3	2021	link	bigBed/ bed 3+	meta, essential

regulatory elements ORegAnno	ORegAnno 3.0	2015	link	bigBed/ bed 3+	meta
Candidate Regulatory Elements	ENCODEv3	2020	link	bigBed/ bed 9+	meta, essential
VISTA Enhancers	ENCODEv3	2016	link	bigBed bed3+; hg19	meta, essential
Promoters from EPD-new	UCSC	2013	link	BED	meta
ReMap	UCSC	2022	link	NA	NA
CpG islands	UCSC	-		BED	meta
Ultra Conserved Non-coding Elements	UCNE base	2013	link	BED; hg19	meta, essential
CTCF binding sites	tomado de "encodeCcre-Combined" de UCSC	2020	link	BED	meta, essential (exclu-sión)
Conserved Non-coding Sequences	dbCNS	2021	link	fasta	NA
Conserved Non-coding Sequences between Human and Mouse	ANCORA-dbCNS	2008	link	BED	meta

Tabla 3.8: Bases de datos de regiones regulatorias para la actualización de la base de datos cistarget feather. Genoma de referencia hg38 a menos que se indique lo contrario.

Inicialmente se tuvo la intención de dividir la colección etiquetada como "essential" en dos: "tissues" y "cell lines", provenientes de las "Candidate Regulatory Elements" de la versión 3 de ENCODE; donde la primera correspondería a experimentos (regiones de cromatina abierta) en los tejidos de interés del estudio presente, es decir: pulmón, corazón, hígado, intestino y riñón y la segunda a datos de líneas celulares (A549 entre ellas) y experimentos de tejido primario de células inmunológicas. Sin embargo, en pasos subsecuentes del análisis se comprobó que ambas subcolecciones contenían el mismo conjunto de regiones genómicas, lo cual era una posibilidad dado que ambas subcolecciones pertenecían a muestras biológicas muy diversas. Por tal razón, se perdió el motivo inicial detrás de la elaboración de estas dos subcolecciones y se dejaron como una sola colección: "essential".

Como nota importante, de las bases de datos enlistadas en la Tabla 3.8, solo los motivos del factor transcripcional y de represión CTCF se usaron como regiones a excluir.

Procesamiento de secuencias El procesamiento de las regiones regulatorias que se describe a continuación requirió que se partiera de archivos en formato BED (tipo de formato que se utiliza para anotar coordenadas genómicas indicando en su formato básico: el cromosoma, el inicio de la región por la primera base y el término indicando la última base), por lo que aquellas bases de datos que partieran de un formato distinto (mencionadas en la Tabla 3.8) fueron convertidas a formato BED utilizando el programa $ucsc - bigbedtobed$ (Kuhn et al. (38)). Asimismo, para las bases de datos que partieran del genoma de referencia hg19 se realizó un *liftOver* (proceso por el que se transforman las coordenadas genómicas entre distintos ensambles de genomas) a la referencia hg38 utilizando el programa $liftover$ (Kuhn et al. (38)).

Para delinear las regiones regulatorias del genoma humano con las colecciones previamente mencionadas, se tomó como referencia el procesamiento de secuencias especificado en el artículo de cistarget (Imrichová et al. (30)), añadiendo modificaciones y su respectiva metodología. Para ejecutarlo, se utilizaron los programas de la suite de herramientas para el manejo de archivos BED: bedtools versión 2.27.1 (Quinlan (52)): sort, merge, annotate, genomecov, closest; comandos de unix: sort, cut, cat; awk y R. El procesamiento se presenta esquematizado en la Figura 3.2 y constó de los pasos detallados a continuación:

1. Unir regiones que estuvieran sobrepuestas.
2. Remover regiones que fueran cubiertas en un 80 % por cualquier región codificante (esto es, exónica) del genoma o por un 20 % elementos de tipo *insulator* (aislantes).

   Las regiones codificantes corresponden a la anotación de gencode versión 38 disponible en la base de datos de UCSC Genome Browser usando como referencia el ensamblaje de genoma hg38.

   Para *insulator elements*, se utilizaron los sitios de pegado del represor transcripcional CTCF ya que se encuentran en la periferia de los subcompartimientos del genoma conocidos como TADs donde se ha demostrado que CTCF tiene actividad de *insulator* (véase sección 2.2).
3. Remover regiones más cortas que 30 pb.
4. Extender regiones más cortas que 1000 pb a 1000 pb en una dirección que previniera su sobreposición con *insulator elements* o exones. Se razonó que una manera de hacer esto sin tomar arbitrariedad en la dirección, era unir las regiones que estuvieran cercanas.

   4.1. Se calculó d como la longitud mediana entre regiones.

   4.2. Unir regiones que estuvieran a máximo una distancia d (315 pb en este caso).

   4.3. Quedarse con aquellas regiones que fueran cubiertas en un 10 % o menos por *insulator elements* o exones (este filtro es más estricto que el anterior porque las regiones son más grandes) y separar aquellas que tuvieran un mayor porcentaje (por simplicidad denominadas "regiones fo').

4.4. Tomar regiones fo y regresarlas a su estado previo a la unión de regiones (paso 4b).

4.5. Repetir los pasos 4b a 4d para las regiones fo pero con una d más corta hasta que el promedio de las regiones fuera más cercano a 1000 pb o quedaran muy pocas regiones fo.

4.6. Unir las regiones regulatorias con las que se decidió quedarse y las últimas fo.

5. Asignar los genes diana de las regiones regulatorias. Esto se realizó de tres maneras: (1) Por anotación previa de las bases de datos de UCNE o EDPnew Promoters, (2) por proximidad en general: se supuso que el gen más cercano de cada región regulatoria era diana y (3) por proximidad para regiones putativamente *enhancers*, se tomaron los cuatro genes más cercanos al razonar que los *enhancers* regulan al conjunto de genes encapsulados en un mismo TAD.

5.1. Para el primer caso, cualquier región incluyendo las regiones regulatorias de UCNE o EDPnew compartieron la anotación de genes de estas bases de datos.

5.2. El segundo caso fue aplicado a todas las regiones regulatorias de ambas colecciones.

5.3. Para el tercer caso, para definir las regiones "putativamente *enhancers*" se tomaron todas las regiones regulatorias excluyendo aquellas que fueran cubiertas por 15 % o menos por los promotores reportados por EDPnew.

5.4. Se unieron las anotaciones.

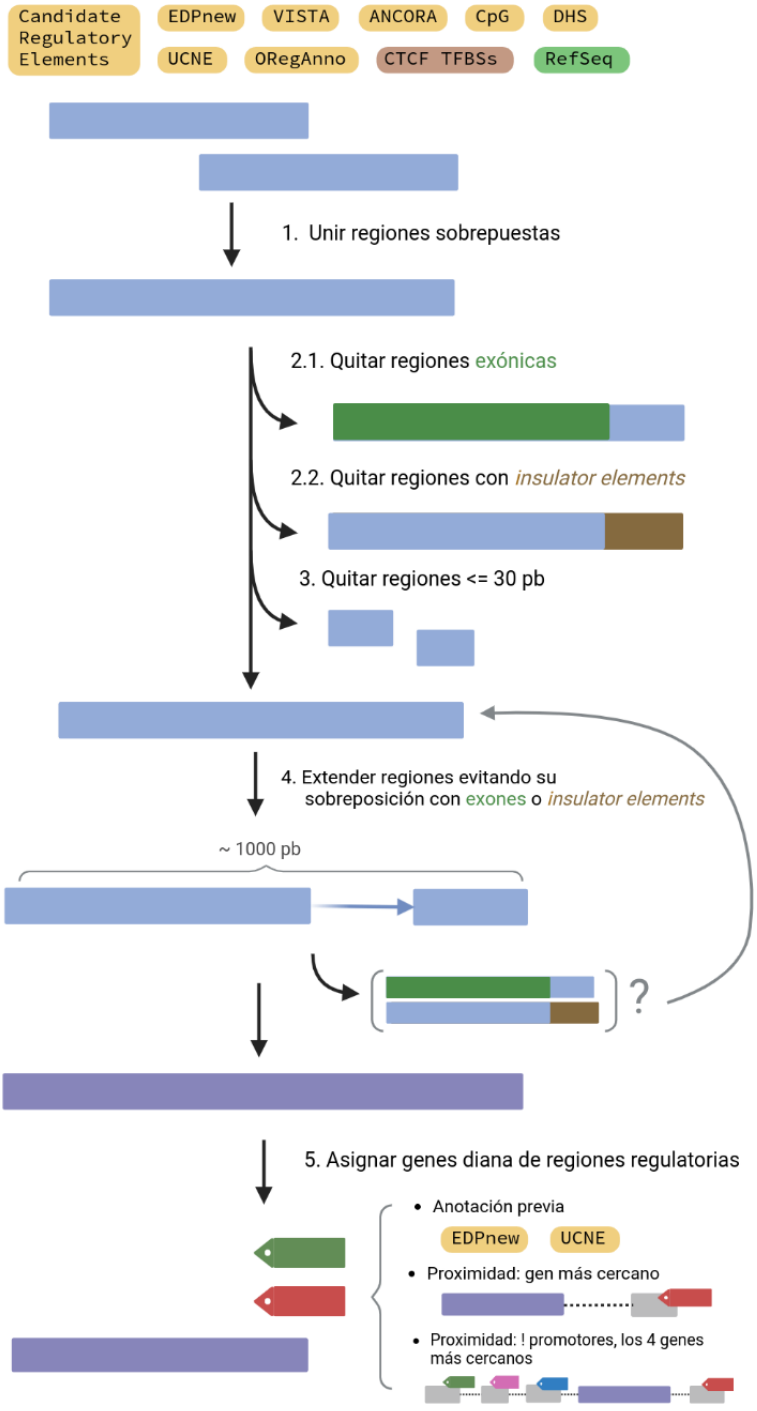


Figura 3.2: Esquema del procesamiento de regiones regulatorias para su uso en la actualización de cistarget. En la parte superior se mencionan las bases de datos de secuencias regulatorias utilizadas. Los rectángulos azules representan las regiones en procesamiento y los morados las regiones regulatorias resultantes. Se obtuvo un archivo BED de las regiones regulatorias putativas con la anotación de genes diana.

Finalmente, se calculó la cobertura del genoma usando como referencia el genoma hg38 descargado del UCSC browser y se obtuvieron las secuencias fasta de las regiones regulatorias candidatas utilizando el programa $fetch - sequences$ de RSAT.

El código a detalle puede ser consultado en la sección "Candidate Regulatory Regions: Processing" en el reporte correspondiente de github.

Datos finales

- Colección "essential':
  - Número total de regiones: 1,161,542
  - Promedio de longitud: 753 pb
  - Mediana de longitud: 337 pb
  - Cobertura del genoma: 27 %

3.5.1.3. Generación de base de datos cistarget

Se utilizó el programa *create_cistarget_motif_databases.py* generado por los autores de CISTARGET y SCENIC. Los archivos de entrada fueron las regiones regulatorias candidatas en formato fasta y con anotación de genes (meta y essential), el directorio conteniendo todas las matrices PWMs individuales en formato cluster-buster, un archivo indicando el nombre de todas matrices a utilizar y el prefijo del archivo de salida tipo feather.

Se generaron las bases de datos cistarget feather actualizadas nombradas essential y meta.

Adicionalmente, se creó una matriz de anotación de motivos siguiendo el ejemplo de la matriz consultada en línea que fue necesaria para el paso dos de la *pipeline* de SCENIC (véase sección 2.4.3).

3.5.2. Generación de redes de regulación con SCENIC

3.5.2.1. Estandarización de *pyscenic pipeline* para datos de *bulk*.

Se requirió hacer algunas adaptaciones de la *pipeline* de SCENIC dado que esta espera que los datos de entrada sean provenientes de la tecnología de *single-cell* RNA-seq pero los que usamos en el estudio presente son de *bulk*. Esta estandarización ocurrió en distintos pasos: en la preparación de los archivos de entrada, en la prueba de la *pyscenic pipeline* ejecutándolo solo una vez en línea de comandos y cuando finalmente se ocupó la *scenic multiruns pipeline* de $nextflow$ para ejecutarlo iterativamente.

Como primer paso, se probó la *pyscenic pipeline* en línea de comandos. En particular, se pretendía explorar la posibilidad de usar GENIE3 (en vez de la opción por default GRNBoost2) en el primer paso de SCENIC, ya que fue diseñado para datos provenientes de *bulk* y se ha visto que obtiene resultados más precisos (Pratapa et al.

(51)), sin embargo, se ha visto que consume un mayor tiempo de ejecución usando datos de *single-cell* (que suelen tener una dimensión más grande que los de *bulk*). Además, se tenía la intención de usar esta *pipeline* para encontrar resultados que se guiaran menos por parámetros iniciales (la semilla del algoritmo aleatorio para generar redes) y que en cambio, fueran más consistentes para tener más certeza sobre que las redes generadas recapitularan las existentes en las células. Por lo tanto, se verificó la diferencia en el tiempo de ejecución entre ambos programas usando los datos de Blanco-Melo.

Al probarlo una vez, el resultado fue que la ejecución de GENIE3 duró 8 horas contra 1 hora de GRNBoost2. Usando la misma semilla para el programa y los mismos archivos de entrada, los resultados diferían bastante, con GENIE3 encontrando el quíntuple de las interacciones TF-gen. Esto puso en perspectiva que al usar GENIE3 múltiples veces, el tiempo de ejecución aumentaría significativamente y que a pesar de que se encontraron más interacciones usando GENIE3, se razonó que al usar múltiples veces GRNBoost2, este sería capaz de capturar las interacciones importantes faltantes. También es importante mencionar que muchas de estas interacciones TF-gen, encontradas por GENIE3 y no por GRNBoost2, se generan para la estimación de un parámetro, proceso que se ve optimizado en el algoritmo de GRNBoost2, por lo que no son necesariamente relevantes.

A partir de esos resultados, se decidió utilizar el programa de GRNBoost2 por la significativa diferencia en el tiempo de ejecución y se razonó que usar esta herramienta múltiples veces, como veremos a continuación, sería suficiente para recuperar resultados más precisos.

3.5.2.2. Preparación de archivos de entrada

Para usar la *pipeline* de *scenic multiruns* de *vsn-pipelines* era necesario dar como archivo de entrada un *loom file* con la versión >= 3 de la paquetería loompy en python. Los archivos tipo *loom* fueron creados para contener información ómica con distintas capas de información (ya sean metadatos, matrices de cuentas, transformaciones de estas y gráficas) que serían muy pesados por si solos, por lo que el tipo de formato que se utiliza es de HDF5 que son menos pesados que el típico formato csv.

Utilizando python, se tomaron las tres matrices de cuentas génicas (tras corregir por efectos de lote) en formato csv correspondientes a los tres análisis descritos en la sección 3.1.4 y (1) se convirtieron los valores "NA" por 0, y (2) se guardaron en loom files individuales en una matriz de genes (filas) por muestras (columnas) como se describe en el código reportado en github.

3.5.2.3. *scenic multiruns pipeline*

Con el objetivo de generar redes de regulación, o en este caso subredes de regulación conocidas como regulones, que fueran consistentes con los datos, es decir, que recuperaran verídicamente los patrones de expresión génica de los datos, se utilizó la *scenic*

multiruns pipeline (parte de las *vsn-pipelines* versión 0.25.0) ocupando $nextflow$ versión 20.10.0 y *singularity* versión 3.7.0 para ejecutar pyscenic 100 veces por análisis y de esta manera, tomar solamente aquellos regulones y los correspondientes genes diana que se hayan reproducido al menos el 10 % y el 5 % de las veces, respectivamente.

A continuación se proveen algunos detalles técnicos para utilizar estas herramientas:

- Se ocuparon y cargaron los programas utilizados anteriormente.
- Se cargaron las *vib-singlecell-nf/vsn-pipelines* de nextflow 0.25.0 utilizando el comando $nextflow\ pull$.
- Se generó el archivo de configuración de la *pipeline* (o estrucutura de trabajo) indicando el uso de las entradas "loom, scenic, scenic_multiruns, scenic_use_cistarget_motifs, scenic_use_cistarget_tracks, hg38, singularity" en la opción $-profile$ con el comando $nextflow\ config$. Estas entradas indican: el tipo de archivo de entrada (loom), que la pipeline requerida es la de SCENIC, y en particular las opciones para correrla iterativamente, que se ocuparán la información de motivos y *tracks* para el segundo paso de pyscenic, que el genoma de referencia es el hg38 y que se ocupará singularity como herramienta para ejecutar este análisis.
- Se estableció una semilla de 777 (la elección de este número fue arbitraria) para posibilitar la reproducción de este análisis con resultados similares.
- Opciones cistarget:
  - Se fijó el uso de las bases de datos feather de cistarget versión 9 con su tabla de anotación correspondinete disponibles en la página web al declararlo bajo los argumentos de : $params.sc.scenic.cisTarget.motifsDb$ y $params.sc.scenic.cisTarget.motifsAnnotation$.
  - De manera relevante, se usó la lista de factores transcripcionales de esa misma base de datos que contiene 1892 TFs bajo el argumento $params.sc.scenic.grn.tfs$.
  - Otra **nota importante** es que las opciones mencionadas aquí son las que deben cambiarse para usar otra base de datos de cistarget, por ejemplo, la base de datos actualizada mencionada en la sección anterior.
- Se establecieron los parámetros de multiruns: $numRuns = 100$ (númpero de iteraciones de la pipeline), $min\_genes\_regulon = 5$ (mínima cantidad de genes diana por regulon) y $min\_regulon\_gene\_ocurrence = 5$ (mínima cantidad de ocurrencia de un gen diana para un regulon dado).

Como nota, dado que esta *pipeline* requiere de muchos programas, surgieron algunos problemas de incompatibilidad de versiones entre ellos en el proceso de estandarización del uso de la *pipeline* en sí. Se recomienda usar las mismas versiones de los programas en futuras ocasiones para evitar este tipo de inconvenientes, a menos de que los conflictos entre versiones se hayan especificado como resueltos en los nuevos lanzamientos de pyscenic.

De manera muy relevante, se utilizó el cluster SGE de DNA ubicado en el LAVIS para ejecutar estos programas. Se sometieron a un qsub y se solicitaron 160 GB de memoria RAM para cada trabajo.

En la Tabla 3.9 se resume el tiempo de ejecución y el números de regulones reconstruidos por análisis.

Análisis	Tiempo de cómputo	Núm. regulones
Infección de SARS-CoV-2 en comparación a otros virus en líneas celulares	11 días 11 hrs 19min 14 seg	898
COVID-19 en distintos tejidos	10 días 15 hrs 35 min 30seg	947
Pulmón con COVID-19 en comparación con sano	6 días 12 hrs 57 min 31 seg	780

Tabla 3.9: Tiempo de ejecución y número de regulones por análisis tras correr SCENIC 100 veces.

3.6. Selección de Regulones

Como se vio en la sección anterior y en el marco teórico (véase 3.5.2), el resultado de pyscenic fue un archivo formato *loom* conteniendo la matriz tamaño m muestras x n regulones con el *score* de activación AUC (véase sección 2.4.3) por regulon, donde en promedio se encontraron 875 regulones por análisis. La pregunta que surgió entonces fue: ¿cuáles de estos regulones son los relevantes durante la progresión de la enfermedad a partir la infección por SARS-CoV-2?

Para responder a esa pregunta, se emplearon tres metodologías para seleccionar regulones que se consideran *relevantes* para explorar en análisis posteriores, es decir, sobre los regulones seleccionados basándonos en las siguientes metodologías se elabora el resto del trabajo de la presente tesis. Las metodologías o pasos de filtrado fueron:

1. Mencionado previamente, el filtrado de regulones por reproducibilidad en la *SCENIC multiruns pipeline*.

2. Se buscaron los regulones diferencialmente activados (DA), con las pruebas estadísticas Mann Whitney U y Kolmogrov-Smirnov, en muestras infectadas. Particularmente, se buscaron aquellos regulones DA que tuvieran un *Fold Change* positivo (*i.e.* que estuvieran sobre-regulados) (véase seccción 3.6.2).

3. Se buscaron los regulones más específicos a una condición dada, específicamente, en las muestras infectadas. Esto, basado en una prueba adicional llamada *Regulon Specificity Score* descrita más adelante (véase sección 3.6.3).

Se generaron tres scripts tanto para realizar estas pruebas de selección de regulones como para analizar sus resultados a partir de, esencialmente, el archivo de salida loom y una tabla de metadatos de las muestras. Con el motivo de facilitar su uso, dos de ellos solo contienen el código necesario para realizar pruebas y filtros en funciones ($difregs\_.py$ y $FDRthreshold\_functions.R$) y están integrados en la estructura de otro, el principal, llamado $reticulateEnrich.R$; de esta manera solo se interactúa con este último para la especificación de los archivos de entrada mencionados, las variables para correr los otros programas y para realizar el análisis de resultados de las pruebas. Esta estructura permite además que estos scripts sean fácilmente adaptados a otros datos y en concordancia con ese objetivo, se encuentran debidamente documentados y disponibles en github.

En las siguientes subsecciones se describen a detalle estas metodologías y pruebas.

3.6.1. Filtrado de Regulones por reproducibilidad en SCENIC

Como se mencionó en secciones pasadas, es posible filtrar tanto los regulones por número de veces que aparecen en las múltiples iteraciones de la *pipeline* de pyscenic como los genes diana de dichos regulones, también, por número de veces que se encuentran como regulados por el TF que guía el regulon en cuestión.

Al revisar el archivo loom de salida después de ejecutar pyscenic 100 veces, la matriz de scores de activación AUC por regulón incluía a todos los regulones encontrados en todas las iteraciones de SCENIC sin filtro, así mismo, la lista de genes diana por regulón se encontraba sin filtro; sin embargo contenía la información del número de iteraciones, con lo que fue posible generar el código necesario en python para aplicar dichos filtros. La elección del número mínimo de iteraciones de regulones y el mínimo de ocurrencia de genes diana se dejó como variable de elección al usuario pero por defecto se toman aquellos regulones que aparecen al menos en el 10 % de las iteraciones de *scenic multiruns*.

A continuación se muestra en un histograma el número de veces que se generaron los regulones en cada análisis antes de aplicar este paso de filtrado, mostrando así la reproducibilidad de los regulones generados por SCENIC aquí.

- En la Figura 3.3 se muestra el histograma del análisis de líneas celulares.
- En la Figura 3.4 se muestra el histograma del análisis de COVID-19 en distintos tejidos.
- En la Figura 3.5 se muestra el histograma del análisis de Pulmón con COVID-19 en comparación con sano.

También, se muestra en un histograma el número de genes diana por regulon encontrado en cada análisis después de aplicar este paso de filtrado.

- En la Figura 3.6 se muestra el histograma del análisis de líneas celulares.

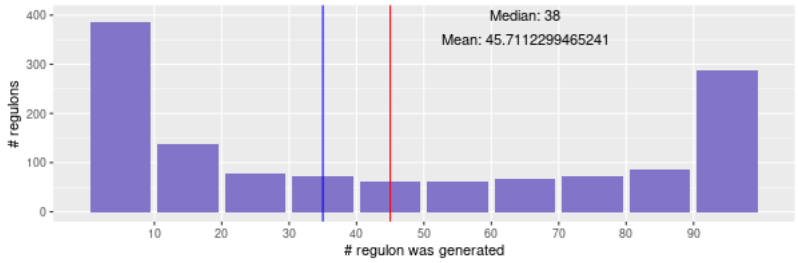


Figura 3.3: Histograma de reproducibilidad por regulon en el análisis de líneas celulares. Se menciona el promedio (*mean*) y la mediana (*median*).

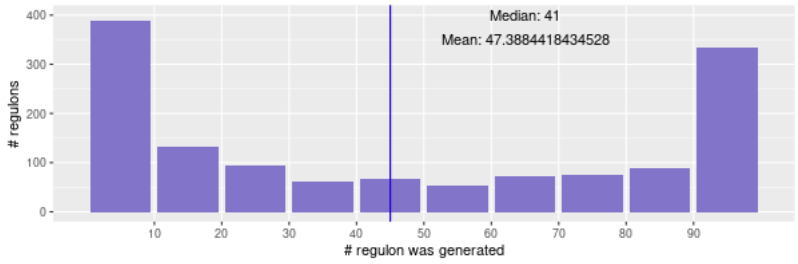


Figura 3.4: Histograma de reproducibilidad por regulon en el análisis de COVID-19 en tejidos. Se menciona el promedio (*mean*) y la mediana (*median*).

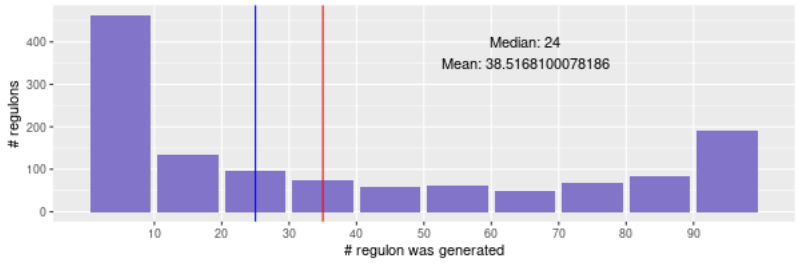


Figura 3.5: Histograma de reproducibilidad por regulon en el análisis de COVID-19 en pulmón. Se menciona el promedio (*mean*) y la mediana (*median*).

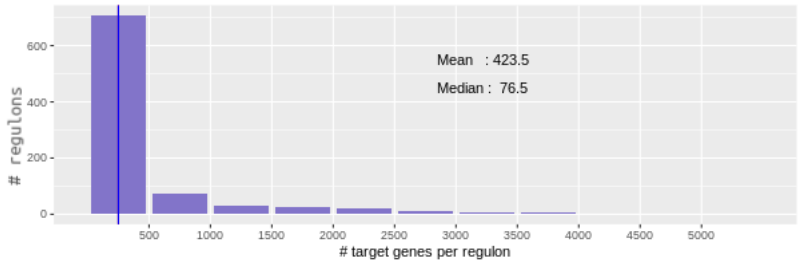


Figura 3.6: Histograma del número de genes diana encontrados por regulon en el análisis de líneas celulares. Se menciona el promedio (*mean*) y la mediana (*median*). Las líneas verticales indicando estas métricas colapsan estas son muy cercanas.

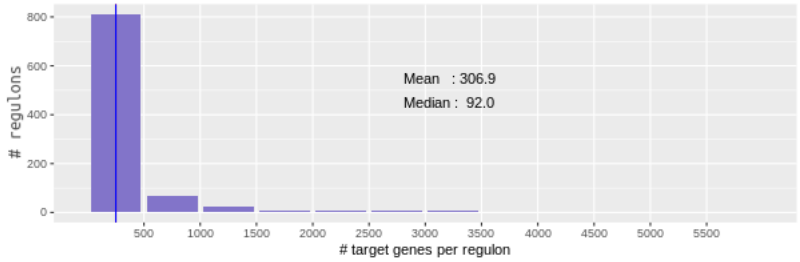


Figura 3.7: Histograma del número de genes diana encontrados por regulon en el análisis de COVID-19 en tejidos. Se menciona el promedio (*mean*) y la mediana (*median*).

- En la Figura 3.7 se muestra el histograma del análisis de COVID-19 en distintos tejidos.
- En la Figura 3.8 se muestra el histograma del análisis de Pulmón con COVID-19 en comparación con sano.

3.6.2. Regulones Diferencialmente Activados

En términos generales, se buscaron los regulones que estuvieran diferencialmente activados (DA) en las muestras infectadas o de biopsias COVID-19 en comparación con sus respectivos controles utilizando dos pruebas estadísticas no-paramétricas: Mann Whitney U (MWU) y Kolmogrov-Smirnov (KS) para dos muestras.

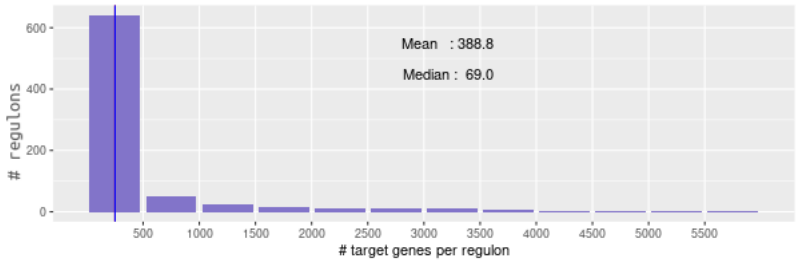


Figura 3.8: Histograma del número de genes diana encontrados por regulon en el análisis de COVID-19 en pulmón. Se menciona el promedio (*mean*) y la mediana (*median*).

3.6.2.1. Pruebas estadísticas elegidas

Se utilizaron pruebas no-paramétricas porque las características de nuestras mediciones no permitían hacer grandes suposiciones sobre ellas, en específico: (1) la cantidad de muestras es pequeña (la mínima cantidad de número de muestras por condición fue dos, véase Tabla 3.1), (2) no se conocía la distribución que siguen las mediciones de actividad de regulones por condición, y (3) como se mencionó en la sección 2.4.3, la métrica de actividad por regulon AUC de pyscenic es solo relativa a la expresión génica en cada muestra.

La prueba MWU hace la pregunta de si dos conjuntos de observaciones de dos variables X y Y provienen de la misma población (hipótesis nula) o de poblaciones distintas (hipótesis alternativa), donde las variables son independientes entre sí y las observaciones son (o pueden ser) ordinales o rankeadas. Por otro lado, la prueba KS para dos muestras hace esta misma pregunta pero tomando observaciones continuas.

La pregunta específica realizada entonces por análisis, por condición y por regulon fue: ¿Los scores de activación AUC de muestras infectadas pertenecen a una población o distribución distinta que los scores de activación AUC de muestras control?

3.6.2.2. Corrección de p-valores por comparación múltiple

Por análisis, los p-valores resultantes de hacer esta pregunta para todos los regulones se muestran en un histograma por condición analizada contra su respectivo control. Nótese entonces que decidir qué regulones están en DA solamente basándonos en un umbral de estos p-valores (e.g. p-valor $<= 0.05$) habría sido incorrecto y podría llevarnos a realizar errores Tipo I (donde se rechaza la hipótesis nula cuando esta era cierta); ya que al haber realizado la pregunta anterior múltiples veces (por cada regulon), se obtuvo una distribución de p-valores que a su vez está sujeta a la probabilidad de que aparezcan valores pequeños por azar. Por esa razón debe hacer una corrección de p-valores para ajustar por esta probabilidad.

Se realizó una corrección de los p-valores utilizando el procedimiento de Benjamini-

Hochberg (BH) cuyo resultado puede interpretarse como el *False Discovery Rate (FDR)*, esto es, la proporción de pruebas que aparecieron como significativas que en realidad no lo son. Es decir, si se tiene un FDR <= 0.1, se espera que el 10 % de los positivos (o p-valores ajustados) tomados sean falsos positivos.

El umbral específico aplicado sobre el p-valor ajustado para tomar regulones en DA cambió por condición. En particular, se buscó que el umbral máximo de FDR (entre 0.19 y 0.05) elegido permitiera que nos quedáramos con alrededor de 100 regulones por condición analizada por cada prueba (MWU y KS) dado que los regulones encontrados en DA se sometieron a filtros adicionales que subsecuentemente redujeron el número final de regulones "relevantes" por condición.

3.6.2.3. Histogramas de p-valores de pruebas de DA por análisis

Cuando la hipótesis nula es verdadera, se espera que la distribución de p-valores resultantes de la prueba estadística sea Uniforme en el rango [0,1]. En cambio, cuando la hipótesis nula es falsa, la distribución de p-valores tiende a estar sesgada al cero.

Infección de SARS-CoV-2 en comparación a otros virus en líneas celulares

- Histogramas de p-valores de la prueba MWU por experimento: Fig 3.9. Nótese que el eje x cambia entre gráficas. Referencias del experimento en la Tabla 3.1.
- Histogramas de p-valores de la prueba KS por experimento: Fig 3.10.
- Histogramas de p-valores de la prueba MWU por experimento y corregidos por comparación múltiple por BH: Fig 3.11.
- Histogramas de p-valores de la prueba KS por experimento y corregidos por comparación múltiple por BH: Fig 3.12.

La mayoría de los histogramas parece tener una forma de U, lo cual indica que muchas pruebas están acorde a la hipótesis nula, es decir, que los regulones no cambian entre condición control e infectada. Se puede notar que la prueba MWU fue más estricta que la de KS en los histogramas de p-valores sin corrección. Posteriormente, tras la corrección de BH, el sesgo de p-valores cercanos a 0 disminuye, indicando una reducción en el número de regulones que se encontrarán como DA. Particularmente importante, hay una significativa reducción en los p-valores del experimento "A549 infectado por SARS-CoV-2 y con vector transducido expresando ACE2 humano" y los p-valores de "Calu-3 infectado por SARS-CoV-2" aunque parecen buenos, por la escala se puede ver que no son tan bajos.

COVID-19 en distintos tejidos

- Histogramas de p-valores de la prueba MWU por tejido: Fig 3.13 A.
- Histogramas de p-valores de la prueba KS por tejido: Fig 3.13 B.

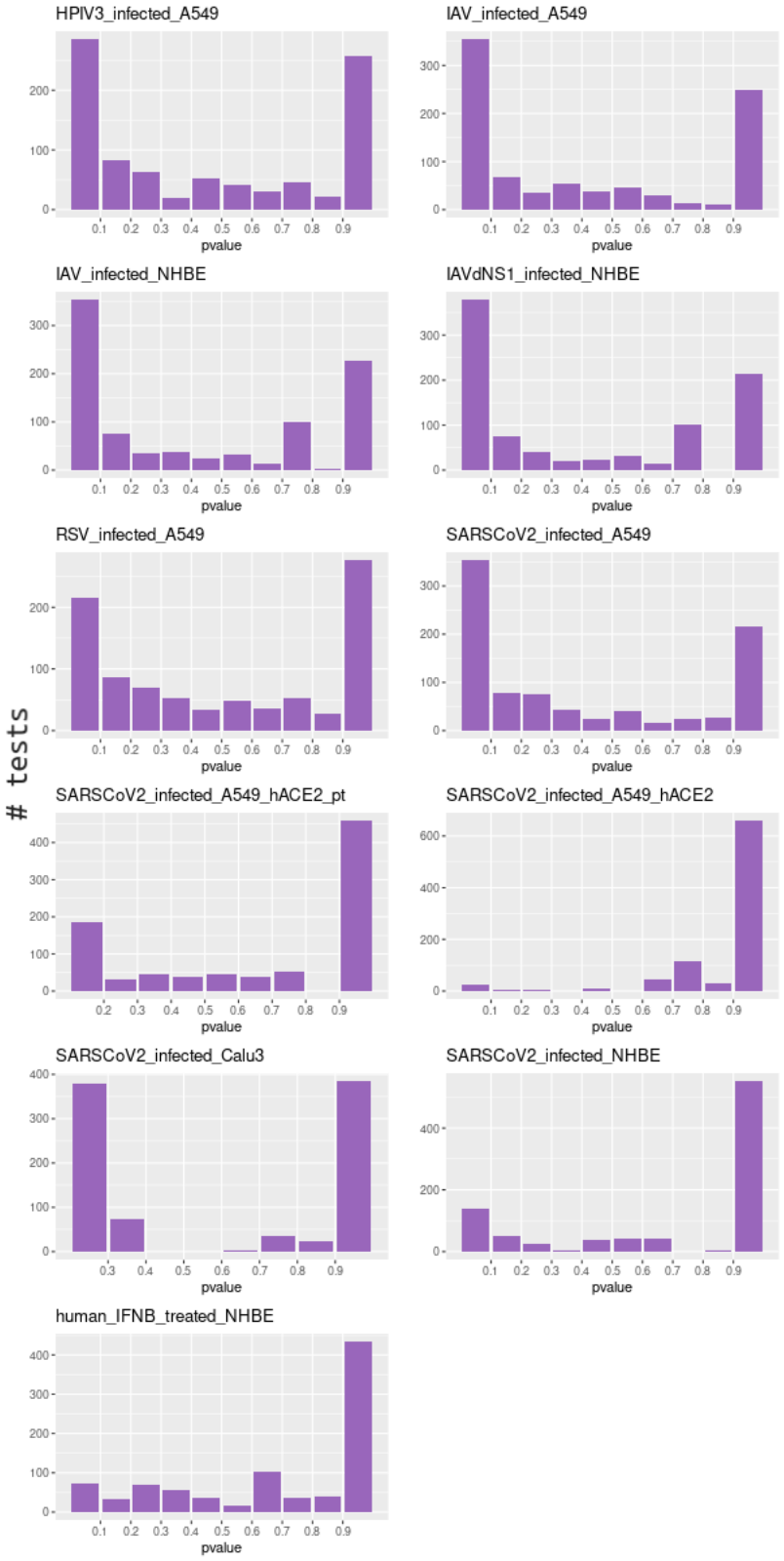


Figura 3.9: Histogramas de P-valores resultantes de prueba Mann Whitney U de Activación Diferencial de regulones por experimento.

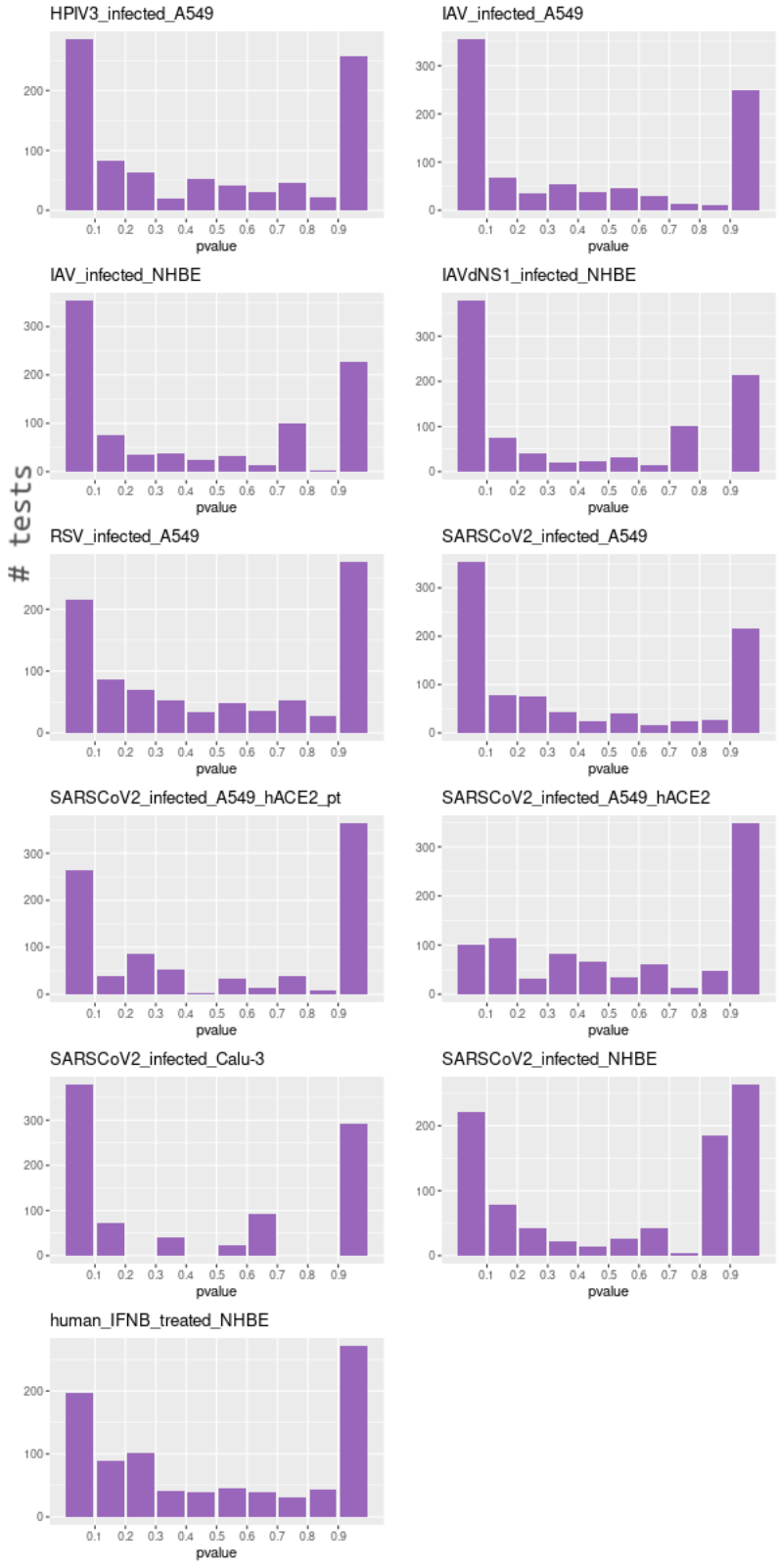


Figura 3.10: Histogramas de P-valores resultantes de prueba Kolmogrov-Smirnov de Activación Diferencial de regulones por experimento.

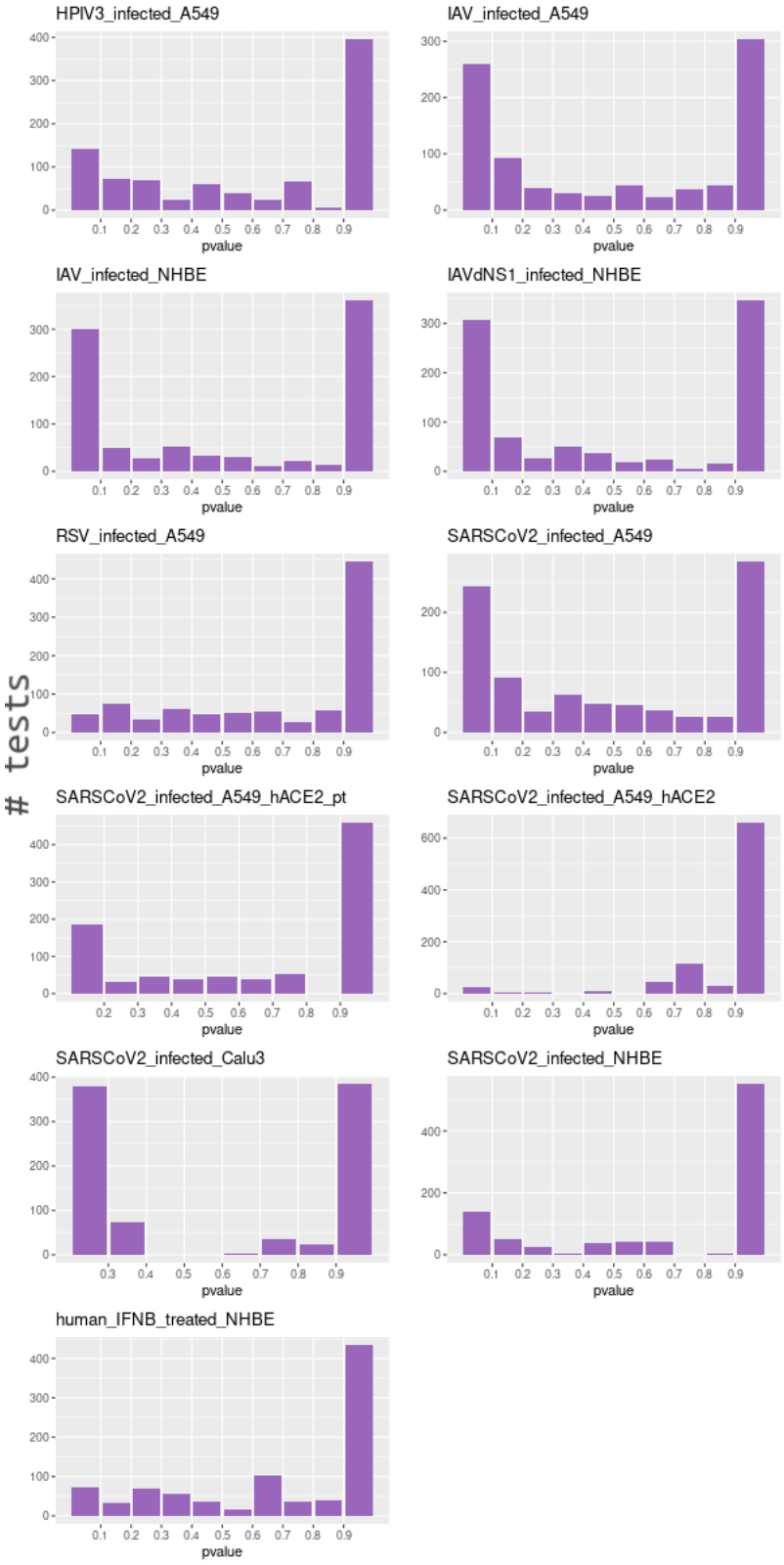


Figura 3.11: Histogramas de P-valores resultantes de prueba MWU corregidos por BH.

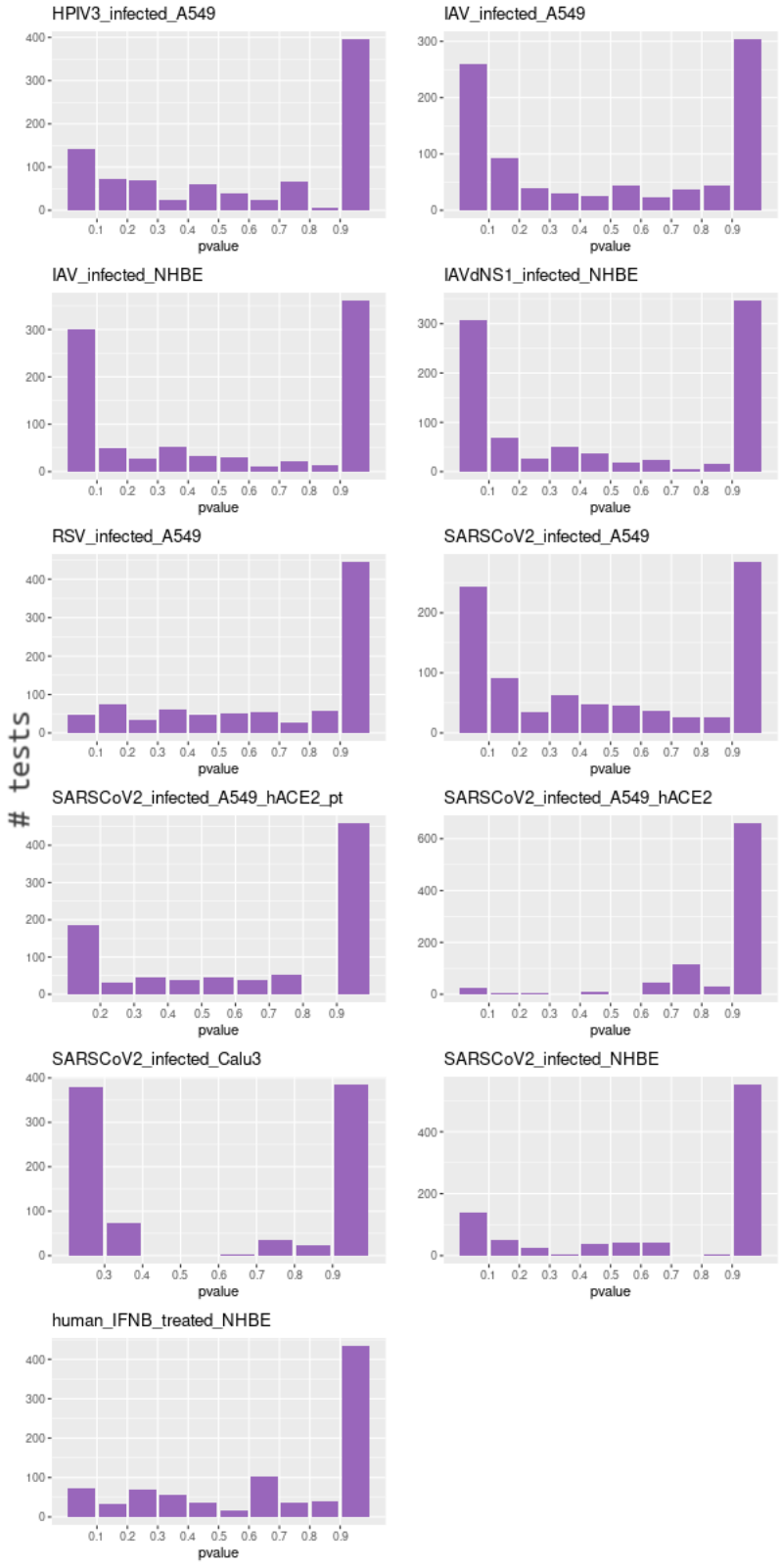


Figura 3.12: Histogramas de P-valores resultantes de prueba KS corregidos por BH.

- Histogramas de p-valores de la prueba MWU por tejido y corregidos por comparación múltiple por BH: Fig 3.14 A.
- Histogramas de p-valores de la prueba KS por tejido y corregidos por comparación múltiple por BH: Fig 3.14 B.

Todos los histogramas de p-valores, tanto con como sin corrección por comparación múltiple, muestran un sesgo al cero, lo cual indica que debe haber muchos regulones que estén diferencialmente activados en los tejidos COVID-19 en comparación con tejido sano.

Pulmón con COVID-19 en comparación con sano

- Histograma de p-valores de la prueba MWU: Fig 3.15 A.
- Histograma de p-valores de la prueba KS: Fig 3.15 B.
- Histograma de p-valores de la prueba MWU y corregidos por comparación múltiple por BH: Fig 3.15 C.
- Histograma de p-valores de la prueba KSy corregidos por comparación múltiple por BH: Fig 3.15 D.

Las pruebas MWU y KS parecen haber tenido resultados muy similares. En las gráficas sin corrección se nota el sesgo de p-valores no corregidos dado que los p-valores se distribuyen semi uniformemente excepto cerca del cero, lo que indica que hay regulones en activación diferencial. Al corregir los p-valores hay una baja de alrededor de 100 p-valores cercanos al 0, pero aún quedan bastantes.

3.6.2.4. Regulones sobre- ó sub- regulados

Para tomar los regulones sobre- y subregulados se calculó el log *Fold Change (LFC)* de las muestras infectadas sobre las muestras control. Se define que aquellos regulones con un LFC > 0 están sobreregulados, mientras que aquellos con un LFC < 0 están subregulados.

La realización de estas pruebas automática a partir de los archivos de entrada mencionados (la matriz AUC y la tabla de metadatos), así como el ajuste de p-valores se codificaron en python ocupando las funciones de los paquetes *scipy.stats* y *statsmodels.stats.multitest* como se puede consultar en el código disponible en github. El código de elección del umbral FDR en R también está disponible.

3.6.3. Regulones más específicos por condición

Adicionalmente, se utilizó la métrica de *Regulon Specificity Score (RSS)* (Suo et al. (58)), la cual cuantifica la especificidad de los regulones entre muestras. A *grosso modo*, el RSS encuentra los reguladores esenciales por condición al usar la Divergencia de

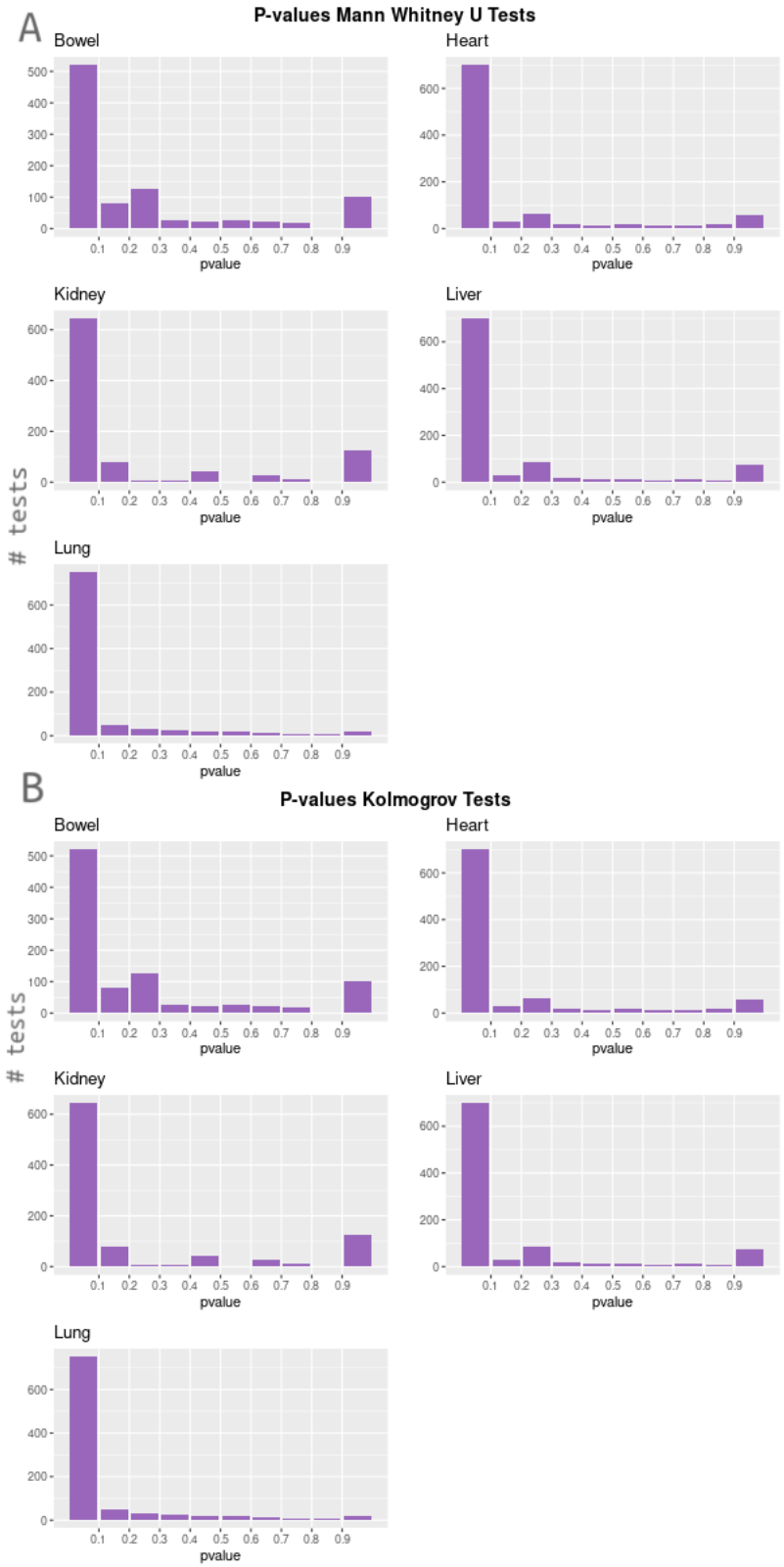


Figura 3.13: Histogramas de P-valores resultantes de pruebas de Activación Diferencial de regulones por tejido. A) MWU B) KS

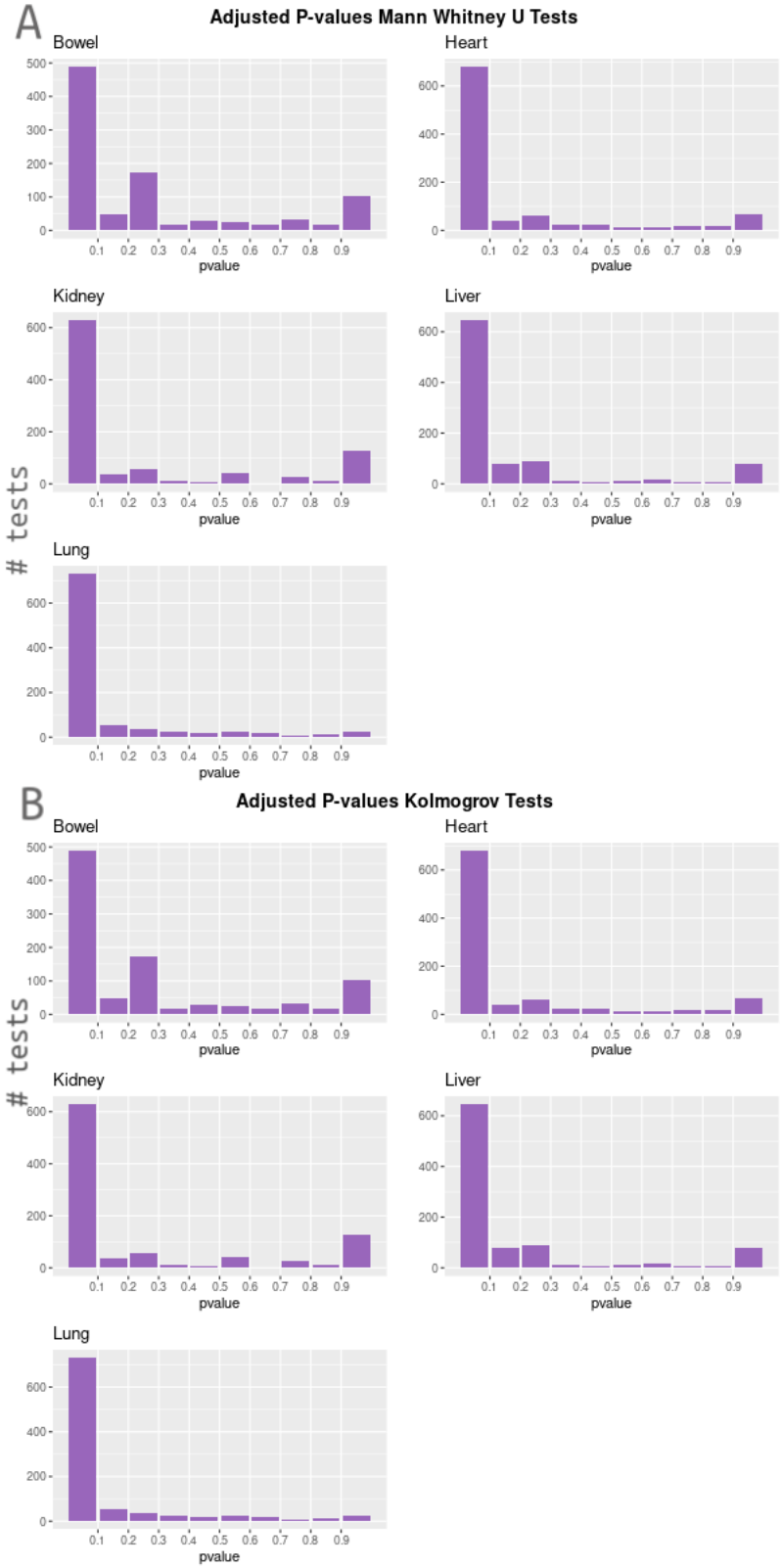


Figura 3.14: Histogramas de P-valores corregidos por BH. A) MWU B) KS

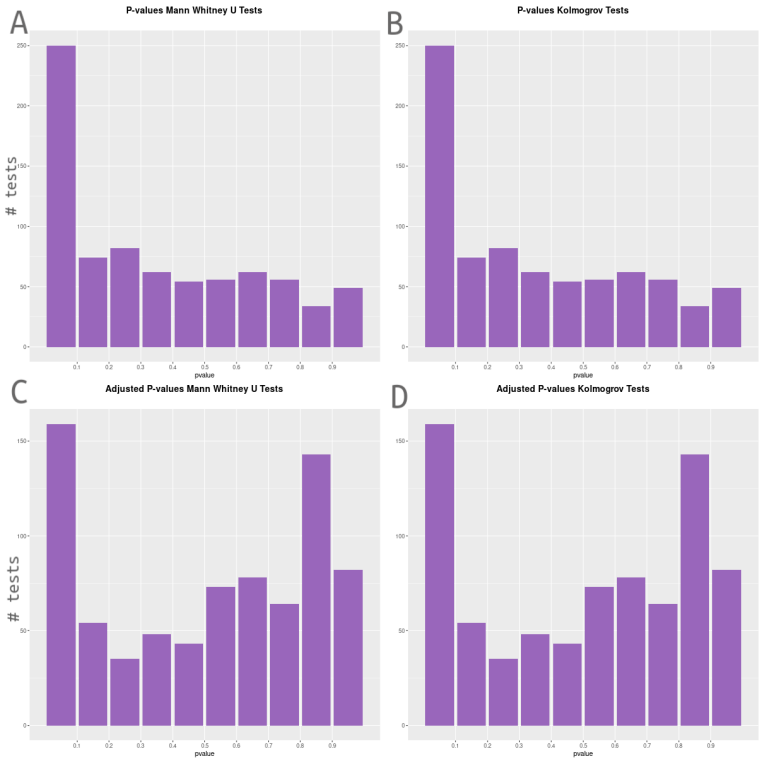


Figura 3.15: Histogramas de P-valores resultantes de pruebas de Activación Diferencial de regulones. Pruebas A) MWU B) KS y P-valores corregidos por BH de pruebas C) MWU y D) KS.

Jensen-Shannon, que a su vez se utiliza para cuantificar al diferencia entre dos distribuciones de probabilidad, y lo hace al evaluar cada regulon comparando cierta condición dada contra todas las demás.

Por análisis, se calculó el RSS para que se compararan todas las condiciones entre sí. Es decir, en el análisis de líneas celulares se calculó el RSS comparando las métricas AUC entre todos los experimentos (Tabla 3.1); en el análisis de tejidos, se compararon todos tejidos con COVID-19 o sanos; y en el análisis de pulmón, se comparó pulmón COVID-19 contra pulmón sano. La función en python para calcular el RSS está disponible con la paquetería de pyscenic y se añadió a los scripts mencionados anteriormente.

Para decidir un umbral sobre los regulones que tomar en base al RSS, se computó el logaritmo base 2 de *Fold Change (LFC)* entre muestras infectadas en comparación con sus respectivos controles y se tomaron aquellos regulones que tenían un LFC >= 0.001.

3.6.4. Resumen de filtrado de regulones por análisis

Algunos regulones seleccionados por los tres filtros mencionados anteriormente se pueden visualizar en los correspondientes *volcano plots* en la sección 4.3. Estas gráficas se realizaron en R con el paquete *ggplot2*, cuyo código está disponible en la función *volcanos_per_exp* del script $regs - analysis.R$ subido a github.

En la Tabla 3.10 se muestra un resumen del filtrado de regulones por análisis. La cantidad de regulones por condición en líneas celulares infectadas por SARS-CoV-2 o tejido COVID-19 tras aplicar estos filtros se muestran en las Tablas 3.11 o 3.12, respectivamente.

Filtro	Experimentos líneas celulares	Tejidos	Pulmón
Regulones encontrados por SCENIC	1309	1367	1279
Filtro de regulones por iteraciones de SCENIC	898	947	779
Activación Diferencial: MWU	638; 542	566	67
Activación Diferencial: KS	638; 542	566	67
Activación Diferencial: MWU + KS	611; 534	554	56
Especificidad (RSS) LFC >= 0.001	326; 234	401	55
Sobre-regulados (logFC(AUC) >= 0)	288; 176	356	4

Tabla 3.10: Resumen filtrado de regulones por análisis. Para líneas celulares se indica la cantidad de regulones unicos tras las pruebas en todos los experimentos y después solo aquellos de experimentos infectados por SARS-CoV-2.

Experimento en línea celular	Núm. regulones después de Filtro "Especificidad"	Núm. regulones después de Filtro "Sobre-regulados'
A549 infectado con SARS-CoV-2	9	6
A549 infectado con SARS-CoV-2 expresando hACE2	17	14
A549 infectado con SARS-CoV-2 expresando hACE2 con tratamiento Ruxolitinib	13	13
Calu-3 infectado con SARS-CoV-2	102	149
NHBE infectado con SARS-CoV-2	7	4

Tabla 3.11: Resumen filtrado de regulones por condición en análisis de líneas celulares.

Tejido COVID-19	Núm. regulones después de Filtro "Especificidad"	Núm. regulones después de Filtro "Sobre-regulados'
Pulmón	380	336
Corazón	199	184
Hígado	219	208
Riñón	82	82
Intestino	206	193

Tabla 3.12: Resumen filtrado de regulones por condición en análisis de Tejidos COVID-19.

El umbral de FDR aplicado por cada experimento en líneas celulares infectadas por

SARS-CoV-2 o tejido COVID-19 se indica en las Tablas 3.13 o 3.14, respectivamente. Se tomó un filtro alto en la línea celular de Calu-3 porque con un FDR más bajo no se encontraban regulones por lo que existe el riesgo de un mayor número de falsos positivos, sin embargo, estos regulones se pueden descartar en análisis posteriores. Para el análisis de Pulmón el umbral de FDR fue <= 0.01.

Experimento	FDR	Núm. regulones
A549 infectado con SARS-CoV-2	0.1	243
A549 infectado con SARS-CoV-2 expresando hACE2	0.14	31
A549 infectado con SARS-CoV-2 expresando hACE2 con tratamiento Ruxolitinib	0.14	184
Calu-3 infectado con SARS-CoV-2	0.24	378
NHBE infectado con SARS-CoV-2	0.11	141

Tabla 3.13: Umbrales superiores de FDR aplicados para las pruebas de DA en experimentos de líneas celulares.

Tejido COVID-19	FDR	Núm. regulones
Pulmón	0.0005	514
Corazón	0.0015	339
Hígado	0.0045	366
Riñón	0.0245	93
Intestino	0.00295	331

Tabla 3.14: Umbrales superiores de FDR aplicados para las pruebas de DA en tejidos COVID-19.

Reportes semi auto-reproducibles. Los resultados de esta sección también se pueden consultar en los reportes semi- auto-reproducibles de RMarkdown para cada uno de los tres análisis (líneas celulares, tejidos y pulmón). A lo que se hace referencia aquí con auto-reproducibles es que el esqueleto de RMarkdown es prácticamente idéntico para los tres reportes, donde lo que cambia esencialmente son los parámetros declarados en el encabezado del reporte. Estos parámetros son para declarar los archivos con los

que se produce el reporte, de manera que al cambiar los parámetros de los archivos cambia el reporte solo con respecto a los datos iniciales, lo cual es muy conveniente si se quisiera repetir el análisis o hacer el mismo análisis con datos nuevos. El "semi-" se añadió porque hay algunos cambios en los reportes aparte de aquellos en los parámetros que son específicos de los datos. Estos reportes se encuentran en la carpeta 04-regulons_selection del repositorio en github del proyecto.

3.7. Exploración de Regulones

Después de seleccionar los regulones considerados como *relevantes* con base en a los filtros y pruebas mencionadas en la sección anterior, se realizó un análisis exploratorio de los regulones encontrados por medio de distintos tipos de visualizaciones (generadas en R con paquetes de bioconductor(Reimers and Carey (53))) y pruebas mencionados a continuación. Para cada análisis (líneas celulares, tejidos y pulmón) se hizo esta exploración.

3.7.1. Comparación de Regulones distintamente compartidos entre condiciones

Upset Plot: Un *upset plot* es un gráfico con la misma finalidad que el diagrama de Venn, utilizado cuando se tienen más de dos conjuntos y las intersecciones son más difíciles de visualizar. Básicamente, enlista los conjuntos distintos e indica, al unir con puntos y rayas dos o más conjuntos, la cantidad de elementos en común con una gráfica de barras en la parte superior.

Se visualizó la cantidad de regulones compartidos entre distintos experimentos tratados e infectados o tejidos con COVID-19 por medio de un gráfico tipo *upset* generado con el paquete de $ComplexHeatmap$ (Gu (24)) utilizando la opción *distinct*, indicando que solo se mostraran los regulones compartidos entre n número de experimentos que no se compartieran entre otras combinaciones de n experimentos, es decir, los "regulones distintamente compartidos".

Tabla complementaria de regulones distintamente compartidos: De manera complementaria al punto anterior, se generó el código necesario en R para saber cuáles eran los regulones que estaban "distintamente compartidos" entre cualquier par de experimentos o tejidos infectados. Este código genera una tabla tamaño $m\ x\ m$, donde m es el número de experimentos o tejidos infectados y muestra tanto (1) los regulones que solo están presentes en una condición dada (los regulones "distintamente compartidos" por una sola condición o bien, los de un solo punto en el gráfico de *upset*), como (2) los regulones que están distintamente compartidos entre cualesquiera dos experimentos distintos (aquellos que son uniones de dos en el gráfico de *upset*). A la función en R generada para realizar esta tarea se le llamó $distinctlySharedRegs$.

3.7.2. Heatmap de métricas AUC y RSS

Los gráficos llamados *heatmap* o mapas de calor, básicamente son tablas de c columnas x f filas que muestran a través de una escala de colores las diferencias en las cantidades de una métrica dada entre un grupo de f condiciones comparadas, de manera que visualmente resulte más sencillo comparar la métrica en cuestión. Una práctica común es hacer *clustering* o agrupamiento de las columnas y filas para detectar grupos.

Se generaron *heatmaps* con el paquete de $ComplexHeatmap$ tanto de las métricas AUC (mostrando la actividad del regulon; véase la sección 2.4.3) y RSS (*score* de especificidad del regulon) para visualizar los distintos perfiles de actividad o de presencia de los regulones *relevantes* (es decir, los regulones diferencialmente activados (DA) y DA + específicos por condición) de las condiciones de interés, es decir, aquellos relevantes en la infección por SARS-CoV-2.

Anotaciones de los Factores de Transcripción Adicionalmente, se tuvieron dos preguntas con respecto a los TFs guía de los regulones (recordemos que un regulon está definido como un TF guía y sus respectivos genes diana).

Primero, se tuvo la pregunta de si se vería un patrón de agrupamiento similar entre los regulones DA (o DA y específicos) por condición y los Dominios de Unión al ADN (DBD) de los TFs guía de esos regulones por condición. Dicho de otra manera, si los regulones relevantes de cada condición tenían un DBD distinto.

Para contestar esta pregunta aprovechando la visualización tipo *heatmap*, se anotaron los DBDs de los TFs guía de cada regulón y se agruparon los regulones por DBD. Se añadió una barra indicadora del DBD en la parte inferior del gráfico.

La anotación se tomó de la base de datos de cisbp, mencionado anteriormente, y se codificó la función en R $family\_annot$ para facilitar la consulta de los DBDs de TFs.

La segunda pregunta respecto a los TFs guía fue simplemente cuáles de los regulones tenían un TF guía con rol "inmunológico". Para ello se consultó la base de datos GeneCards (Safran et al. (54)) y se clasificó manualmente a cada TF.

3.7.3. Enriquecimiento de términos biológicos

Se realizaron análisis de enriquecimiento de términos biológicos provenientes de distintas bases de datos: Por Ontología de Genes (GO, por sus siglas en inglés) de Procesos Biológicos (BP, por sus siglas en inglés) (Consortium (12)), de vías KEGG (vías de señalización celulares tomadas de la base de datos de KEGG (Kanehisa et al. (33))) y en algunos casos de enfermedades utilizando la base de datos del paquete $DOSE$ de bioconductor (Yu et al. (72)), de los regulones (*i.e.* los TFs con sus genes regulados) relevantes por condición (infecciones o tejidos COVID-19) por análisis. Esto, utilizando el paquete $clusterProfiler$ (Wu et al. (68)) de bioconductor.

Dependiendo de cómo se realizó este análisis, se generaron distintas visualizaciones que se mencionan a continuación realizados con el paquete de *enrichplot* (Wu et al. (68)).

Dotplot En general, un *dotplot* o gráfico de punto en español, es un gráfico que en el campo de la bioinformática se utiliza para visualizar dos métricas simultáneamente; donde una métrica se puede codificar en la escala de colores (como en el caso del *heatmap*) y la otra métrica en el tamaño del punto. Particularmente en el caso de enriquecimiento de términos biológicos utilizando la función *dotplots* del paquete *enrichplot* (Wu et al. (68)), se grafican los términos biológicos enriquecidos con mayor significancia estadística, donde la escala de color del punto denota la significancia estadística de la prueba de enriquecimiento y el tamaño del punto la cuenta de genes que son relevantes para el término biológico dado.

Cuando la prueba de enriquecimiento se realizó sobre todos los regulones (TFs y genes diana) de cada condición se generaron *dotplots* para visualizar los resultados.

Cloudplot Un *cloudplot* o gráfico de nube en español es un gráfico que muestra la sobre-representación de palabras de una lista de palabras dada, donde típicamente el incremento del tamaño de la fuente denota que esa palabra o frase apareció más veces.

Cuando la prueba de enriquecimiento se realizó sobre cada regulon de cada condición, se generó un *cloudplot* para visualizar los resultados.

3.7.4. Reportes auto-reproducibles por análisis

Los resultados de esta sección también se pueden consultar en los reportes semi auto-reproducibles de RMarkdown para cada uno de los tres análisis. Estos reportes se encuentran en la carpeta 05-regulons_exploration del repositorio en github del proyecto.

3.8. Contribuciones a RSAT

En paralelo al trabajo mencionado hasta ahora, se colaboró con los desarrolladores de la plataforma de RSAT (véase sección 2.3.3.2) en dos proyectos mencionados a continuación.

3.8.1. network-interactions

Se desarrolló la *pipeline* bioinformática llamada $network - interactions$, cuya función principal es la construcción de una red de regulación a partir de una lista de factores transcripcionales de interés y un conjunto de regiones genómicas correspondientes a las regiones regulatorias de genes específicos. En esencia, se buscan los motivos de los TFs en las regiones regulatorias de los genes para establecer conexiones regulatorias directas TF-gen diana. Este programa puede funcionar también como paso de corrección de redes regulatorias hechas *a priori*, donde se remueven aquellas conexiones que no estén sustentadas por el paso de *pattern-matching*. De manera importante, el programa actualmente funciona para organismos cuyos genomas también estén disponibles en el UCSC Genome Browser, es decir, alrededor 104 organismos hasta Febrero del 2023.

network − interactions puede utilizarse desde línea de comandos o desde la página web de RSAT para organismos del reino metazoa (servidor metazoa).

Este programa se desarrolló principalmente en el lenguaje de perl dentro de la infraestructura de programación de RSAT. Para el desarrollo de la página web se optimizó y desarrolló código a partir del código del programa *matrix − clustering* de RSAT; esto ocupando html, jquery y php. Los códigos desarrollados se pueden consultar con la documentación debida desde github.

- Código principal del programa *network − interactions* : network-interactions.
- Código de página web del programa: network-interactions_form.cgi.
- Código de página web de resultados del programa: network-interactions.cgi.

También se colaboró en la escritura de la sección correspondiente a este trabajo en el artículo.

3.8.2. REST-API

Una REST-API es una aplicación de servicio de interfaz para algún programa, particularmente programas web donde se facilita el proceso de consulta y respuesta del mismo (red (1)).

Anteriormente se había desarrollado el REST-API para el RSAT servidor fungi que se puede consultar aquí y recientemente, se desarrolló la versión en python para hacer consultas a RSAT desde *scripts* de python.

Como contribución directa y para garantizar el buen funcionamiento de estas nuevas tecnologías fue necesario probar cada uno de los 48 programas de RSAT que estaban accesibles en la REST-API en web como a través de los *scripts* de python. Para probar el funcionamiento adecuado de estos programas se utilizaron los DEMOs (accesibles en la página web de cada programa) dado que representan análisis bioinformáticos estandarizados. Posteriormente, se reportó el estado de cada uno y en dado caso, el tipo de error que generaban: conflictos de permisos en el servidor, error en código original del programa, o error en la API. Se realizó el seguimiento de errores o depuración de código y consiguientemente, se realizó un reporte y se consultó con los desarrolladores originales las posibles soluciones de los errores. En el caso de los *scripts* de python, se solucionaron directamente los problemas de código como puede verse aquí.

Adicionalmente, se realizó un manual de uso para el programa de *matrix−scan*, uno de los más citados en esta plataforma por la comunidad científica como introductorio para usar estos servicios de interfaz. Trabajo que puede ser consultado aquí.

Para finalizar, en la sección de resultados 4.5 se muestra el artículo publicado en la revista Nucleic Acids Research conteniendo el trabajo mencionado en esta sección.

Capítulo 4

Resultados

4.1. Datos recuperados

Un resumen de los datos públicos recuperados para este estudio se muestra en la Figura 4.1. Los datos analizados y procesados (mostrados la Figura 3.1) pueden utilizarse para futuros análisis.

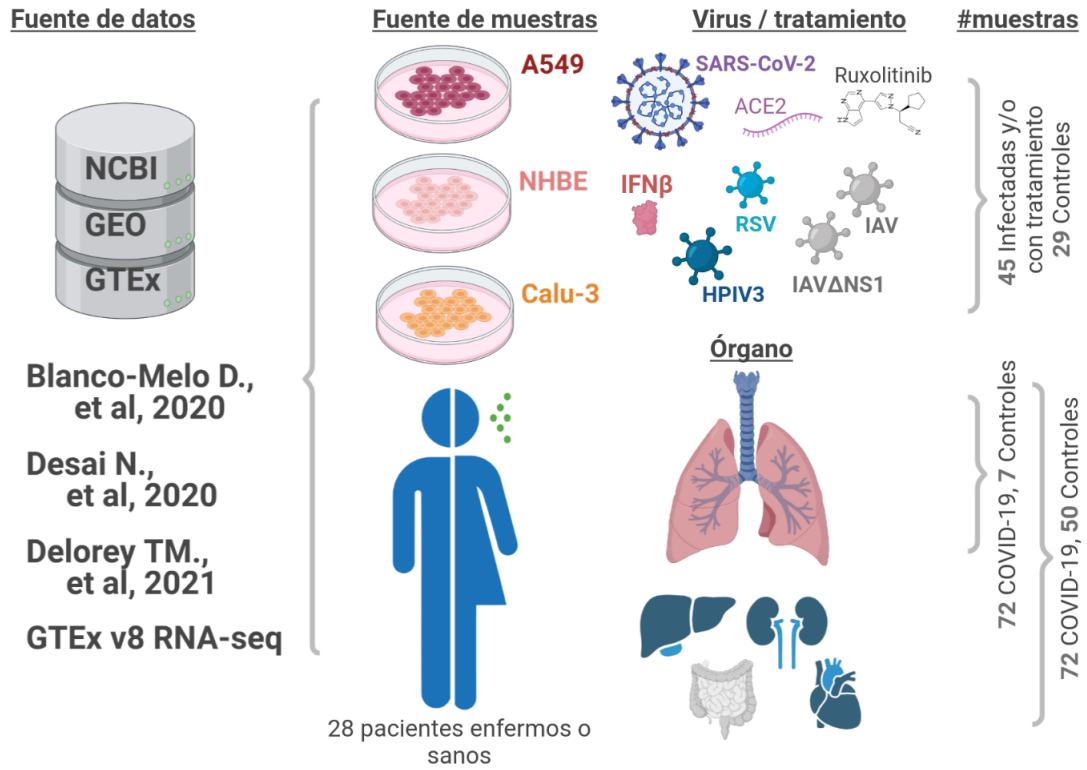


Figura 4.1: Resumen de datos utilizados. En "Data origin", se menciona de qué bases de datos se descargaron los datos y de qué estudios en particular provinieron. En "Sample origin" e "Infection virus or organ", se menciona el tipo de muestra: se recuperaron datos de líneas celulares infectados con distintos virus (mostrado en la parte superior) y datos de biopsias de pacientes con COVID-19 o sanos (mostrados en la parte inferior, los tejidos corresponden a pulmón, hígado, riñón, intestino y corazón). En la última columna se anota el número de muestras por análisis (véase sección 3.1.4), *i.e.*, para líneas celulares: 45 controles y 29 infectados o tratados, para pulmón: 72 biopsias de COVID-19 y 7 controles, y para el análisis de todos los tejidos de biopsias: 72 de COVID-19 y 50 controles.

4.2. Exploración de datos con PCA

4.2.1. Perfil Transcriptómico de Líneas Celulares

4.2.1.1. Infección de SARS-CoV-2 en comparación a otros virus en líneas celulares

En este análisis, los datos provenían de un sólo estudio (Blanco-Melo et al. (4)), por lo que, como se mencionó antes, sólo fue necesaria una correción de lotes por rondas de secuenciación.

Aquí se aborda el tercer objetivo de la exploración de datos (véase sección 3.4): Hacer una descripción general del perfil transcriptómico de las muestras.

En general, en la Figura 4.2 panel A podemos ver que:

- Todas las muestras de un mismo experimento se agrupan entre sí, que es lo esperado dado que las réplicas biológicas deben parecerse entre sí ya que provienen de condiciones iguales. Esto es un buen indicador de que las comparaciones entre experimentos serán certeras.
- En la mayoría de los casos, las infecciones causadas por distintos virus dan lugar a un perfil transcriptómico particular, teniendo su propio cúmulo en la gráfica de PCA.
- Como excepción a lo anterior, las muestras de Calu-3 infectadas con SARS-CoV-2 parecen tener un perfil muy parecido al de A549 infectado por HPIV3.
- La mayoría de las muestras infectadas con SARS-CoV-2 (SC2-inf-A549, SC2-inf-A549+ACE2, SC2-inf-NHBE) se agrupan en el cúmulo superior izquierdo, con las muestras de NHBE tratadas con IFN-β y las muestras control de A549 expresando ACE2 (quizás la expresión de ACE2 influenció de manera similar el perfil que a las infectadas con SARS-CoV-2)
- Como se esperaría, las muestras de NHBE infectadas por IAV y IAVdNS1 (menos patogénico) se encuentran en posiciones opuestas.

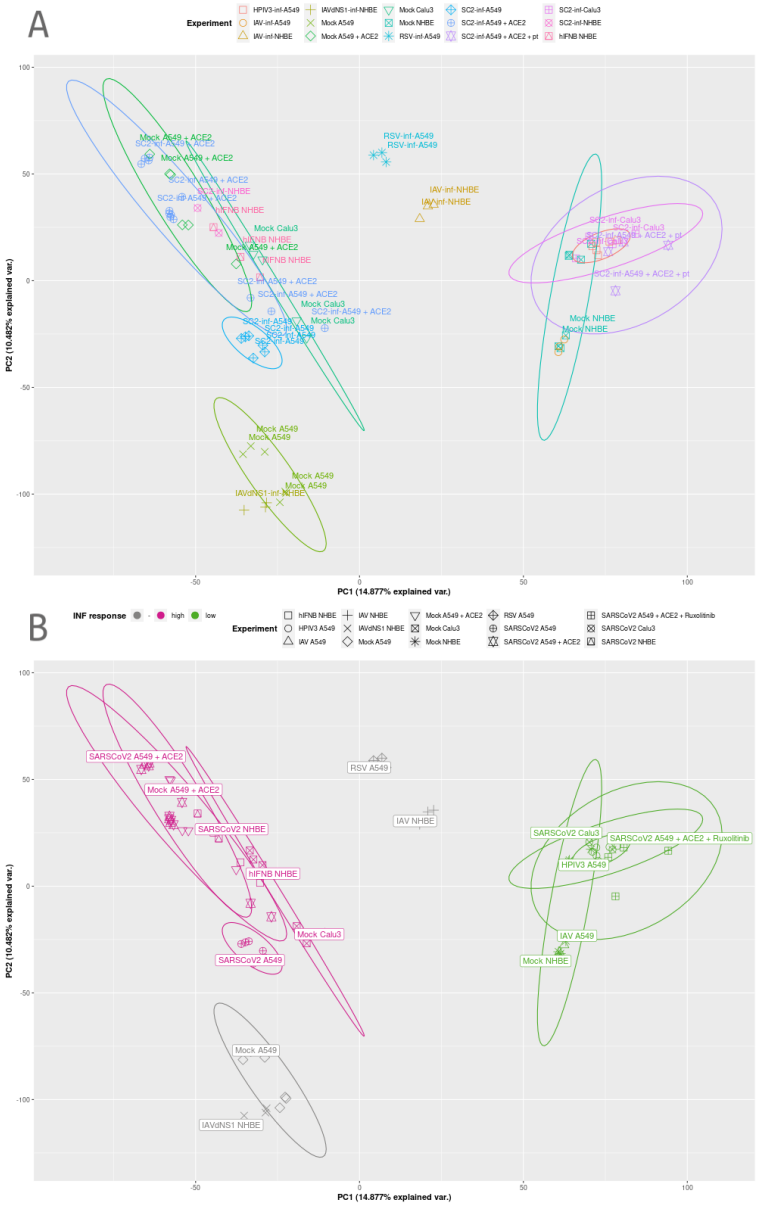


Figura 4.2: PCA de líneas celulares infectadas. Se muestran los componentes principales 1 y 2 explicando el 25.359 %de la variabilidad génica. La descripción de las abreviaciones en las etiquetas puede consultarse en la Tabla 4.1. A) Las muestras de un mismo experimento se identifican por la combinación particular de una figura de punto, un color y su etiqueta. B) Se muestra una etiqueta de experimento por cúmulo de muestras. Se colorean las muestras por supuesta respuesta a interferón (rosa es alta y verde baja).

Abreviación	**Experimento**
Mock Calu-3	Control Calu-3
Mock NHBE	Control NHBE
Mock A549	Control A459
Mock A549+ACE2	Control A549 transducido con vector expresando ACE2 humano
SC2-inf-A549	SARS-CoV-2 infección a A459
SC2-inf-A549+ACE2	SARS-CoV-2 infección a A549 transducido con vector expresando ACE2 humano
SC2-inf-A549+ACE2+pt	SARS-CoV-2 infección a A459 transducido con vector expresando ACE2 humano (pretratamiento Ruxolitinib)
SC2-inf-Calu3	SARS-CoV-2 infección a Calu-3
SC2-inf-NHBE	SARS-CoV-2 infección a NHBE
RSV-inf-A549	RSV infección a A459
IAVdNS1-inf-NHBE	IAVdNS1 infección a NHBE
IAV-inf-NHBE	IAV infección a NHBE
IAV-inf-A549	IAV infección a A459
HPIV3-inf-A549	HPIV3 infección a A459
hIFNB NHBE	NHBE tratado con IFN-β

Tabla 4.1: Descripción de etiquetas en gráficos de PCA de líneas celulares.

La respuesta hospedera a la infección por SARS-CoV-2 es compleja. La dispersión de las muestras en el PCA 4.2 depende tanto del experimento de procedencia: línea celular, infección y el tratamiento. Lo que se puede observar de manera más puntualizada en el panel B, en código de colores, es que las muestras infectadas por SARS-CoV-2 están repartidas en dos cúmulos: el verde y el rosa. La coloración de estos cúmulos está dada por 3 experimentos. Las muestras de NHBE tratadas con IFN-β se encuentran en el cúmulo rosa (lado izquierdo) y sus controles en el cúmulo verde (lado derecho), que deben tener una menor respuesta interferón; las muestras con tratamiento de Ruxolitinib, que inhibe a IFN, también se encuentran en el cúmulo verde. Puede ser que, aunque compleja y llevada por la expresión de más de 20 mil genes, los experimentos SARSCoV2 A545, SARSCoV2 A545 + ACE2 y SARSCoV2 NHBE podrían tener una mayor respuesta a interferón y se parecen más entre sí. Tiene sentido que el experimento SARSCoV2 Calu3 se encuentre lejano pues al ser una línea celular cancerosa, a diferencia de las otras, esta debe tener un perfil transcripcional muy distinto relacionado a cáncer.

4.2.1.2. Similitud entre células epiteliales de *pseudobulk* y muestras de línea celular A549

Como se mencionó en la sección 3.4, se tenía como objetivo evaluar la similitud transcripcional de las células epiteliales (AT1 o AT2) infectadas por SARS-CoV-2 de BAL provenientes de la tecnología *single-cell* con las muestras de líneas celulares provenientes de epitelio pulmonar (en particular, la línea celular A549 ha sido usada como modelo de AT2 (Khan et al. (36))). Esto con el fin de ser capaces de comparar nuestros resultados con aquellos vistos en estudios de *single-cell* y saber a qué nivel los resultados de experimentos en líneas celulares se podrían extrapolar a los observados en el tejido de interés.

Precisamente, en la Figura 4.3 panel A se tiene el PCA correspondiente a este análisis, se puede apreciar como el perfil transcripcional del agregado celular epitelial (identificado por la etiqueta "Epithelial", en azul) coincide con el de las líneas celulares (también en azul). Se puede notar cómo las células no epiteliales y las muestras de biopsias pulmonares (en rojo) se localizan, en su gran mayoría, fuera de las distribuciones de perfil transcriptómico asumidas para datos de líneas celulares. En particular, el agregado celular epitelial de *pseudobulk* se encuentra en la distribución de muestras de experimentos de A549 y NHBE control pero cerca además de las muestras de A549 infectadas por SARS-CoV-2 lo cual es congruente con lo esperado. Adicionalmente, se puede ver como el agregado de células inmunológicas se encuentra cerca de las muestras pulmonares de COVID-19.

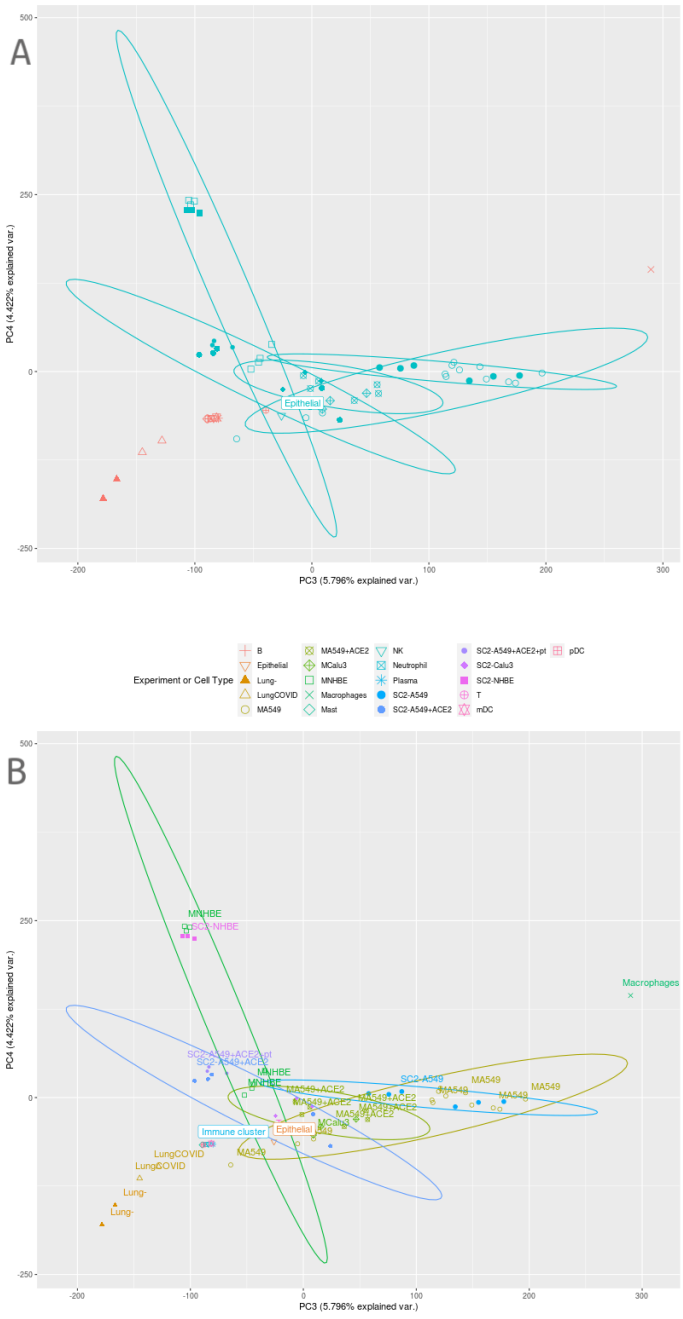


Figura 4.3: PCA de líneas celulares infectadas con SARS-CoV-2 y células inmunes y epiteliales provenientes de pacientes con COVID-19. Componentes principales 3 y 4 explicando el 10.218 % de la variabilidad génica. A) Se colorean las muestras por ser tipo Epitelial (rojo es no, azul es sí). La etiqueta con fondo blanco "Epithelial" corresponde a las células epiteliales de *pseudobulk*. B) Las muestras de un mismo experimento o tipo celular se identifican por una figura de punto, un color y su etiqueta. Se etiquetan con fondo blanco los agregados celulares de *pseudobulk*, *i.e.*, las células inmunológicas como "Immune cluster" y "Epithelial" al epitelial.

4.2.2. Perfil Transcriptómico de COVID-19 en tejidos

En este análisis los datos provenían de dos fuentes (Desai et al. (16) y Aguet et al. (3)) y de manera relevante, estas mismas fuentes corresponden a las condiciones biológicas (COVID-19 y sano, respectivamente), por lo que llevar a cabo el paso de corrección por lotes era particularmente importante. Se evaluó la eficacia en que uno y dos pasos de corrección lograron remover el efecto de lote de las cuentas génicas. El primero tomando en cuenta variaciones técnicas en el mismo *dataset* y el segundo entre ellos. Como se puede ver en la Figura 4.4 panel A, al aplicar la corrección solo una vez, los datos de tejido sano se conglomeran y la variación de la expresión génica entre sus muestras es prácticamente imperceptible a comparación con las muestras de tejido infectado. Después de una segunda corrección (Figura 4.4 panel B), se puede ver la disminución del efecto de lote al notar como el perfil transcripcional de las muestras sanas se desfasan por tejido.

Evidentemente, el componente principal 1 separa los datos tanto por estado COVID-19 o sano, y por el *dataset* del que provienen. Después, los datos se agregan por tejido. Sin embargo, hay una notable heterogeneidad entre las muestras de pulmón con COVID-19.

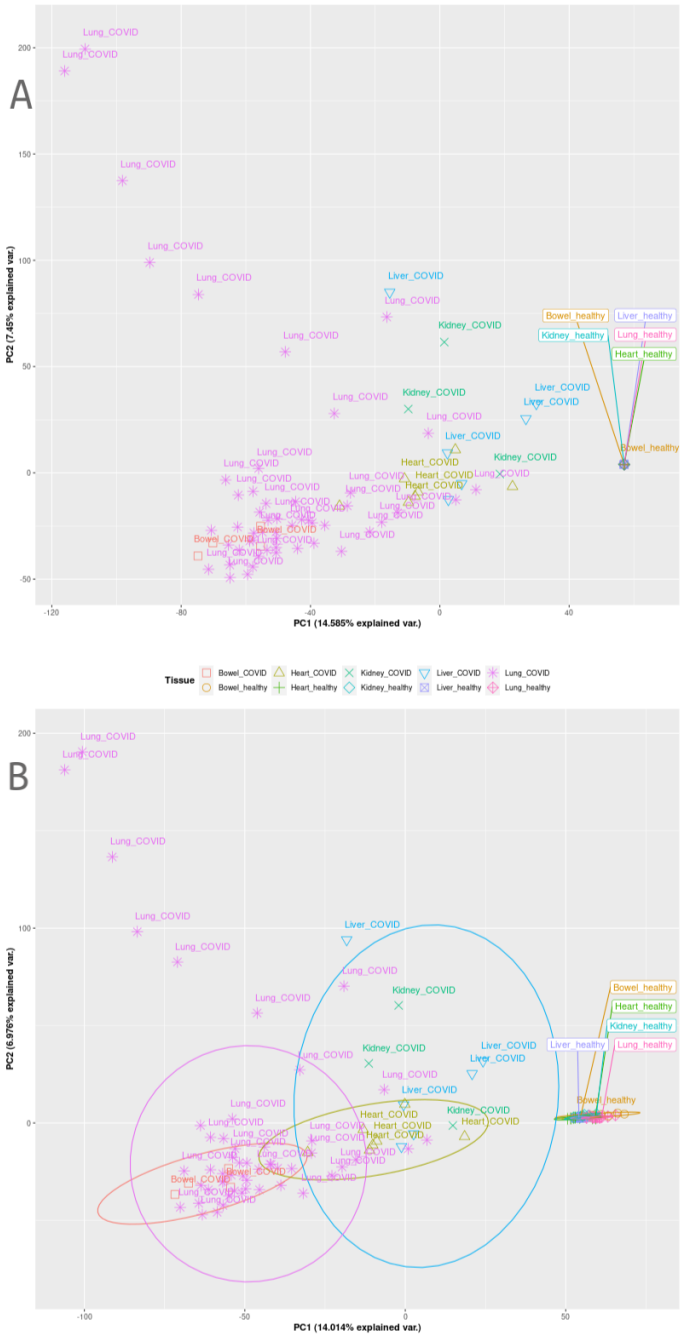


Figura 4.4: PCA de distintos tejidos: pulmón, corazón, hígado, intestino y riñón con COVID-19 o sano. Componentes principales 1 y 2, explicando el 20.99 % de la variabilidad. Las muestras provenientes de un mismo tejido y estado (sano: "healthy" o "COVID") se identifican por una figura de punto, un color y su etiqueta. Corrección por *batches* aplicada A) 1 o B) 2 veces.

De manera importante, al visualizar los datos desde los componentes principales 3 y 4 (Figura 4.5), se puede ver cómo los datos se mezclan entre *datasets*.

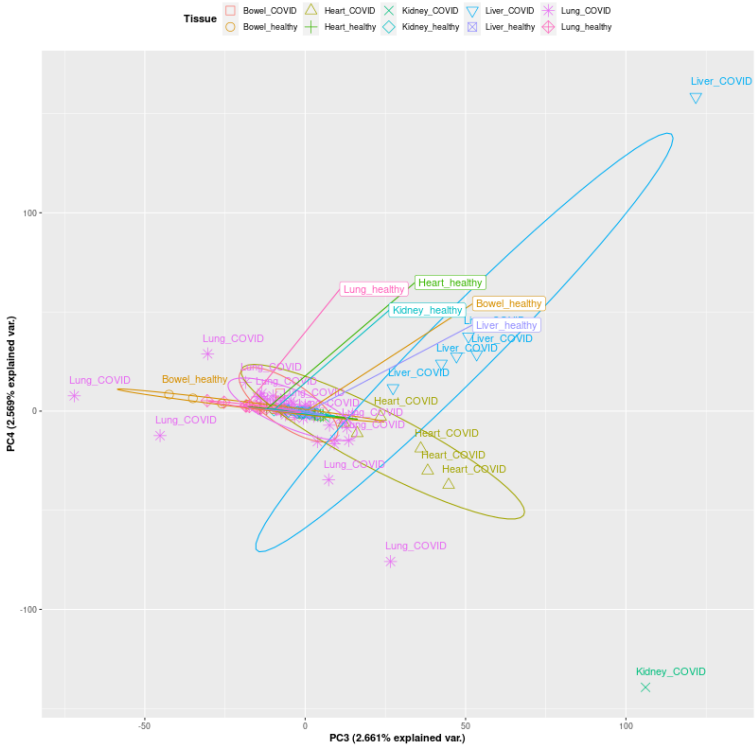


Figura 4.5: PCA de distintos tejidos: pulmón, corazón, hígado, intestino y riñón con COVID-19 o sano. Componentes principales 3 y 4, explicando el 5.23 % de la variabilidad. Las muestras provenientes de un mismo tejido y estado (sano: "healthy" o "COVID") se identifican por una figura de punto, un color y su etiqueta. Corrección por *batches* aplicada 2 veces.

Se concluyó que la mejor opción era proseguir con análisis posteriores a partir de los datos corregidos dos veces.

4.2.3. Perfil Transcriptómico de Pulmón COVID-19

Como se menciona en la sección 3.1.4 en este análisis los datos provienen de tres fuentes distintas. A partir del resultado de la sección anterior, se tomó como base aplicar dos correcciones de *batch*, como fue aplicado aquí.

Como se puede ver en la Figura 4.6 panel A, los datos más difíciles de integrar fueron los de Delorey, lo cual tiene sentido porque además de provenir de una fuente distinta, la *pipeline* de pre-procesamiento de datos aplicada fue distinta al no tener los

datos “crudos” disponibles. Como mejora, en el panel B se puede ver como a pesar de la gran heterogeindad entre las muestras de Pulmón con COVID-19, estas se mezclan entre *datasets*, rescatando así el perfil transcriptómico representando las condiciones biológicas y no la variabilidad técnica detrás. Asimismo, las muestras de biopsias de pulmón sano se acercan entre sí.

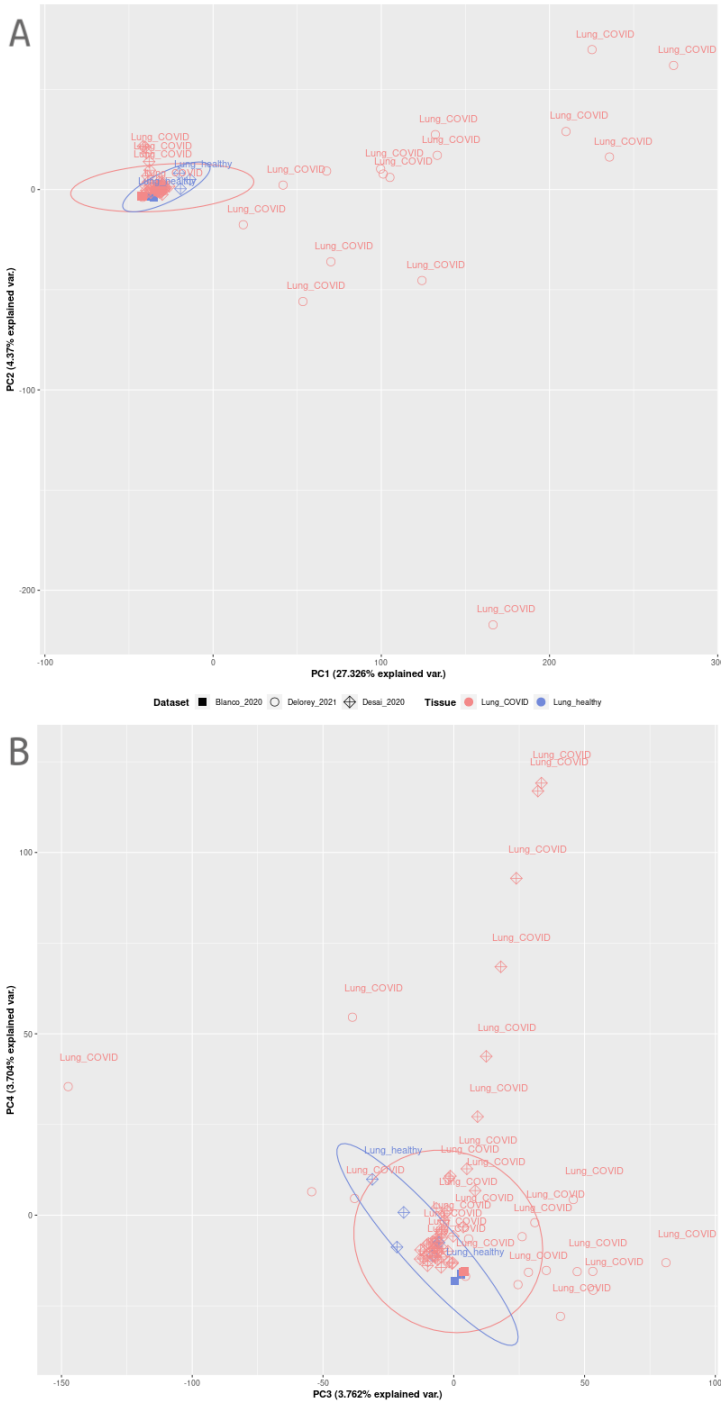


Figura 4.6: PCA de Pulmón COVID.19. Las muestras provenientes de un mismo tejido y estado (sano: "healthy" o "COVID") se identifican por una figura de punto y un color, y su etiqueta, respectivamente. Corrección por *batches* aplicada 2 veces. A) Componentes principales 1 y 2, explicando el 31.696 % de la variabilidad génica. B) Componentes principales 3 y 4, explicando el 7.23 % de la variabilidad.

Combinar *datasets* en un análisis supone el jugar con un balance: por un lado, tener una mayor cantidad de muestras contribuyendo a un resultado, y por lo tanto tener mayor confianza sobre el mismo, y por otro lado, tener una mayor cantidad de efectos de lote que contender, que a pesar del uso de algoritmos de corrección, estos sólo pueden ser disminuidos hasta cierto punto como se puede ver en el resultado anterior.

Como se vio en la Figura 4.6 panel B, es posible rescatar la variabilidad biológica de este conjunto de datos. Razonamos que usarlos tras dos correcciones de lote sería suficiente para proseguir con análisis posteriores al revisar a detalle el funcionamiento del paso siguiente en la *pipeline*: generación de la red de regulación con base en a la co-expresión usando *GRNboost*2 (véase sección 3.5.2). Brevemente, el algoritmo hace una regresión por cada gen diana usando un algoritmo basado en árboles que captura relaciones de regulación no lineales, lo cual hace posible que contenga con efectos de lote entre muestras.

4.3. Regulones Diferencialmente Activados

A continuación se muestran los resultados de la selección de regulones relevantes, es decir, que podrían ser importantes durante la patogénesis del COVID-19. Lo anterior, con base en (1) las pruebas de Activación Diferencial (AD) usando las pruebas no paramétricas de MWU y KS; y (2) la distinción entre los regulones sobre y sub activados en muestras infectadas por SARS-CoV-2 usando el log Fold change (infectado/control).

De manera específica, se mencionan específicamente en una tabla los regulones diferencialmente sobre- o subactivados que tenían la mayor significancia estadística. Es importante recalcar que para análisis posteriores se toman todos los regulones que resultaron en AD mientras que aquí sólo se mencionan los "top". Nótese que el FDR utilizado por condición es distinta como se detalla en la sección 3.6.4.

4.3.1. Análisis: Infección de SARS-CoV-2 en comparación a otros virus en líneas celulares

La Tabla **??** menciona los Top regulones en DA (referidos por el Factor Transcripcional que los guía) que se identificaron como sobre-activados en los experimentos de líneas celulares. Por otra parte, en la Tabla 4.8 menciona los Top regulones en DA que se identificaron como sub-activados.

Experimento	p-value	Núm. regulones	Top regulones
SARSCoV2-inf-A549	0.00236410353	6	BCLAF1,GABPB1,MAX,PSMD12,STAT1,TFAP2C
SARSCoV2-inf-A549 +ACE2	0.08098845599	19	ATF3,CEBPD,EGR3,ETV6,ETV7,FOXK1,HIVEP2,HOXB6,IRF1,IRF5,IRF8,IRF9,MEF2B,NKX3-1,SOX2,STAT1,STAT2,TP53,XBP1
SARSCoV2-inf-A549 +ACE2+ Ruxolitinib	0.1306239432	81	AIRE,ATF3,BCLAF1,BHLHE40,BRF2,CEBPD,CLOCK,DLX3,DMRT1,DMRTA1,E2F7,EGR1,EHF,ELF1,ELF4,ELK3,ELK4,EMX1,ETS1,ETS2,ETV6,ETV7,FOXB1,FOXK1,FOXK2,FOXN1,FOXO1,GTF2B,HEY1,HIVEP2,HOXB6,IKZF2,IRF1,IRF2,IRF5,IRF6,IRF7,IRF8,IRF9,IRX5,KLF5,KLF6,KLF7,MAX,NEUROD2,NFE2L2,NFKB1,NKX2-2,NKX3-1,NR3C2,PBX3,POLR3A,POLR3G,POU2F3,POU3F1,PPARD,RUNX2,RXRG,SATB1,SIX2,SMARCA5,SOX2,SOX9,STAT1,STAT2,TBX3,TCF12,TFAP2A,TFAP2C,TWIST1,XBP1,ZBTB41,ZFX,ZNF274,ZNF460,ZNF502,ZNF583,ZNF701,ZNF730,ZNF773,ZXDC
SARSCoV2-inf-Calu3	0.2375661376	223	ALX4,ARID5B,ARNT2,ARNTL,ARNT,ATF3,ATF4,ATF6B,BACH2,BATF,BCL11A,BCL3,BHLHE40,BHLHE41,BRCA1,BRF2,CEBPD,CEBPG,CIC,CREB5,CREBL2,CUX1,DBP,DDIT3,DLX3,DLX4,DRGX,E2F1,E2F2,E2F8,EGR2,EGR3,EHF,ELF1,ELF4,ELK1,ELK3,ELK4,EMX1,EP300,ESX1,ETS1,ETS2,ETV3,ETV4,ETV6,ETV7,FEV,FOSL1,FOSL2,FOXC2,FOXF2,FOXJ2,FOXK1,FOXL1,FOXM1,FOXN1,FOXO1,FOXO3,FOXO4,GABPB1,GATA2,GATA3,GATA5,GATA6,GBX1,GLI1,GMEB1,GMEB2,GRHL1,GRHL2,GSC2,GTF2B,GTF2IRD1,GTF3A,HAND1,HDAC2,HES1,HINFP,HIVEP1,HIVEP2,HLF,HMGB1,HOXA6,HOXB4,HOXB6,HOXD10,HOXD8,HOXD9,HSF2,IKZF2,ING3,IRF1,IRF2,IRF4,IRF5,IRF6,IRF7,IRF8,IRF9,IRX2,JUN,KDM5A,KLF11,KLF12,KLF13,KLF14,KLF15,KLF16,KLF2,KLF5,KLF6,KLF7,KLF8,KLF9,LHX1,MAFB,MAX,MECOM,MEIS1,MESP1,MIOS,MITF,MLXIP,MXI1,MYEF2,MYOP,NFE2L1,NFE2L2,NFE2,NFIC,NFKB1,NFYA,NFYC,NHLH1,NHLH2,NKX3-1,NPAS2,NR1H3,NRF1,OLIG3,OVOL1,PAX8,PHF21A,PHF2,PHF8,PKNOX1,POLE3,POLR2A,POLR3A,POU2F1,POU4F1,POU5F1B,POU5F1,POU6F1,PPARA,PPARD,PRRX2,RAD21,RARA,RREB1,RUNX2,SATB1,SETDB1,SIX1,SIX2,SMAD1,SMARCA5,SMARCC2,SMC3,SNAPC4,SOX15,SOX2,SOX4,SOX9,SP1,SP2,SP7,SPDEF,SREBF2,SRF,STAT1,STAT2,STAT5B,TAF7,TBP,TBX18,TBX6,TCF7L1,TEAD4,TFAP2A,TFAP2C,TFDP1,TFE3,TFEB,TGIF2,THAP1,TRIM28,USF1,VAX2,VDR,WT1,YBX1,YY2,ZBTB33,ZEB1,ZFY,ZNF134,ZNF143,ZNF208,ZNF267,ZNF316,ZNF384,ZNF429,ZNF471,ZNF550,ZNF554,ZNF560,ZNF607,ZNF672,ZSCAN31,ZXDC
SARSCoV2-inf-NHBE	0.1061465721	99	AHCTF1,AIRE,ALX1,ATF1,BCL3,CEBPB,CEBPG,CREB1,CREBL2,DLX4,DMRTA1,DUXA,EBF1,EGR2,ERF,ETV1,ETV6,ETV7,FOXA2,FOXF2,FOXK2,FOXP3,FOXQ1,GATA2,GATA3,HDAC2,HLF,HNF1A,HNF1B,HOXA2,HOXA5,HOXA7,HOXA9,HOXB5,HOXB6,HOXB7,HOXB8,HOXB9,HOXC10,HOXC4,HOXC5,HOXC6,HOXC8,ING3,IRF1,IRF4,IRF5,IRF8,IRF9,ISL2,LHX1,LHX4,LHX8,LHX9,MEOX2,MIOS,MSX1,MSX2,MYEF2,NFIL3,NFYB,NHLH1,NKX3-1,NR1I2,NR4A1,NRF1,ONECUT2,PAX2,PAX7,PBX1,PDX1,POU2F3,RBPJ,RFX5,SMC3,SP1,SP2,SP4,SP7,STAT1,STAT2,TAF7,TAL1,TBPL1,TBPL2,TBP,TBX3,TET1,TFE3,TFEB,UBTF,UQCRB,VSX1,WT1,ZEB1,ZFHX2,ZFY,ZNF502,ZNF580

Figura 4.7: Top regulones diferencialmente sobre-activos en líneas celulares tratadas o infectadas con distintos virus que afectan el sistema respiratorio.

Experimento	p-value	Núm. regulones	Top regulones
SARSCoV2-inf-A549	0.00236410353	19	ERF,HES6,HNF1A,HOXD3,MAZ,MNX1,MXD3,MZF1,NFYC,NR2F1,SNAI3,SP5,TCF3,TRIM28,USF1,USF2,ZNF486,ZNF52 4,ZNF787
SARSCoV2-inf-A549 +ACE2	0.0809884559	2	DLX2,SOX21
SARSCoV2-inf-A549 +ACE2+ Ruxolitinib	0.1306239432	74	AKR1A1,ALX4,BHLHA15,CDX2,CEBPA,DLX2,DLX4,E2F1,E2F2,E2F8,EBF1,ENO1,ERF,FOXD3,FOXM1,FOXP1,GATA5,GLI1,GRHPR,HES6,HLX,HNF4G,HOXA6,HOXC11,HOXC5,HOXC9,HOXD4,HSF1,LHX1,LHX2,LMX1B,MAZ,MLXIP,MNX1,MXD3,MYEF2,MZF1,NFATC2,NPDC1,NR1D1,NR1H3,NR1H4,NR2F1,NR2F6,ONECUT2,PDX1,POLE4,POU3F4,PRRX1,PRRX2,RFX1,RFXAP,SETDB1,SIX5,SMARCC2,SNAI3,SOX21,SP6,SREBF1,STAT5B,TAF1,TEAD2,TET1,TGIF1,TRIM28,USF1,VAX1,ZBED1,ZFHX3,ZNF276,ZNF524,ZNF579,ZNF672,ZNF787
SARSCoV2-inf-Calu3	0.2375661376	125	AKR1A1,ARNTL2,ARX,AR,ASCL1,ASCL2,ATF6,BBX,BDP1,BNC2,BORCS8-MEF2B,CCNT2,CDX2,CPEB1,DLX2,DMRT1,DUX4,DUXA,E4F1,ENO1,EOMES,ESRRA,EVX2,FBXL19,FOXA2,FOXA3,FOXC1,FOXD2,FOXD3,FOXF1,FOXG1,FOX3,GLIS2,GRHPR,GSC,HES5,HES6,HESX1,HEY2,HLX,HNF1A,HOXA2,HOXA9,HOXB7,HOXB8,HOXB9,HOXC10,HOXC11,HOXC13,HOXC4,HOXC6,HOXC8,HOXD4,HSF1,ISL2,KDM4D,LBX1,LHX4,LHX5,LMX1B,MAZ,MEF2C,MEOX2,MYOD1,MZF1,NEUROD1,NFATC2,NFE2L3,NKX6-2,NPDC1,NR0B1,NR1H2,NR1H4,NR1I2,NR2E1,NR2F1,NR2F6,NR4A1,NR5A2,OLIG2,ONECUT2,OTP,PAX7,PAX9,PBX1,PDX1,PHOX2A,PRDM1,PRRX1,RARG,RBBP5,RFX7,SIN3A,SIX5,SNAI3,SOX13,SOX21,SP6,SP9,STAT4,STAU2,TAL1,TCF3,TCF4,TET1,TLX1,TLX2,TP73,UBP1,VAX1,VEZF1,VSX1,VSX2,ZFHX2,ZNF219,ZNF264,ZNF276,ZNF441,ZNF486,ZNF513,ZNF524,ZNF606,ZNF721,ZNF768,ZNF787
SARSCoV2-inf-NHBE	0.1061465721	26	BATF,BBX,CKMT1B,DMRT1,ESRRA,HES5,HSF2,IRF3,IRX2,KLF5,MLXIP,MXD4,NKX2-1,OVOL2,PKNOX1,RARG,SIX2,SREBF2,TAGLN2,TFAP2A,TLX3,TWIST1,ZNF471,ZNF560,ZNF773,ZXDC

Figura 4.8: Top regulones diferencialmente sub-activos en líneas celulares tratadas o infectadas con distintos virus que afectan el sistema respiratorio.

4.3.2. Análisis: COVID-19 en distintos tejidos

La Tabla 4.9 menciona los Top regulones en DA (referidos por el Factor Transcripcional que los guía) que se identificaron como sobre-activados en los tejidos de pacientes COVID-19. Por otra parte, en la Tabla 4.10 menciona los Top regulones en DA que se identificaron como sub-activados.

Tejido COVID-19	p-value	Núm. regulones	Top regulones
Intestino	0.024246638651	190	ARID5A,ARNT2,ARX,AR,ATF3,ATF7,BACH1,BACH2,BATF3,BATF,BCL3,BCLAF1,BDP1,BHLHE40,BHLHE41,BRF2,CEBPB,CEBPD,CEBPZ,CHD2,CLOCK,CREB1,CREB3L1,CREB3L2,CREB3,CREBL2,CSNK2B,E2F4,EGR1,EGR3,EHF,ELF3,ELF4,ELK3,ELK4,ENO1,EP300,ESR1,ETS1,ETS2,ETV3,ETV4,ETV6,EZR,FOXA1,FOXA3,FOXC1,FOXF2,FOXI2,FOXK1,FOXN1,FOXN2,FOXO1,FOXP2,FOXQ1,GLIS2,GRHPR,HDAC2,HES1,HEY1,HINFP,HIVEP2,HIVEP3,HLX,HOXB2,HOXB4,HOXC9,ID1,IRF1,IRF2,IRF4,IRF8,IRF9,IRX3,JDP2,JUNB,JUND,JUN,KDM5A,KLF12,KLF17,KLF2,KLF3,KLF4,KLF5,KLF6,KLF7,KLF8,KLF9,MAFG,MAFK,MAZ,MECP2,MEIS2,MIOS,MSC,MSX1,MTHFD1,MXD4,MYBL1,MYBL2,NANOG,NANOS1,NEUROD1,NFATC2,NFIL3,NFKB1,NFYA,NHLH1,NKX1-2,NKX2-4,NKX3-1,NR1I2,NR2F2,NR2F6,NR3C1,NR3C2,ONECUT1,OSR1,PAX9,PKNOX1,POLD2,POLR2A,POLR3A,POU3F2,RAD21,RARA,RBBP5,RBPJ,RELA,RELB,RFXAP,RREB1,RUNX3,SIN3A,SMAD6,SMARCC1,SNAI1,SOX4,SOX7,SOX9,SP1,SP2,SPI1,SREBF2,SREBF1,SRF,STAT1,STAT2,STAT3,TAF7,TBX21,TBX4,TCF21,TEAD2,TEF,TFDP1,TFE3,THRA,THRB,TMEM33,UBTF,VSX2,YBX1,YY1,YY2,ZBTB33,ZBTB7A,ZBTB7B,ZFHX3,ZFX,ZIC5,ZNF148,ZNF318,ZNF358,ZNF362,ZNF426,ZNF428,ZNF460,ZNF467,ZNF524,ZNF579,ZNF671,ZNF726,ZNF782,ZNF785,ZNF787,ZNF878,ZNF91,ZXDA
Corazón	0.001067047553	36	BDP1,ENO1,FOXF1,FOXI2,FOXN1,GLIS2,GTF2A1,HMX3,HOXB2,IRX3,JAZF1,MECP2,MIOS,MTHFD1,MXD4,MYBL2,NFATC2,NKX2-4,OSR1,PAX9,RFXAP,RUNX3,SMARCC1,SNAI1,SOX11,STAT5B,TCF21,TRPS1,YBX1,YY2,ZFHX3,ZNF382,ZNF569,ZNF785,ZNF787,ZNF813
Riñón	0.028818002626	78	BCL11B,BDP1,BHLHE22,CSNK2B,ENO1,EZR,FOXC2,FOXF1,FOXI2,FOXJ2,FOXK1,FOXK2,FOXQ1,FUBP1,GLIS2,GTF2A1,GTF3A,HEY1,HIVEP3,HOXB2,HOXB6,ID1,IRX3,JAZF1,KLF8,MECP2,MEIS1,MIOS,MSX1,MTHFD1,MXD1,MXD4,MYBL2,NFAT5,NFATC2,NHLH1,NKX2-4,NR1I2,NR3C2,OSR1,PAX9,PIR,PLAGL2,PLG,POU2AF1,PRRX2,PSMC2,RBAK,RFX7,RFXAP,RUNX2,RUNX3,SIX1,SMAD3,SMAD6,SMARCC1,SMC3,SNAI1,SOX11,SOX7,STAT5B,TBX4,TCF21,TMEM33,TRPS1,UBP1,VEZF1,YBX1,YY2,ZIC5,ZNF134,ZNF426,ZNF429,ZNF569,ZNF625,ZNF782,ZNF785,ZNF787
Hígado	0.003588663719	198	ARID5A,ARX,AR,ATF1,ATF3,ATF5,ATF7,BACH1,BACH2,BARX2,BATF3,BATF,BCL3,BCLAF1,BDP1,BHLHE40,BHLHE41,BRF2,CEBPA,CEBPB,CEBPD,CEBPZ,CHD2,CREB1,CREB3L1,CREB3L2,CREB3,CREBL2,CSNK2B,E2F2,E2F4,EBF2,EGR1,EGR3,EHF,ELF1,ELF3,ELF4,ELK3,ELK4,EMX2,ENO1,EOMES,EP300,ESR1,ETS1,ETS2,ETV3,ETV4,ETV6,EZR,FOSL1,FOSL2,FOXF2,FOXI2,FOXK1,FOXN1,FOXO1,FOXP2,FOXQ1,GLIS2,GMEB1,GRHPR,GTF3A,HDAC2,HES1,HEY1,HINFP,HIVEP2,HIVEP3,HLX,HOXA7,HOXB2,HOXB4,HSF1,IKZF1,IRF1,IRF2,IRF4,IRF8,IRF9,IRX3,JAZF1,JDP2,JUNB,JUND,JUN,KDM4B,KDM5A,KLF12,KLF17,KLF1,KLF3,KLF4,KLF5,KLF6,KLF9,MAFG,MAFK,MAX,MAZ,MECP2,MIOS,MLXIP,MSX1,MXD4,MYBL1,NFE2L1,NFE2L2,NFE2,NFIL3,NFKB1,NFYA,NKX3-1,NR1H2,NR1I2,NR2F2,NR2F6,NR3C1,OSR1,PAT21,PAX9,PIR,PLG,POLR2A,POLR3A,POLR3G,RAD21,RARA,RBBP5,RBPJ,RELA,RELB,RFX1,RFXAP,RREB1,SIN3A,SMAD3,SMAD5,SMARCC1,SMARCC2,SNAI1,SNAI3,SOX11,SOX4,SOX6,SOX9,SP1,SP2,SPI1,SREBF1,SREBF2,SRF,STAT1,STAT2,STAT3,STAT4,TAF7,TEAD2,TEF,TFAP2C,TFAP4,TFDP1,TFE3,THRA,TMEM33,UBTF,YBX1,YY1,YY2,ZBTB33,ZBTB7A,ZBTB7B,ZFHX3,ZFP64,ZFX,ZIC5,ZNF134,ZNF148,ZNF319,ZNF358,ZNF362,ZNF382,ZNF429,ZNF444,ZNF460,ZNF467,ZNF513,ZNF524,ZNF671,ZNF785,ZNF787,ZNF813,ZNF84,ZNF878,ZNF91,ZXDA
Pulmón	3.59E-07	113	ARID3A,ATF1,BATF,BCL3,BCLAF1,BHLHE40,BRF2,CEBPB,CEBPD,CEBPZ,CHD2,CLOCK,CREB1,CREB3L1,CREB3L2,CREBL2,E2F4,EGR1,EGR3,EHF,ELF3,ELF4,ELK3,ELK4,EP300,ESR1,ETS1,ETS2,ETV3,ETV4,ETV6,FLI1,FOSL2,FOXC1,FOXF2,FOXK1,FOXN1,FOXN2,FOXO1,HDAC2,HIVEP3,HLX,HOXB2,HSF4,IRF1,IRF2,IRF4,IRF9,JDP2,JUNB,JUND,JUN,KDM5A,KLF12,KLF1,KLF3,KLF4,KLF5,KLF6,KLF9,MAFB,MAFG,MAFK,MECP2,MIOS,MSC,MSX1,NFIL3,NFYA,NHLH1,NKX3-1,NR2F6,NR3C1,ONECUT1,OSR1,POLR2A,POLR3A,PPARGC1A,RAD21,RARA,RBBP5,SIN3A,SMARCC2,SOX4,SP1,SP2,SREBF2,SRF,TAF7,TEAD2,TEF,TFDP1,TFE3,THRA,THRB,UBTF,YBX1,YY1,ZBTB33,ZBTB7A,ZBTB7B,ZFHX3,ZIC5,ZNF148,ZNF319,ZNF358,ZNF362,ZNF429,ZNF467,ZNF785,ZNF787,ZNF91,ZXDA

Figura 4.9: Top regulones diferencialmente sobre-activos en tejidos con COVID-19.

Tejido COVID-19	p-value	Núm. regulones	Top regulones
Intestino	0.02424663865	103	AHCTF1,ARNTL2,CARF,CPEB1,CREM,CRX,DBX1,DBX2,DDIT3,DLX3,DMRTA2,E2F8,EBF2,EGR2,EMX1,EN2,EVX2,FERD3L,FIGLA,FOXN3,FOXO3,GABPA,GATA3,GFI1B,GFI1,GLI1,GMEB2,GSX1,GTF3C2,HDX,HIC1,HNF4G,HOXA2,HOXA3,HOXA4,HOXA6,HOXC11,HOXC4,HOXC6,HOXD3,LHX1,LHX6,LMX1A,MBD1,MEF2A,MEF2D,MESP1,MSRA,MXI1,MZF1,NFIC,NFYC,NKX2-1,NPDC1,NR1H4,NRL,OSR2,PAX8,PHOX2A,PLAG1,POLE3,POU2F2,PROP1,REL,RFX3,RFX4,RFX5,RFXANK,RORC,SALL1,SETDB1,SNAPC4,SP6,SPIB,STAT4,STAT5A,TAL1,TBX20,TFEB,THAP1,TP53,USF1,XBP1,ZBED1,ZBTB40,ZEB1,ZHX3,ZNF121,ZNF154,ZNF233,ZNF239,ZNF30,ZNF37A,ZNF502,ZNF559,ZNF582,ZNF583,ZNF594,ZNF607,ZNF836,ZSCAN20,ZSCAN4,ZSCAN5B
Corazón	0.00127163590	88	AHCTF1,AKR1A1,CARF,CPEB1,CREM,DBX2,DDIT3,DLX3,DMRTA2,E2F8,EGR2,EN2,FERD3L,FOXA2,FOXO3,FOXP2,GABPA,GFI1B,GFI1,GMEB2,GTF3C2,HDX,HES2,HIC1,HNF4G,HOXA3,HOXA4,HOXA6,HOXC11,HOXC4,HOXC6,HOXD3,HOXD4,LHX1,LHX6,LMX1A,MBD1,MEF2A,MEF2D,MESP1,MSRA,MXI1,NFIC,NFYC,NKX2-1,NPDC1,NRL,OSR2,PHOX2A,PLAG1,POLE3,POU2F2,PROP1,REL,RFX3,RFX5,RFXANK,SALL1,SETDB1,SNAPC4,SP6,SPIB,STAT5A,TAL1,TFEB,THAP1,USF1,XBP1,ZBED1,ZBTB40,ZEB1,ZFY,ZHX3,ZNF121,ZNF154,ZNF233,ZNF239,ZNF30,ZNF37A,ZNF559,ZNF582,ZNF583,ZNF594,ZNF607,ZNF681,ZNF836,ZSCAN20,ZSCAN4,ZSCAN5B
Riñón	0.03302888419	4	ARID3C,POU5F1,ZNF239,ZNF37A
Hígado	0.00358866371	99	AHCTF1,AHR,AKR1A1,ARNTL2,BSX,CARF,CERS5,CPEB1,CREM,CUX1,DBX2,DDIT3,DLX1,DLX3,DMRTA2,E2F8,EMX1,EN2,ESX1,EVX2,FERD3L,FOXI1,FOXO3,GABPA,GFI1B,GFI1,GTF3C2,HDX,HES2,HIC1,HNF4G,HOXA3,HOXA4,HOXA6,HOXB3,HOXC11,HOXC4,HOXC6,HOXD3,LHX1,LHX6,LMX1A,MBD1,MEF2A,MEF2D,MESP1,MSRA,MSX2,MXI1,NFIB,NFIC,NFYC,NKX2-1,NOBOX,NPDC1,NR1H4,NRF1,NRL,OSR2,PHOX2A,POLE3,POU2F2,PROP1,REL,RFX3,RFX5,RFXANK,SALL1,SETDB1,SMAD9,SNAPC4,SP6,SPIB,TAL1,TBX19,TFEB,THAP1,USF1,XBP1,ZBED1,ZBTB40,ZEB1,ZHX3,ZNF121,ZNF233,ZNF239,ZNF30,ZNF37A,ZNF559,ZNF582,ZNF583,ZNF594,ZNF607,ZNF681,ZNF836,ZNF846,ZSCAN20,ZSCAN4,ZSCAN5B
Lung	8.51E-13	4	ESRRB,FOXL2,ZNF154,ZNF233

Figura 4.10: Top regulones diferencialmente sub-activos en tejidos con COVID-19.

4.3.3. Análisis: Pulmón con COVID-19 en comparación con sano

La Tabla 4.11 indica el top regulon DA en pulmón con COVID-19 en comparación con pulmón sano que se encontraron para 3 estudios independientes mencionados anteriormente.

Activación Diferencial	p-value	Núm. regulones	Top regulones
Sobre-activado	9.59E-04	1	TP53
Sub-activado	8.26E-07	1	HES7

Figura 4.11: Top regulones diferencialmente sobre y sub-activos en Pulmón con COVID-19.

4.4. Exploración de Regulones

En esta sección se muestran los resultados de la exploración de los regulones seleccionados por los filtros de iteración en SCENIC, las pruebas de activación diferencial (DA) en casos contra controles y la especificidad (medida por RSS) para la condición dada (véase la sección 3.6) a través de visualizaciones como *Upset plots* y *heatmaps* para comparar los perfiles de activación de regulones por condición, y *dotplots* para visualizar los resultados del análisis de enriquecimiento de términos biológicos (véase la sección 3.7) dividido por cada uno de los tres análisis. Primero se exploran mediante estas visualizaciones los regulones resultantes de los dos primeros filtros de selección y después los regulones resultantes tras los tres filtros de selección.

4.4.1. Análisis: Infección de SARS-CoV-2 en comparación a otros virus en líneas celulares

Aunque en esta sección se enfoca en los regulones que resultaron relevantes durante la infección por SARS-CoV-2, es importante notar que aquellos experimentos infectados por otros virus que afectan el sistema respiratorio se utilizaron para comparar y enfocarnos en mecanismos regulatorios y funciones biológicas específicas de la infección por SARS-CoV-2.

4.4.1.1. Regulones sobre-activados

Comparación de Regulones distintamente compartidos entre condiciones
La Figura 4.12 es un Upset plot donde se muestra la cantidad de regulones distintamente compartidos entre cualesquiera dos a cinco experimentos de infecciones virales

en líneas celulares, y en la Tabla 4.13 se muestran cuáles son aquellos regulones distintamente compartidos sólo entre cualesquiera dos experimentos infectados por SARS-CoV-2 (es decir, aquellos que en el upset plot anterior tienen líneas azules encerradas en un rectángulo naranja) (véase sección 3.7.1), siendo estos los de mayor relevancia ya que tenemos certeza de que son las subunidades de regulación que se sobre activan ante la infección de SARS-CoV-2 específicamente.

De los experimentos infectados por SARS-CoV-2, la línea celular Calu3 fue la que compartió más regulones con otras líneas, lo cual tiene sentido porque es la que tuvo un mayor número de regulones DA. Esta infección en la línea A540 tratada con ACE2 y Ruxolitinib tuvo siete regulones DA únicos; *E2F7, NKX2-2, NR3C2, PBX3, TWIST1, ZNF460, ZNF701*, que deben corresponder a procesos regulatorios de la infección distintos de la respuesta a interferón. Similarmente, hay un regulon que se comparte distintamente cuando hay tratamiento de ACE2 pero no de Ruxolitinib, lo cual nos indica que este regulon, *NEUROD2*,debe ser particular de cuando las células son altamente susceptibles a la infección (por la expresión de ACE2) pero que no tiene que ver con interferón.

Hubo dos regulones distintamente compartidos entre los cinco experimentos infectados por SARS-CoV-2, indicando procesos biológicos en común altamente específicos de esta infección pero recurrente en distintas condiciones. Se habla de los posibles procesos biológicos en común en la siguiente sección.

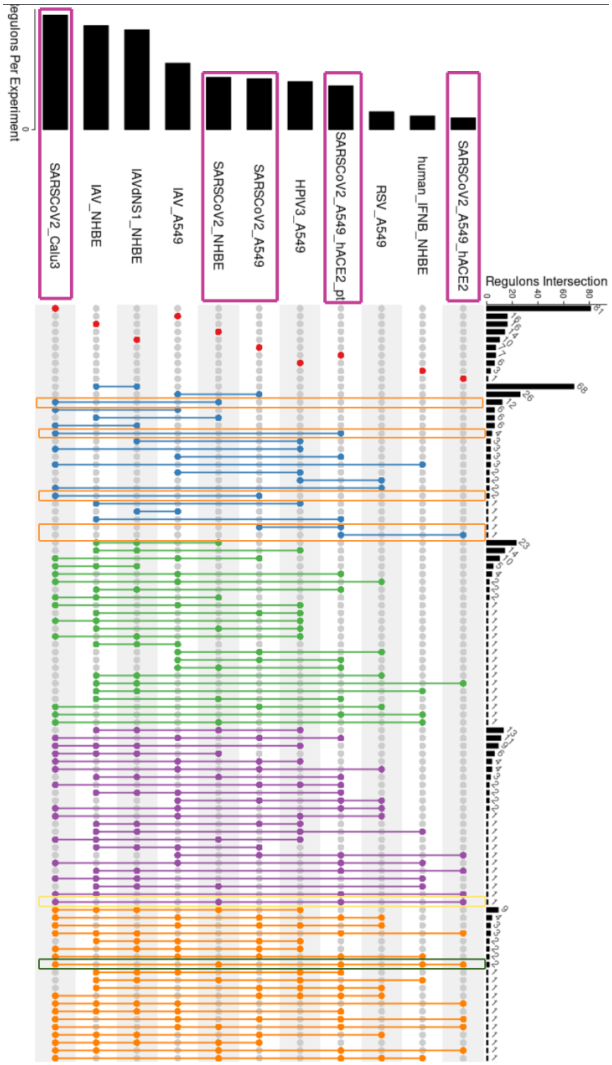


Figura 4.12: Upset Plot de los regulones sobre-activados distintamente compartidos entre condiciones en el análisis de líneas celulares. A la izquierda se enlistan los experimentos de líneas celulares tratadas (para una referencia véase la Tabla 3.1), a su izquierda en barras se muestra la cantidad total de regulones sobre-activados en el respectivo experimento y en la parte superior en un gráfico de barras se muestra la cantidad de regulones distintamente compartidos entre los experimentos indicados por uniones de puntos. Los puntos rojos indican los regulones que sólo están sobre-activados en un experimento dado, las líneas azules los regulones que sólo se comparten en dos experimentos, las líneas verdes los que sólo se comparten en tres experimentos y así sucesivamente. Con cuadros rosas horizontales señalan los experimentos de interés, *i.e.* experimentos infectados por SARS-CoV-2; con cuadros naranjas o rosa verticales se señalan los regulones compartidos entre esos experimentos de interés.

	SARSCoV2-inf-A549	SARSCoV2-inf-A549 +ACE2	SARSCoV2-inf-A549 +ACE2+pt	SARSCoV2-inf-Calu3	SARSCoV2-inf-NHBE
SARSCoV2-inf-A549	FLI1,FOXP4,NFIB,RB1,SNAI1,TLX1,ZFP64	NA	ZNF583	DRGX,TCF7L1	NA
SARSCoV2-inf-A549 +ACE2	NA	TP53	NEUROD2	NA	NA
SARSCoV2-inf-A549 +ACE2+pt	ZNF583	NEUROD2	E2F7,NKX2-2,NR3C2,PBX3,TWIST1,ZNF460,ZNF701	BRF2,FOXN1,MLXIPL,ZXDC	NA
SARSCoV2-inf-Calu3	DRGX,TCF7L1	NA	BRF2,FOXN1,MLXIPL,ZXDC	ATF6B,BRCA1,CIC,CREB3L1,CUX1,DBP,E2F1,E2F2,E2F8,ELK1,ESX1,ETV4,FEV,FOXJ2,FOXL1,FOXM1,FOXO4,GATA5,GATA6,GBX1,GLI1,GRHL2,GSC2,GTF3A,HIVEP1,HMGB1,HOXA6,HOXB4,HOXD1,HOXD10,HOXD8,HOXD9,KLF11,KLF12,KLF14,KLF15,KLF16,MEIS1,MESP1,MITF,MLXIP,MYBL1,MYPOP,NFIC,NFYC,NPAS2,NR1H3,OLIG3,OVOL1,PAX8,PHF2,PHF21A,PHF8,POLE3,POU2F1,POU4F1,POU6F1,PPARA,PRRX2,RAD21,SETDB1,SIX1,SMARCC2,SPDEF,SREBF2,SRF,STAT5B,TBX18,TBX6,TEAD4,TFDP1,TFDP2,TGIF2,TRIM28,USF1,VAX2,YBX1,ZNF354C,ZNF384,ZNF550,ZNF672	ARID5B,CEBPG,CREBL2,HDAC2,LHX1,MIOS,SMC3,SP1,SP2,TBPL2,WT1,ZFY
SARSCoV2-inf-NHBE	NA	NA	NA	ARID5B,CEBPG,CREBL2,HDAC2,LHX1,MIOS,SMC3,SP1,SP2,TBPL2,WT1,ZFY	EBF1,FOXP3,HOXA7,HOXB5,HOXB7,HOXC5,LHX8,NFYB,NR1I2,PAX2,PBX1,SP4,UBTF,ZNF580

Figura 4.13: Tabla complementaria de regulones sobre-activados distintamente compartidos entre condiciones en el análisis de líneas celulares. Los regulones que sólo se encuentran en un experimento dado se muestran en la diagonal de la tabla y su celda está coloreada en rosa fuerte (los puntos rojos en la Fig. 4.12). Los regulones distintamente compartidos entre experimentos distintos están resaltados en color naranja (líneas azules resaltadas en naranja en la Fig. 4.12). Los experimentos que no tuvieron regulones distintamente compartidos se indican por un "NA".

Asimismo, los regulones sub-activados distintamente compartidos entre los distintos experimentos en líneas celulares se muestran en la Figura 4.14 y qué regulones en concreto están siendo sub-regulados específicamente por SARS-CoV-2 se indican en la Tabla 4.15, que son los de mayor interés aquí.

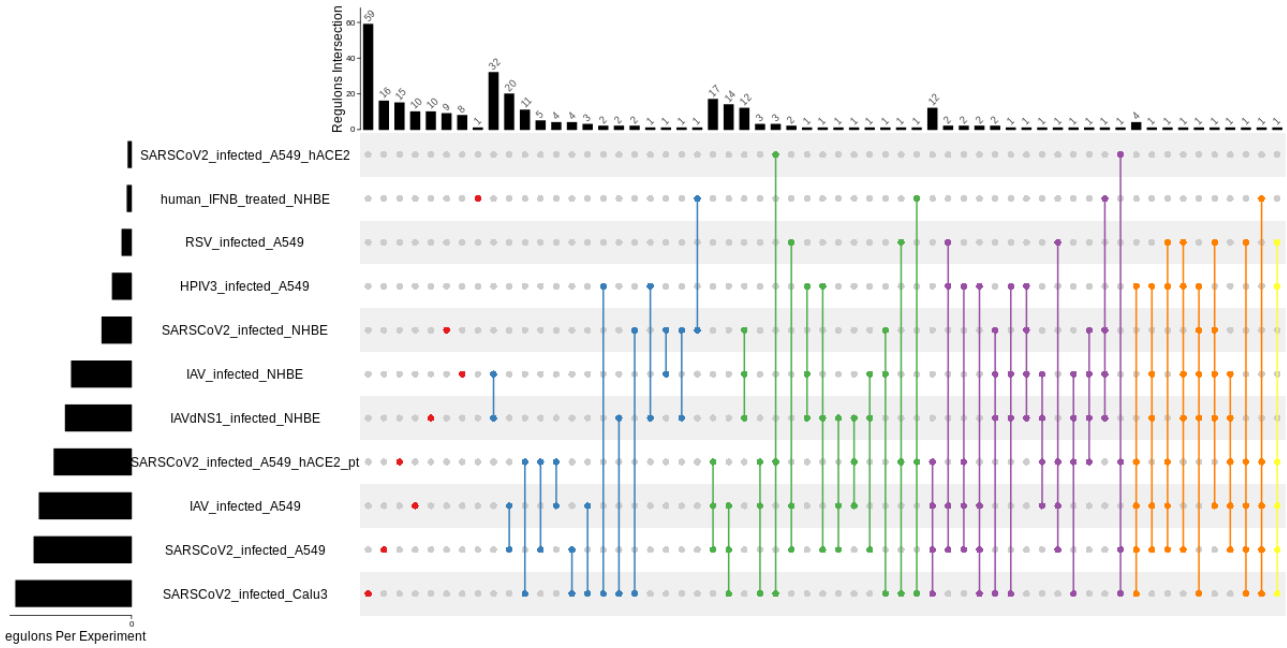


Figura 4.14: Upset Plot de los regulones sub-activados distintamente compartidos entre condiciones en el análisis de líneas celulares. A la izquierda se enlistan los experimentos de líneas celulares tratadas (para una referencia véase la Tabla 3.1), a su izquierda en barras se muestra la cantidad total de regulones sobre-activados en el respectivo experimento y en la parte superior en un gráfico de barras se muestra la cantidad de regulones distintamente compartidos entre los experimentos indicados por uniones de puntos. Los puntos rojos indican los regulones que sólo están sobre-activados en un experimento dado, las líneas azules los regulones que sólo se comparten en dos experimentos, las líneas verdes los que sólo se comparten en tres experimentos y así sucesivamente.

	SARSCoV2-inf-A549	SARSCoV2-inf-A549 +ACE2	SARSCoV2-inf-A549 +ACE2+pt	SARSCoV2-inf-Calu3	SARSCoV2-inf-NHBE
SARSCoV2-inf-A549	CELF6,ELK1,FEV,FOXQ1,GTF3A,HSF4,MESP1,NKX2-5,NR6A1,PATZ1,TAL2,TCF7L2,THRA,ZIC2,ZNF319,ZNF75A	NA	BHLHA15,DLX4,RFX1,SHOX2,TEAD2,ZFHX3	ATF6,FOXA3,FOXD2,MYOD1,ZNF486,ZNF513	NA
SARSCoV2-inf-A549 +ACE2	NA	NA	NA	NA	
SARSCoV2-inf-A549 +ACE2+pt	BHLHA15,DLX4,RFX1,SHOX2,TEAD2,ZFHX3	NA	CEBPA,EBF1,FOXP1,GSC2,HNF4G,HOXB1,HOXC9,LHX1,NR1D1,RFXAP,SREBF1,TAF1,TGIF1,ZBED1,ZNF579,ZNF652	AKR1A1,HSF1,KDM4D,NPDC1,NR4A1,PDX1,PRRX1,SP6,TBX20,TET1,ZNF276,ZNF668,ZNF721	NA
SARSCoV2-inf-Calu3	ATF6,FOXA3,FOXD2,MYOD1,ZNF486,ZNF513	NA	AKR1A1,HSF1,KDM4D,NPDC1,NR4A1,PDX1,PRRX1,SP6,TBX20,TET1,ZNF276,ZNF668,ZNF721	AR,ARID3C,ARX,ASCL1,ASCL2,BDP1,BNC2,BORCS8-MEF2B,CCNT2,CTCF,DUX4,DUXA,EOMES,EVX2,FBXL19,FOXA2,FOXC1,FOXF1,FOXG1,FOXJ3,GSC,HESX1,HOXB9,HOXC4,HOXC8,ISL2,LBX1,LHX4,LHX5,MAFK,MEF2C,MEOX2,NFE2L3,NKX6-2,NR0B1,NR1H2,NR1I2,NR5A2,OLIG2,PAX9,PBX1,PHOX2A,RBBP5,RFX7,SP9,STAT4,STAU2,TAL1,TBX15,TCF4,TLX1,VSX1,ZFHX2,ZNF264,ZNF441,ZNF606,ZNF713,ZNF8	BBX,DMRT1
SARSCoV2-inf-NHBE	NA	NA	NA	BBX,DMRT1	BCL6B,HSF2,IRF3,OVOL2,TWIST1,ZNF773

Figura 4.15: Tabla complementaria de regulones sub-activados distintamente compartidos entre condiciones en el análisis de líneas celulares. Los regulones que sólo se encuentran en un experimento dado se muestran en la diagonal de la tabla y su celda está coloreada en rosa fuerte (los puntos rojos en la Fig. 4.14). Los regulones distintamente compartidos entre experimentos distintos están resaltados en color naranja (líneas azules en la Fig. 4.14). Los experimentos que no tuvieron regulones distintamente compartidos se indican por un "NA".

Cada uno de los experimentos mostró un perfil de activación de regulones distinto, diferenciado en bloques por la infección de SARS-CoV-2. Se muestra que así fue el caso en las distintas líneas celulares utilizadas en las Figuras 4.16, 4.17 y 4.18. Como se muestra en el clusterizado de la activación de regulones, no hubo una coincidencia en el nivel de activación entre los regulones cuyos TFs pertenecen a un mismo DBD.

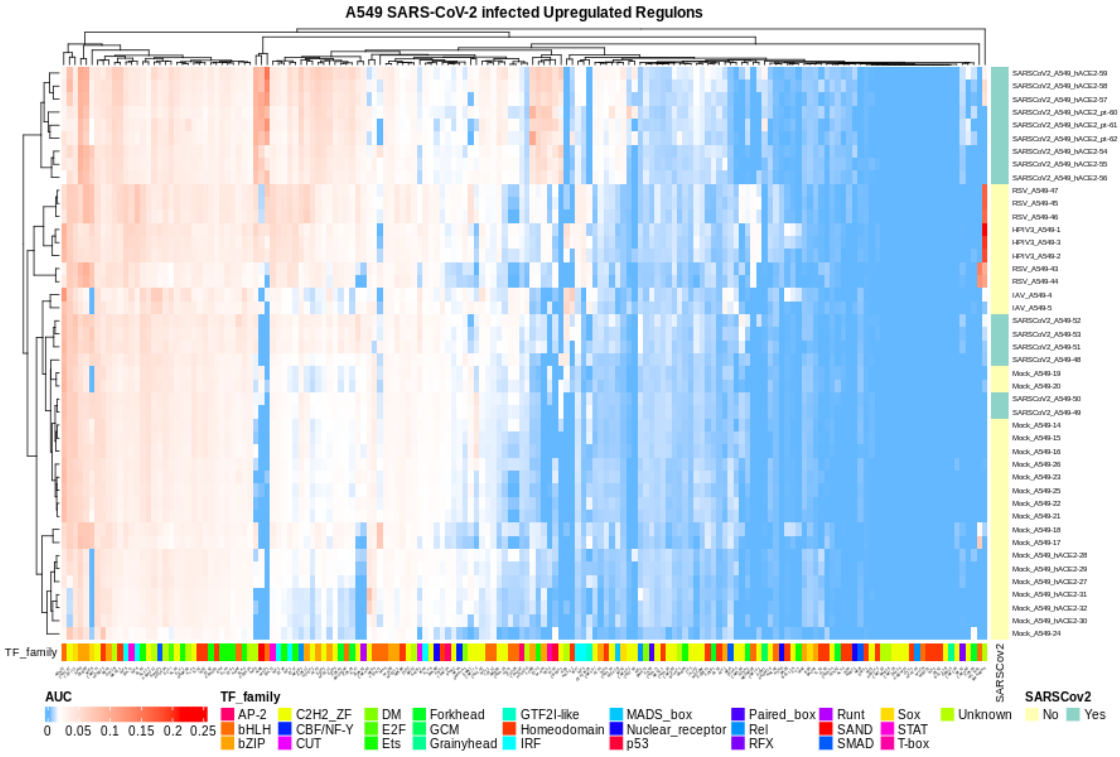


Figura 4.16: Heatmap clusterizado mostrando la activación de regulones (métrica AUC) en las muestras de experimentos en la línea de celular A549. En el eje x están los regulones sobre-activados en las muestras infectadas por SARS-CoV-2, indicados por el nombre de los factores transcripcionales guía de cada uno; mientras que el eje y está cada muestra. En la barra lateral vertical se indica si la muestra pertenece (verde turquesa) o no (amarillo) al experimento de infecciones por SARS-CoV-2. En la barra lateral horizontal se indica por colores la familia de factores transcripcionales a la que pertenece el TF guía del regulon. La leyenda indica las correspondencias de color y familia. La métrica AUC está en el rango de valores [0,0.3] donde estos valores fueron el valor mínimo y máximo de AUC, respectivamente.

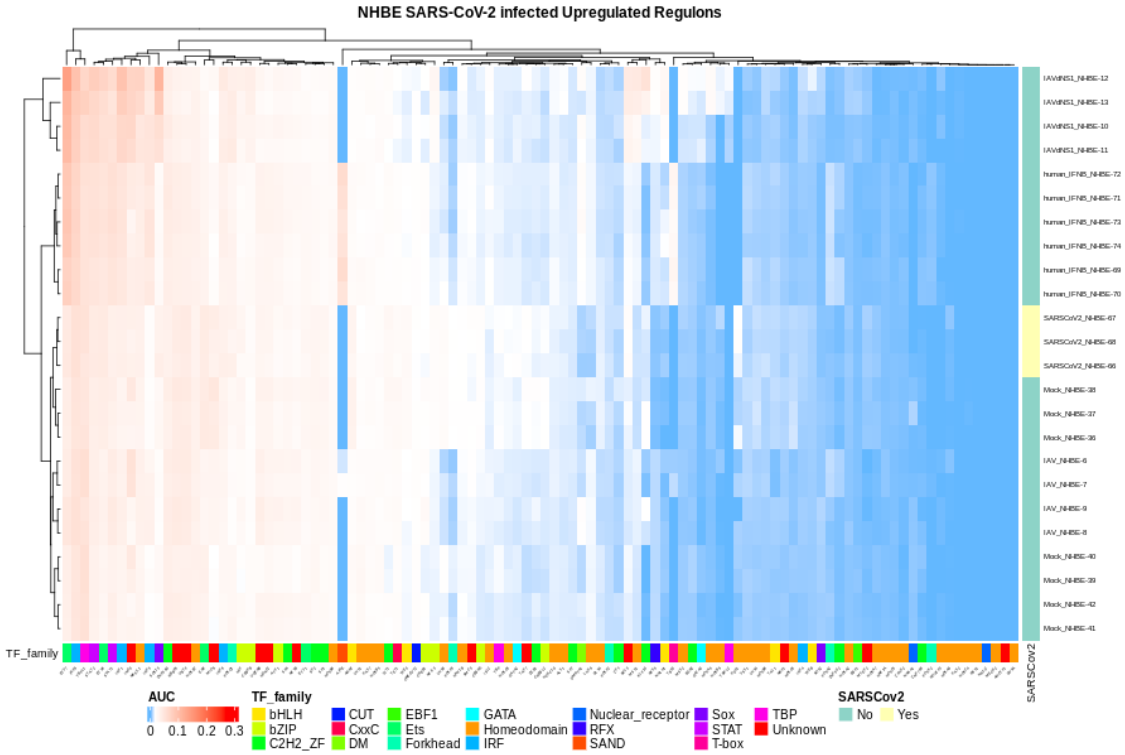


Figura 4.17: Heatmap clusterizado mostrando la activación de regulones (métrica AUC) en las muestras de experimentos en la línea de celular NHBE. En el eje x están los regulones sobre-activados en las muestras infectadas por SARS-CoV-2, indicados por el nombre de los factores transcripcionales guía de cada uno; mientras que el eje y está cada muestra. En la barra lateral vertical se indica si la muestra pertenece (amarillo) o no (verde turquesa) al experimento de infecciones por SARS-CoV-2. En la barra lateral horizontal se indica por colores la familia de factores transcripcionales a la que pertenece el TF guía del regulon. La leyenda indica las correspondencias de color y familia. La métrica AUC está en el rango de valores [0,0.3] donde estos valores fueron el valor mínimo y máximo de AUC, respectivamente.

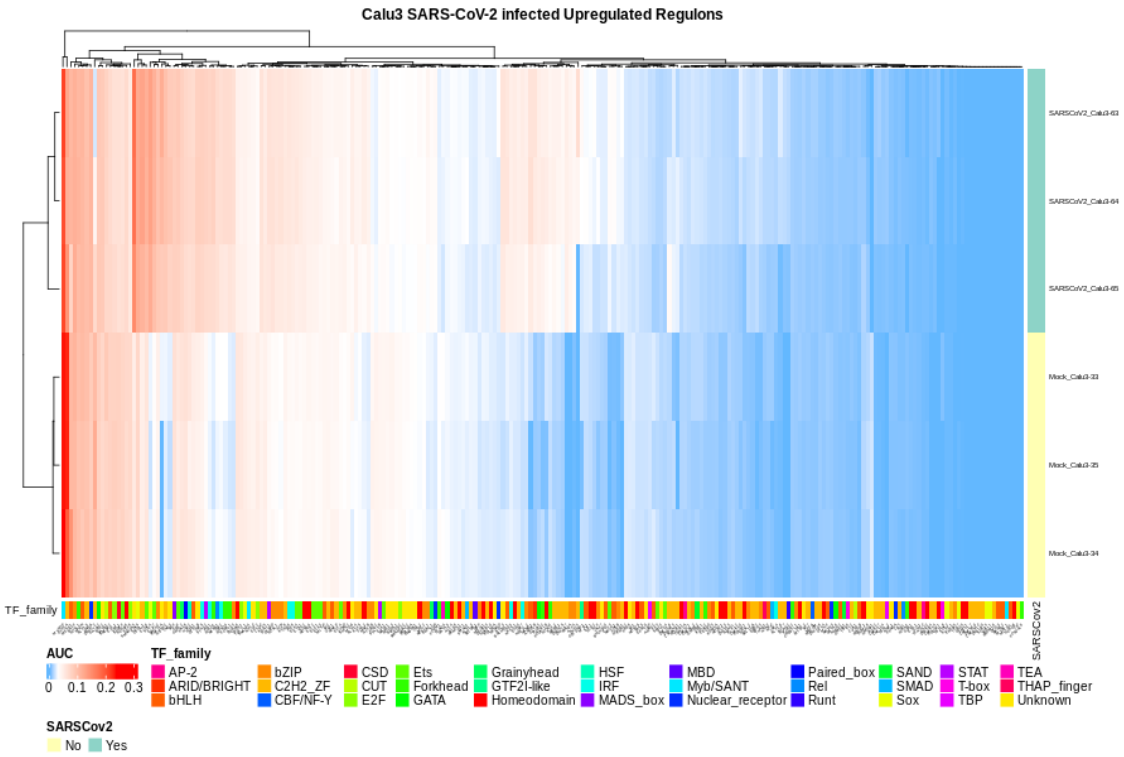


Figura 4.18: Heatmap clusterizado mostrando la activación de regulones (métrica AUC) en las muestras de experimentos en la línea de celular Calu3. En el eje x están los regulones sobre-activados en las muestras infectadas por SARS-CoV-2, indicados por el nombre de los factores transcripcionales guía de cada uno; mientras que el eje y está cada muestra. En la barra lateral vertical se indica si la muestra pertenece (verde turquesa) o no (amarillo) al experimento de infecciones por SARS-CoV-2. En la barra lateral horizontal se indica por colores la familia de factores transcripcionales a la que pertenece el TF guía del regulon. La leyenda indica las correspondencias de color y familia. La métrica AUC está en el rango de valores [0,0.3] donde estos valores fueron el valor mínimo y máximo de AUC, respectivamente.

Búsqueda de la función biológica de los regulones sobre-activados por SARS-CoV-2 y específicos por condición Como se mencionó en el marco teórico de la tesis presente, las subunidades de regulación suelen estar asociadas a subprocesos biológicos concretos que en conjunto dan lugar a uno más grande, como puede ser el desarrollo de la enfermedad de COVID-19 o bien, la recuperación del tejido afectado. Para este motivo, se buscaron los términos biológicos enriquecidos en los regulones sobre-activados ante la infección por SARS-CoV-2 y específicos por condición como se

detalla en la sección 3.7.3.

En general se vio que durante la infección de SARS-CoV-2, los términos biológicos asociados a líneas celulares fueron los siguientes.

- En general en los experimento de la línea A549 (véanse las Figuras 4.19, 4.20 y 4.21), se encontraron enriquecidos términos biológicos relacionados a la apoptosis, a la destrucción de proteínas por el proteosoma y la angiogenesis; lo cuál sugiere una respuesta de destrucción del virus y de reparación del tejido. De las vías KEGG repetidamente se encontraron enriquecidas vías activadas ante la presencia de distintos patógenos en la célula, así como vías con funciones implicadas en el sistema inmune innato (en los patrones de *C-type lektin receptor pathway*), inflamación por liberación de citoquinas (NF-kappaB, IL-17 y TNF) y en procesos apoptóticos (mTOR).
- En la línea NHBE (véase la Figura 4.22) también se encontraron enriquecidos términos relacionados a la potencial destrucción de proteínas virales, apoptosis y la recuperación del tejido por activación del crecimiento celular.

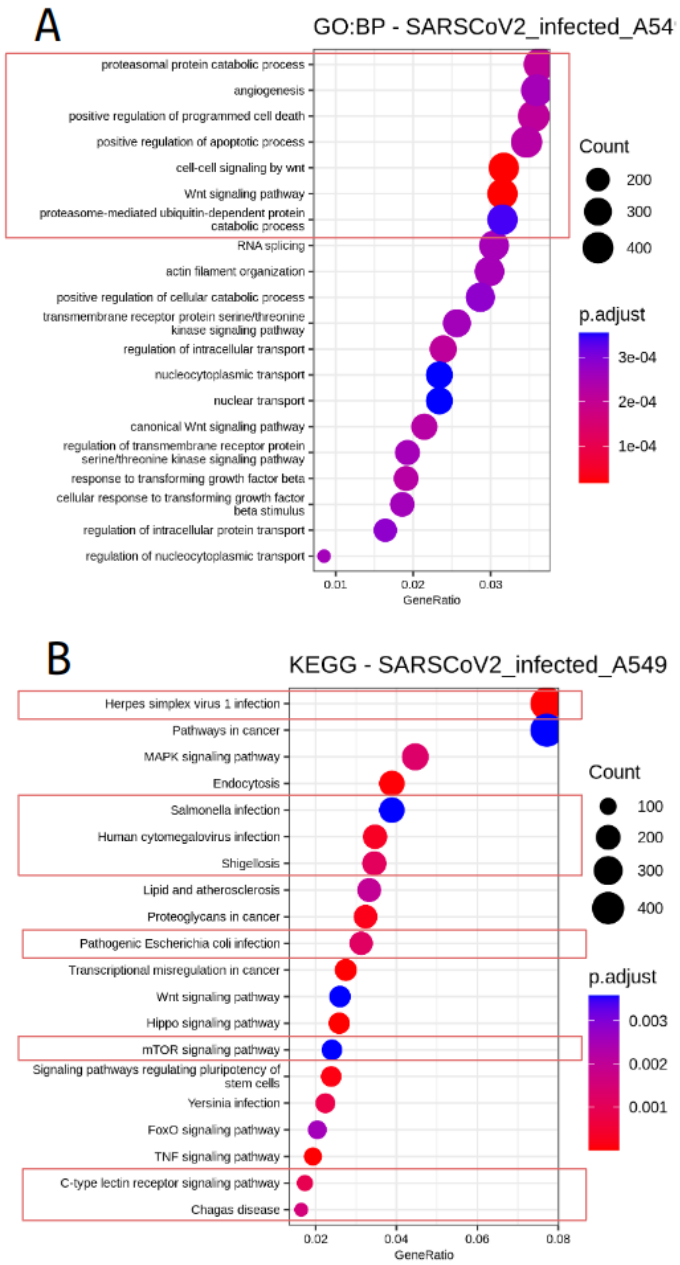


Figura 4.19: Dotplot del enriquecimiento de términos por ontología de genes, procesos biológicos (GO:BP, panel A) y por vías de señalización de KEGG (panel B) en los regulones sobre-activados de la línea celular A549 infectada por SARS-CoV-2. En el eje y está el top 14 de los términos que se encontraron más enriquecidos y en el x la proporción de genes etiquetados con ese término. El tamaño del punto indica la proporción de genes con ese término y el la escala de colores el p-valor ajustado. Los términos interesantes para el estudio presente están encerrados en cuadros rojos.

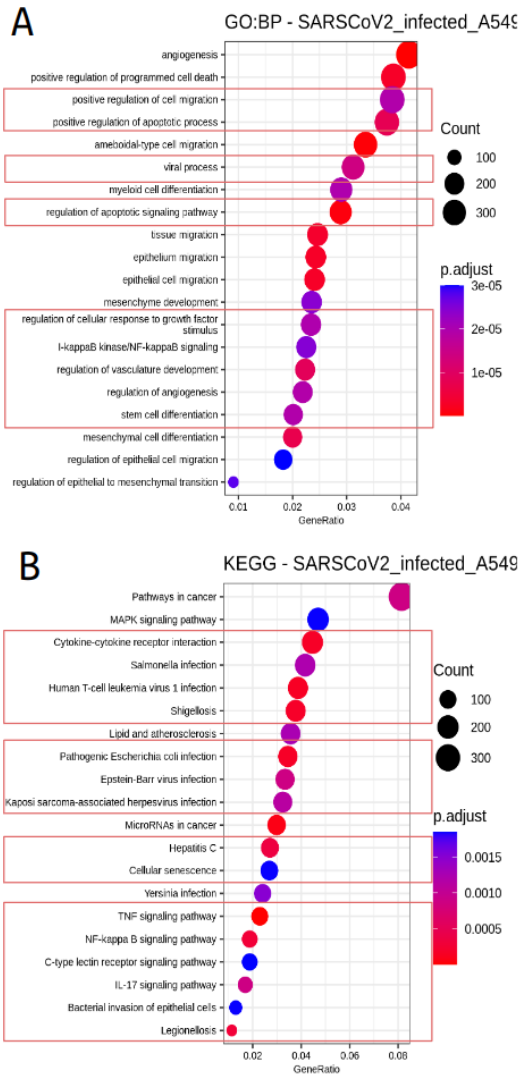


Figura 4.20: Dotplot del enriquecimiento de términos por ontología de genes, procesos biológicos (GO:BP, panel A) y por vías de señalización de KEGG (panel B) en los regulones sobre-activados de la línea celular A549 infectada por SARS-CoV-2 expresando ACE2. En el eje y está el top 14 de los términos que se encontraron más enriquecidos y en el x la proporción de genes etiquetados con ese término. El tamaño del punto indica la proporción de genes con ese término y el la escala de colores el p-valor ajustado. Los términos interesantes para el estudio presente están encerrados en cuadros rojos.

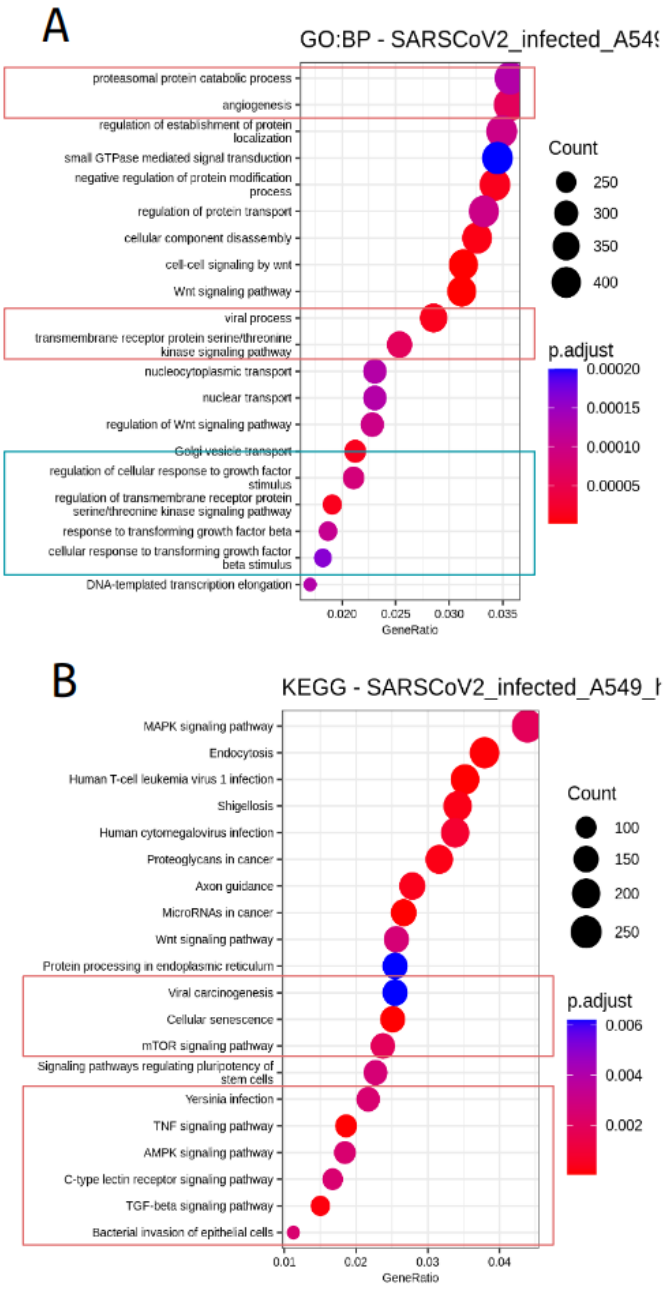


Figura 4.21: Dotplot del enriquecimiento de términos por ontología de genes, procesos biológicos (GO:BP, panel A) y por vías de señalización de KEGG (panel B) en los regulones sobre-activados de la línea celular A549 infectada por SARS-CoV-2 expresando ACE2 y con tratamiento de Ruxolitinib. En el eje y está el top 14 de los términos que se encontraron más enriquecidos y en el x la proporción de genes etiquetados con ese término. El tamaño del punto indica la proporción de genes con ese término y el la escala de colores el p-valor ajustado. Los términos interesantes para el estudio presente están encerrados en cuadros rojos.

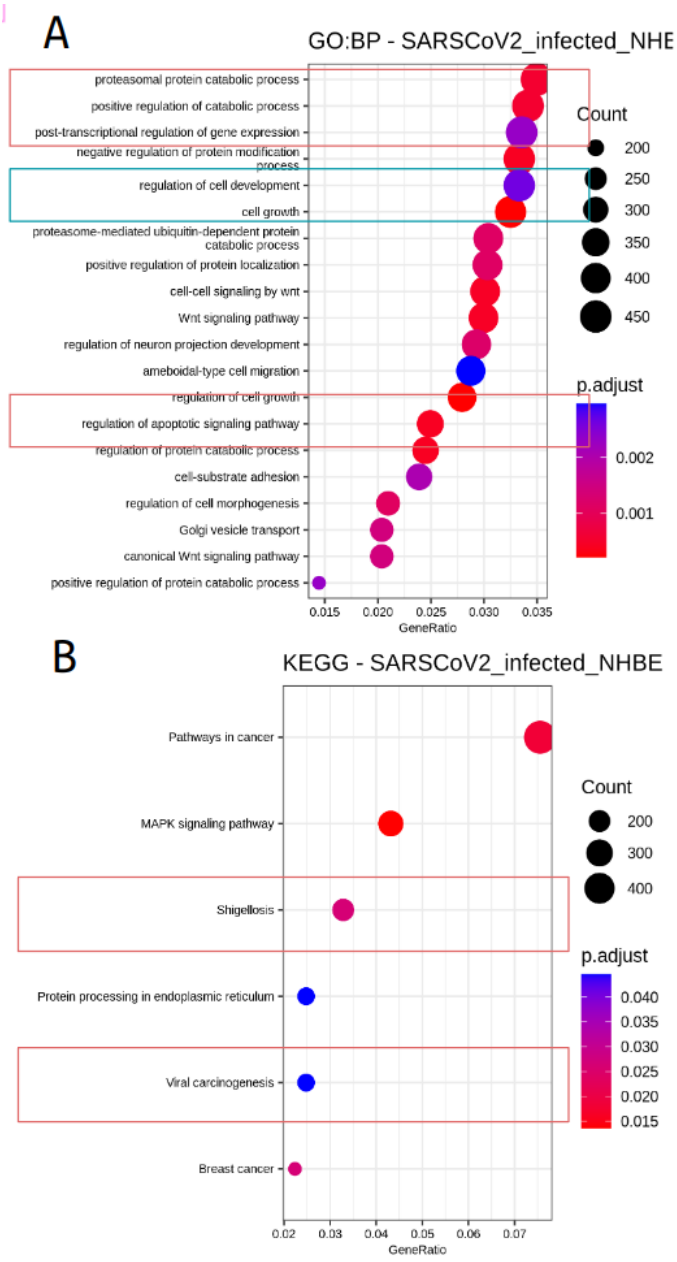


Figura 4.22: Dotplot del enriquecimiento de términos por ontología de genes, procesos biológicos (GO:BP, panel A) y por vías de señalización de KEGG (panel B) en los regulones sobre-activados de la línea celular NHBE infectada por SARS-CoV-2. En el eje y está el top 14 de los términos que se encontraron más enriquecidos y en el x la proporción de genes etiquetados con ese término. El tamaño del punto indica la proporción de genes con ese término y el la escala de colores el p-valor ajustado. Los términos interesantes para el estudio presente están encerrados en cuadros rojos.

4.4.1.2. Regulones sobre-activados y específicos por condición

Se repitió el análisis anterior tomando los regulones que estuvieran diferencialmente activados y que además también estuvieran más presentes en las pruebas infectadas de acuerdo a la métrica RSS de especificidad sobre las muestras infectadas.

Comparación de Regulones distintamente compartidos entre condiciones La Figura 4.23 es un upset plot que muestra todos los regulones distintamente compartidos en los experimentos de infecciones virales en líneas celulares, y en la Tabla 4.24 se muestran aquellos regulones distintamente compartidos solo entre cualesquiera dos experimentos infectados por SARS-CoV-2 (es decir, aquellos que en el upset plot anterior tinene líneas azules encerradas en un rectángulo naranja)(véase sección 3.7.1), que para enfatizar, no sólo son los regulones específicamente sobre-activados durante la infección por SARS-CoV-2 sino que además son aquellos que están más presentes en las muestras infectadas (de acuerdo al uso de la métrica RSS), por lo que estos regulones son de mayor interés biológico.

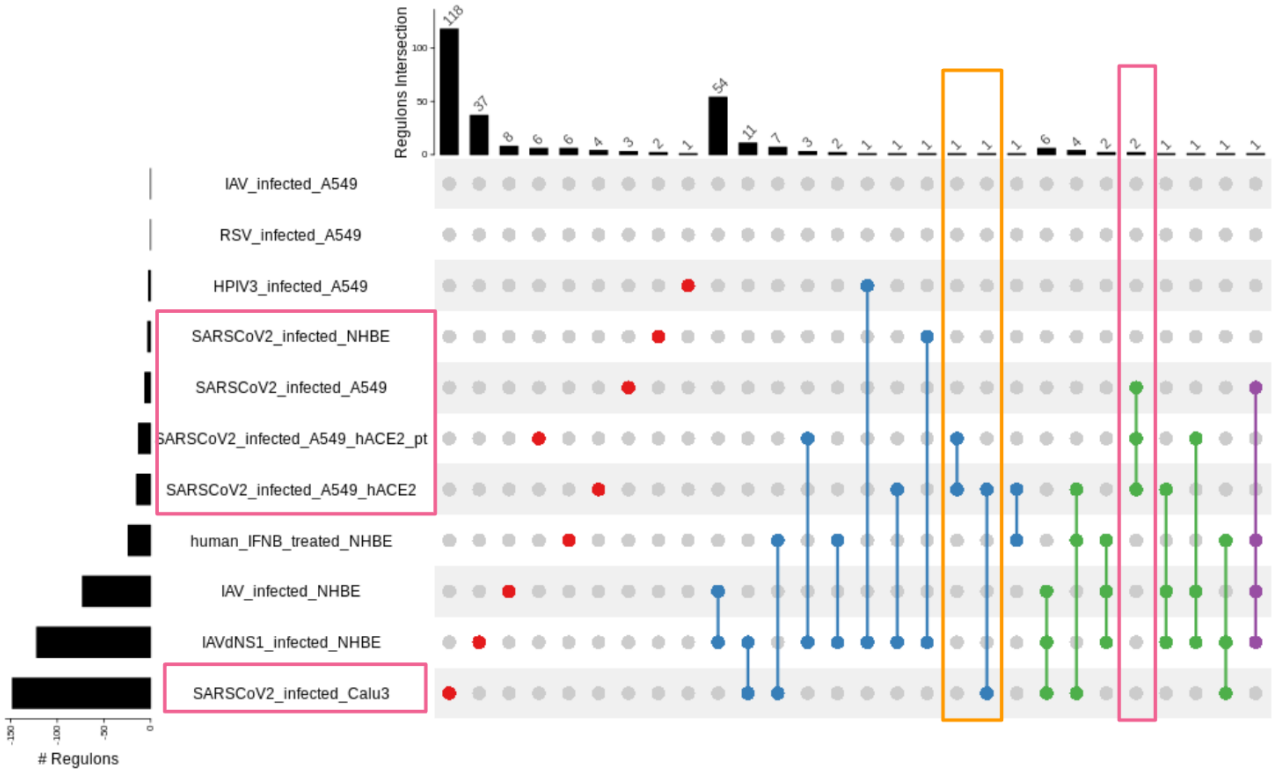


Figura 4.23: Upset Plot de los regulones sobre-activados y más específicos por condición distintamente compartidos entre condiciones en el análisis de líneas celulares. A la izquierda se enlistan los experimentos de líneas celulares tratadas (para una referencia véase la Tabla 3.1), a su izquierda en barras se muestra la cantidad total de regulones sobre-activados en el respectivo experimento y en la parte superior en un gráfico de barras se muestra la cantidad de regulones distintamente compartidos entre los experimentos indicados por uniones de puntos. Los puntos rojos indican los regulones que sólo están sobre-activados en un experimento dado, las líneas azules los regulones que sólo se comparten en dos experimentos, las líneas verdes los que sólo se comparten en tres experimentos y así sucesivamente. Con cuadros rosas horizontales señalan los experimentos de interés, *i.e.* experimentos infectados por SARS-CoV-2; con cuadros naranjas o rosa verticales se señalan los regulones compartidos entre esos experimentos de interés.

	SARSCoV2-inf-A549	SARSCoV2-inf-A549 +ACE2	SARSCoV2-inf-A549 +ACE2+pt	SARSCoV2-inf-Calu3	SARSCoV2-inf-NHBE
SARSCoV2-inf-A549	FOXP4,IRX4,MXD1	NA	NA	NA	NA
SARSCoV2-inf-A549 +ACE2	NA	XBP1,IRF5,STAT1,EMX1	NEUROD2	NA	NA
SARSCoV2-inf-A549 +ACE2+pt	NA	NEUROD2	ZNF460,ZNF730,POU2F3,NR3C2,ZNF701,NKX2-2	NA	NA
SARSCoV2-inf-Calu3	NA	NA	NA	ZNF354C,NHLH2,IKZF2,ZNF208,MECOM,ZNF471,ZNF560,ATF4,FOXN1,PRDM16,VDR,GSC2,ZNF429,KLF8,NFE2L1,PHF21A,ZNF134,GRHL1,RUNX2,GATA6,GMEB1,HIVEP1,ZNF554,SATB1,KLF7,FOXJ2,IRX2,KLF12,TBX6,PPARD,KLF9,BCL11A,CREB5,MITF,POLR2A,IRF6,NFYA,RREB1,FOXO4,TFAP2A,ZNF384,SIX2,EHF,BATF,EP300,TCF7L1,ZBTB33,DDIT3,NFKB1,CUX1,DLX3,ELK3,POLE3,KLF6,MIOS,POLR3A,TGIF2,HES1,YY2,ARNT,MAFB,MEIS1,ARNTL,SOX15,MLXIP,KDM5A,ETV4,ELF4,ZNF143,MAX,YBX1,GABPB1,ETS1,ZNF267,OLIG3,NPAS2,TFAP2C,PKNOX1,POU6F1,ELF1,TFDP1,FOXM1,HSF2,KLF13,HINFP,ELK4,CEBPG,ARNT2,ETS2,SIX1,DRGX,SP1,TFE3,HMGB1,ZFY,KLF16,GTF2IRD1,ZEB1,SNAPC4,SMAD1,NRF1,SP7,CREBL2,FOSB,ZNF550,MAFA,CREB3L1,HOXD10,ZNF607,GATA5,PRRX2,DLX4,FOSL2,SMC3,LHX1,GLI1,FOXF2,TBX18	ARID5B
SARSCoV2-inf-NHBE	NA	NA	NA	ARID5B	NR1I2,AHCTF1

Figura 4.24: Tabla complementaria de regulones sobre-activados y más específicos distintamente compartidos entre condiciones en el análisis de líneas celulares. Los regulones que sólo se encuentran en un experimento dado se muestran en la diagonal de la tabla y su celda está coloreada en rosa fuerte (los puntos rojos en la Fig. 4.23). Los regulones distintamente compartidos entre experimentos distintos están resaltados en color naranja (líneas azules resaltadas en naranja en la Fig. 4.23). Los experimentos que no tuvieron regulones distintamente compartidos se indican por un "NA".

Nuevamente, los experimentos mostraron tener un perfil de activación y especificidad por ciertos regulones diferenciado entre aquellos infectados por SASR-CoV-2 y otras infecciones virales u otros tratamientos como la expresión promovida de IFN en NHBE (véase la Fig. 4.25), experimento que mostró tener un fuerte perfil de activación de casi de todos los regulones relevantes en la infección de SARS-CoV-2, lo cuál tiene sentido porque como se ha mencionado antes en la tesis presente, la fisiopatología del COVID-19 ha sido caracterizada por la alta inducción de la inflamación, proceso biológico subsecuente la activación de interferón.

En particular en la Fig. 4.25 A se puede ver que los regulones guiados por los TFs FOXP4 y ZNF730 se comparten entre todos los experimentos infectados por SARS-CoV-2 y ningún otro. También cabe notar que los regulones guiados por (o regulones, simplemente) SMARCA5, RGXR, NEUROD2 e IRF5 parecen ser particularmente importantes en la infección por SARS-CoV-2 en la línea celular A549.

Por otro lado, para los regulones sobre-activados y más específicos en Calu3 (Fig. 4.25 B) durante la infección de SARS-CoV-2, podemos ver que los regulones VDR, ZNF471, ZNF485, ZNF208, PROM10, ZNF560 y NHLH2 fueron importantes para los experimentos de A549 y Calu3 infectados por SARS-CoV-2 y ningún otro. Los regulones TFAP2C, GTFZIRD1 y DRGX tienen un mayor nivel de RSS en los experimentos de infección por SARS-CoV-2 en NHBE y A549.

También podemos ver que, nuevamente, los regulones más activados en ciertas con-

diciones no coinciden en los DBDs de sus TFs guía, sin embargo, si fue posible ver que algunos de ellos (7 y 22 en la Fig. 4.25 panel A y B, respectivamente) son TFs activos en la respuesta inmune.

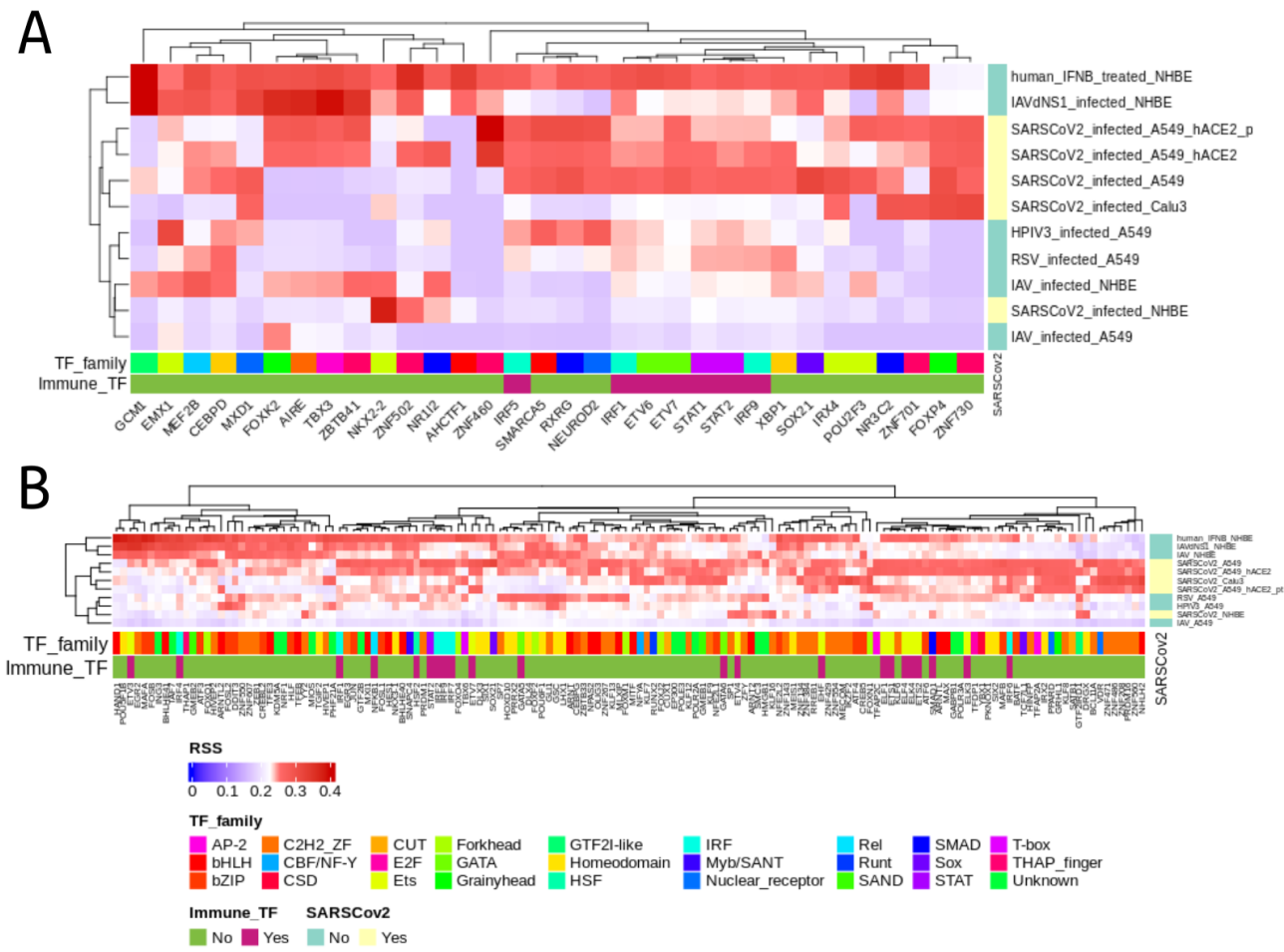


Figura 4.25: Heatmap clusterizado mostrando la especificidad de regulones por experimento o condición (RSS) de los regulones sobre-activados y más específicos en la infección de SARS-CoV-2 en líneas celulares de epitelio pulmonar. En el eje x están los regulones sobre-activados y más específicos en las muestras infectadas por SARS-CoV-2 en A549 y NHBE (panel A) o Calu3 (panel B), indicados por el nombre de los factores transcripcionales guía de cada uno; mientras que el eje y está cada experimento de línea celular. En la barra lateral vertical se indica si la muestra pertenece (amarillo) o no (verde turquesa) al experimento de infecciones por SARS-CoV-2. En la primera barra lateral horizontal se indica por colores la familia de factores transcripcionales a la que pertenece el TF guía del regulon, en la segunda barra horizontal se indica si el TF guia del regulon está asociado con alguna actividad inmunológica (rosa fuerte) o no (verde). La leyenda indica las correspondencias de color y familia, y la correspondencia de color y ausencia/presencia de SARS-CoV-2 y TF inmune. La métrica RSS está en el rango de valores [0,0.4].

Búsqueda de la función biológica de los regulones sobre-activados por SARS-CoV-2 y específicos por condición Una vez que se identificaron los regulones de mayor interés al compararlos, se quiso saber cuáles eran las funciones biológicas asociadas a estos. En general se vio que durante la infección de SARS-CoV-2, los términos biológicos asociados a líneas celulares fueron los siguientes.

- De manera interesante se encontró que los regulones de NHBE (Figura 4.26) tienen varios términos enriquecidos relacionados con el silenciamiento de genes a través de micro RNAs (miRNA). En el enriquecimineto de vías KEGG se encontró que una vía de infección patógena está potencialmente activa.
- Para el caso de los regulones de Calu3 (Figura 4.27), se vio que vías de migración y señalización entre células están potencialmente activas, así como un factor de crecimiento que podría sugerir una respuesta reparativa ante el daño causado por la infección. Asimismo, se encuentran enriquecidas vías activadas ante la presencia de patógenos en la célula, así como nuevamente la vía de señalización TGF, implicada en el crecimiento y contenedora de las citoquinas SMAD.

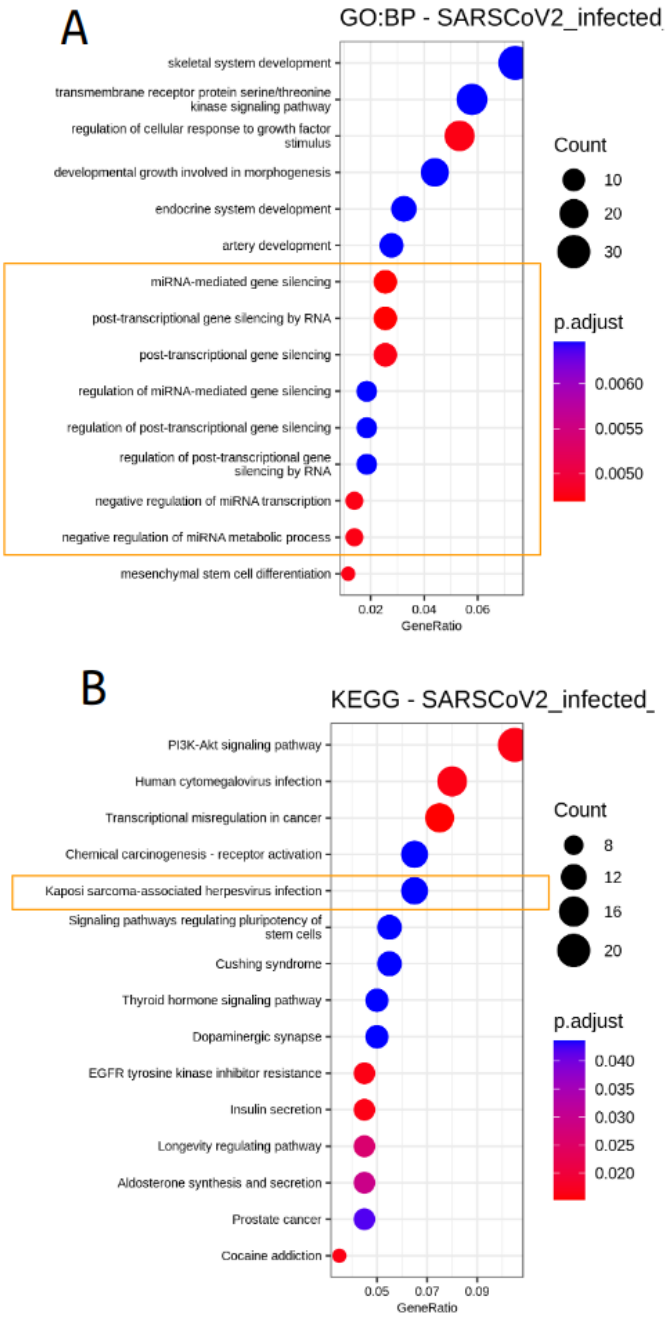


Figura 4.26: Dotplot del enriquecimiento de términos por ontología de genes, procesos biológicos (GO:BP, panel A) y por vías de señalización de KEGG (panel B) en los regulones sobre-activados y más específicos de la línea celular NHBE. En el eje y está el top 14 de los términos que se encontraron más enriquecidos y en el x la proporción de genes etiquetados con ese término. El tamaño del punto indica la proporción de genes con ese término y el la escala de colores el p-valor ajustado. Los términos interesantes para el estudio presente están encerrados en un cuadro naranja.

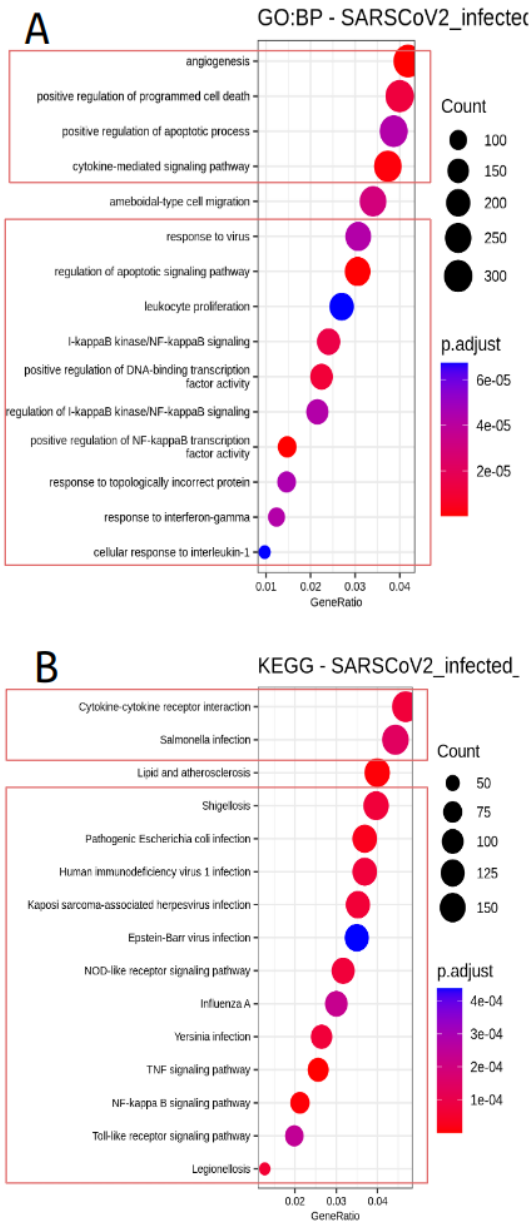


Figura 4.27: Dotplot del enriquecimiento de términos por ontología de genes, procesos biológicos (GO:BP, panel A) y por vías de señalización de KEGG (panel B) en los regulones sobre-activados y más específicos de la línea celular Calu3. En el eje y está el top 14 de los términos que se encontraron más enriquecidos y en el x la proporción de genes etiquetados con ese término. El tamaño del punto indica la proporción de genes con ese término y el la escala de colores el p-valor ajustado.

Particularmente interesante, fue que los regulones RXRG y SMARCA5, que se mencionaron anteriormente por aparecer sólo en experimentos de infección por SARS-CoV-2, están enriquecidos en respuesta a virus y muerte celular (véase Fig. 4.28).

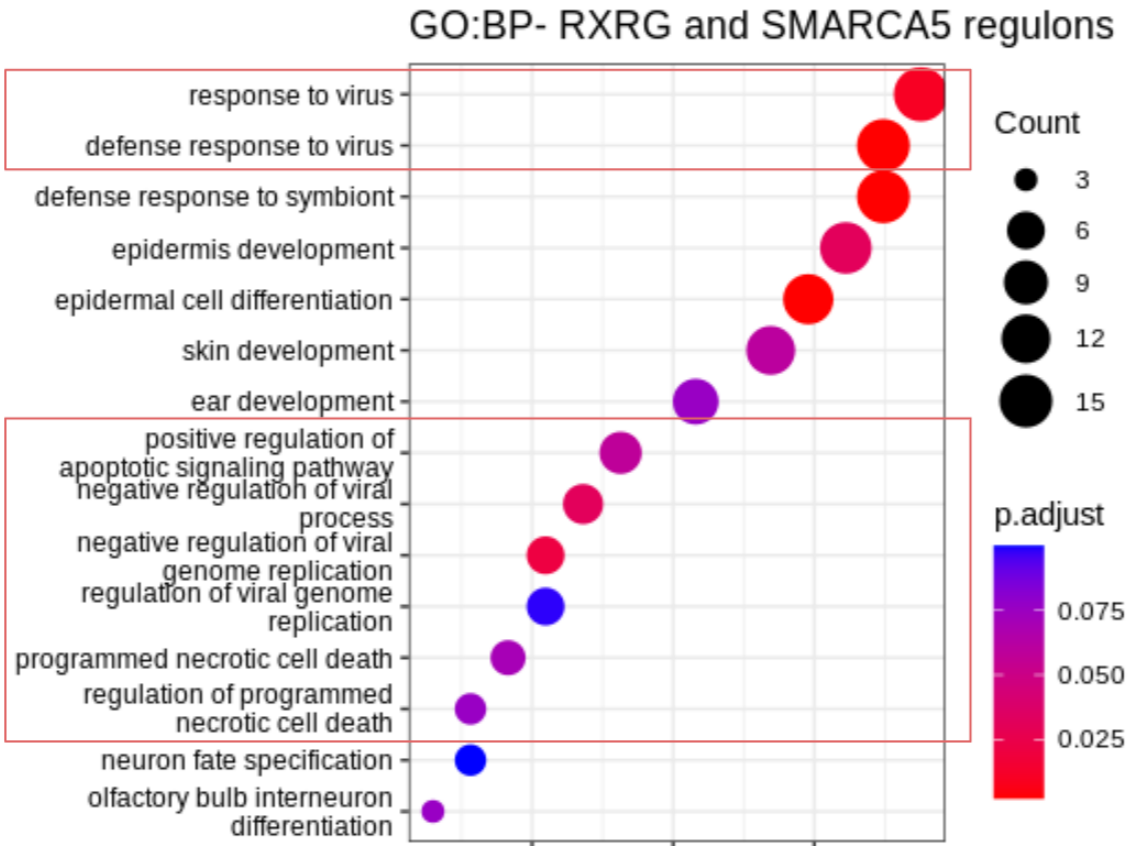


Figura 4.28: Dotplot del enriquecimiento de términos por ontología de genes, procesos biológicos (GO:BP) en los regulones RXRG y SMARCA5. En el eje y está el top 14 de los términos que se encontraron más enriquecidos y en el x la proporción de genes etiquetados con ese término. El tamaño del punto indica la proporción de genes con ese término y el la escala de colores el p-valor ajustado.

4.4.2. Análisis: COVID-19 en distintos tejidos

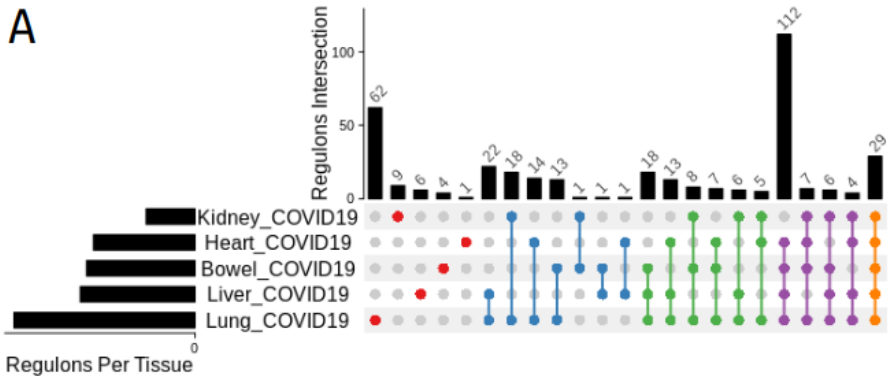
4.4.2.1. Regulones sobre-activados

Comparación de Regulones distintamente compartidos entre condiciones En la Figura 4.29 panel A se muestran los regulones distintamente compartidos en tejidos con COVID-19, y en el panel B se muestran aquellos regulones distintamente

compartidos entre cualesquiera dos tejidos con COVID-19 (véase sección 3.7.1), donde estos resultan de interés ya que indica que hay procesos biológicos específicos que son importantes y se prenden en común ante la infección.

Asimismo, los regulones sub-activados distintamente compartidos entre tejidos con COVID-19 se muestran en la Figura 4.30 A y qué regulones en concreto están siendo sub-regulados específicamente se indican en la Figura 4.30 B, que podrían ser procesos celulares innecesarios ante la infección y necesarios de sub-regular para optimizar recursos. De manera antagónica a lo esperado, se encontró que los regulones guiados por los factores transcripcionales IRF3 e IRF7 se encontraron como sub-activados, pero IRF1, IRF2, IRF4, IRF8 y IRF9 sí se encontraron sobre-activados.

A

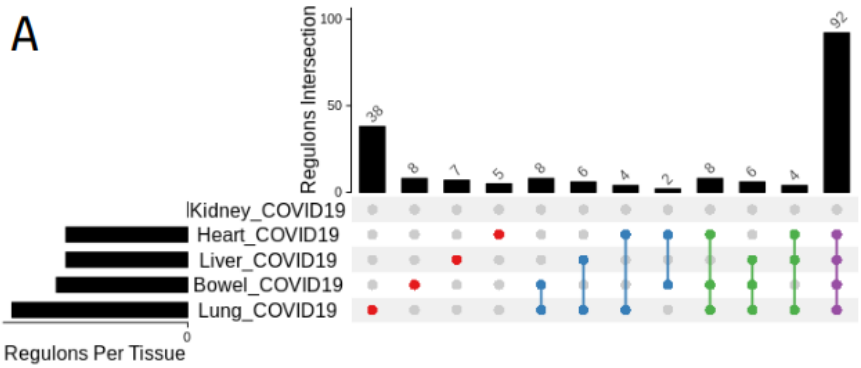


B

	Bowel COVID19	Heart COVID19	Kidney COVID19	Liver COVID19	Lung COVID19
Bowel_COVID19	HOXC9,POLD2,TBX21,ZNF726	NA	ZNF782	STAT4	FOXA1,FOXA3,FOXC1,KLF2,KLF7,MSC,NANOG,NANOS1,NEUROD1,ONECUT1,TCF7L2,THRB,VSX2
Heart_COVID19	NA	HMX3	NA	SNAI3	ERG,FLI1,FOXP3,GRHL2,GTF2B,MLXIPL,NFKB2,NFYB,OTP,REST,RUNX1,SP4,TEAD1,TFEC
Kidney_COVID19	ZNF782	NA	BARHL1,BHLHE22,FUBP1,NFAT5,RBAK,ZNF189,ZNF260,ZNF420,ZNF625	NA	BCL11B,FOXC2,FOXJ3,FOXK2,HOXB6,MEIS1,MXD1,PLAGL2,POU2AF1,PRRX2,PSMC2,RFX7,RUNX2,SIX1,SMC3,UBP1,VEZF1,ZNF667
Liver_COVID19	STAT4	SNAI3	NA	CUX1,EMX2,NFE2,NR1H2,NRF1,ZNF846	ATF5,BARX2,CEBPA,E2F2,EOMES,FOSB,GMEB1,HSF1,IKZF1,KDM4B,MLXIP,PATZ1,POLR3G,RFX1,SOX17,SOX6,STAT2,TCF7L1,TFAP2C,TFAP4,ZFP64,ZNF513
Lung_COVID19	FOXA1,FOXA3,FOXC1,KLF2,KLF7,MSC,NANOG,NANOS1,NEUROD1,ONECUT1,TCF7L2,THRB,VSX2	ERG,FLI1,FOXP3,GRHL2,GTF2B,MLXIPL,NFKB2,NFYB,OTP,REST,RUNX1,SP4,TEAD1,TFEC	BCL11B,FOXC2,FOXJ3,FOXK2,HOXB6,MEIS1,MXD1,PLAGL2,POU2AF1,PRRX2,PSMC2,RFX7,RUNX2,SIX1,SMC3,UBP1,VEZF1,ZNF667	ATF5,BARX2,CEBPA,E2F2,EOMES,FOSB,GMEB1,HSF1,IKZF1,KDM4B,MLXIP,PATZ1,POLR3G,RFX1,SOX17,SOX6,STAT2,TCF7L1,TFAP2C,TFAP4,ZFP64,ZNF513	ALX3,ARID3A,ARNTL,ATF4,ATF6B,BCL6,CCDC25,CRX,DMBX1,E2F1,E2F3,E2F7,ELF2,EN1,ERF,ESRRA,FOXB2,FOXM1,GATA2,GATA6,GCM1,GSX2,HAND1,HEY1,HNF1B,HOXB7,HOXB8,HOXC8,HSF4,MAFB,MECOM,NFATC3,NR1D1,NR1D2,NR2C2,NR2E3,ONECUT3,POU1F1,POU2F1,PPARGC1A,RFX2,RORC,RXRB,SALL2,SIX2,SIX5,SMAD1,SOX1,SOX5,SP3,STUB1,TAF1,TRIM28,ZBTB14,ZFP69B,ZHX1,ZNF214,ZNF282,ZNF384,ZNF436,ZNF653,ZNF76

Figura 4.29: Upset Plot y tabla complementaria de los regulones sobre-activados distintamente compartidos entre condiciones en el análisis de tejidos COVID-19. A) A la izquierda se enlistan los tejidos con COVID-19, a su izquierda en barras se muestra la cantidad total de regulones sobre-activados en el respectivo tejido. En la parte superior en un gráfico de barras se muestra la cantidad de regulones compartidos. B) Los regulones que sólo se encuentran en un tejido dado se muestran en la diagonal de la tabla. Los regulones compartidos entre tejidos distintos están resaltados en color naranja y los tejidos que no tuvieron regulones distintamente compartidos se indican por un "NA".

A



B

	Bowel COVID19	Heart COVID19	Kidney COVID19	Liver COVID19	Lung COVID19
Bowel COVID19	CRX,DBX1,GATA3,GLI1,GSC,GSX1,RORC,TP53	TBX2,ZNF502	NA	NA	FOXN3,HOXA2,MZF1,PAX8,RFX4,ZBTB2,ZNF18,ZNF341
Heart COVID19	TBX2,ZNF502	CDX1,HES5,HOXD4,NPAS2,ZFY	NA	NA	FOXA2,HOXA1,ZNF768,ZSCAN16
Kidney COVID19	NA	NA	NA	NA	NA
Liver COVID19	NA	NA	NA	AHR,CERS5,HOXD1,MSX2,NFIB,NOBOX,SMAD9	BSX,DLX1,ESX1,FOXI1,HOXB3,TBX18
Lung COVID19	FOXN3,HOXA2,MZF1,PAX8,RFX4,ZBTB2,ZNF18,ZNF341	FOXA2,HOXA1,ZNF768,ZSCAN16	NA	BSX,DLX1,ESX1,FOXI1,HOXB3,TBX18	ASCL2,ATOH7,BRF1,CELF4,DMRT2,ESRRB,EVX1,FOXD4L1,FOXL2,GABPB1,GBX1,GLI3,HOXC10,HOXC13,ING4,INSM1,IRF3,IRF7,LEF1,MDM2,MEF2C,PAX1,POU5F1B,POU6F1,SP5,TBX5,TFAP2E,TGIF1,USF2,ZNF235,ZNF3,ZNF439,ZNF540,ZNF585A,ZNF660,ZNF729,ZNF831,ZXDC

Figura 4.30: Upset Plot y tabla complementaria de los regulones sub-activados distintamente compartidos entre condiciones en el análisis de tejidos COVID-19. A) A la izquierda se enlistan los tejidos con COVID-19, a su izquierda en barras se muestra la cantidad total de regulones sobre-activados en el respectivo tejido. En la parte superior en un gráfico de barras se muestra la cantidad de regulones compartidos. B) Los regulones que sólo se encuentran en un tejido dado se muestran en la diagonal de la tabla. Los regulones compartidos entre tejidos distintos están resaltados en color naranja y los tejidos que no tuvieron regulones distintamente compartidos se indican por un "NA".

Hay una clara diferencia en el perfil de activación de los regulones sobre-activados en tejidos de pacientes con COVID-19 en comparación con controles sanos como se puede ver en la Figura 4.31. De nuevo, no hubo una coincidencia en el nivel de activación entre los regulones cuyos TFs pertenecen a un mismo DBD.

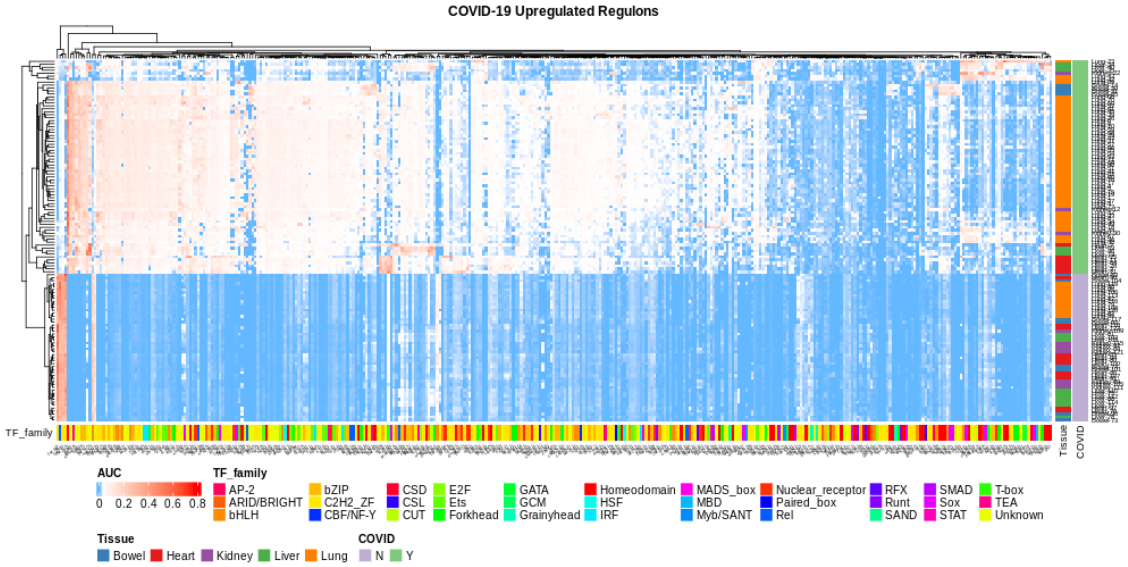


Figura 4.31: Heatmap clusterizado mostrando la activación de regulones (métrica AUC) en las muestras de tejidos COVID-19. En el eje x están los regulones sobre-activados, indicados por el nombre de los factores transcripcionales guía de cada uno; mientras que el eje y está cada muestra. En la barra lateral vertical se indica si la muestra pertenece (verde turquesa) o no (amarillo) a un tejido con COVID-19. En la barra lateral horizontal se indica por colores la familia de factores transcripcionales a la que pertenece el TF guía del regulon. La leyenda indica las correspondencias de color y familia. La métrica AUC está en el rango de valores [0,0.3] donde estos valores fueron el valor mínimo y máximo de AUC, respectivamente.

Búsqueda de la función biológica de los regulones sobre-activados por SARS-CoV-2 y específicos por condición Se quiso saber cuáles eran las funciones biológicas asociadas a este perfil de regulones sobre-activados durante la COVID-19, por lo que se hicieron pruebas de enriquecimiento de términos biológicos. Se encontró que en el riñón (véase la Figura 4.32) parecen estar activándose vías de migración de tejido y vías que participan en la regulación de la apoptosis celular (Hippo y FoxO), lo cuál puede ser indicativo de que estos procesos están ocurriendo en el riñón ante el daño causado.

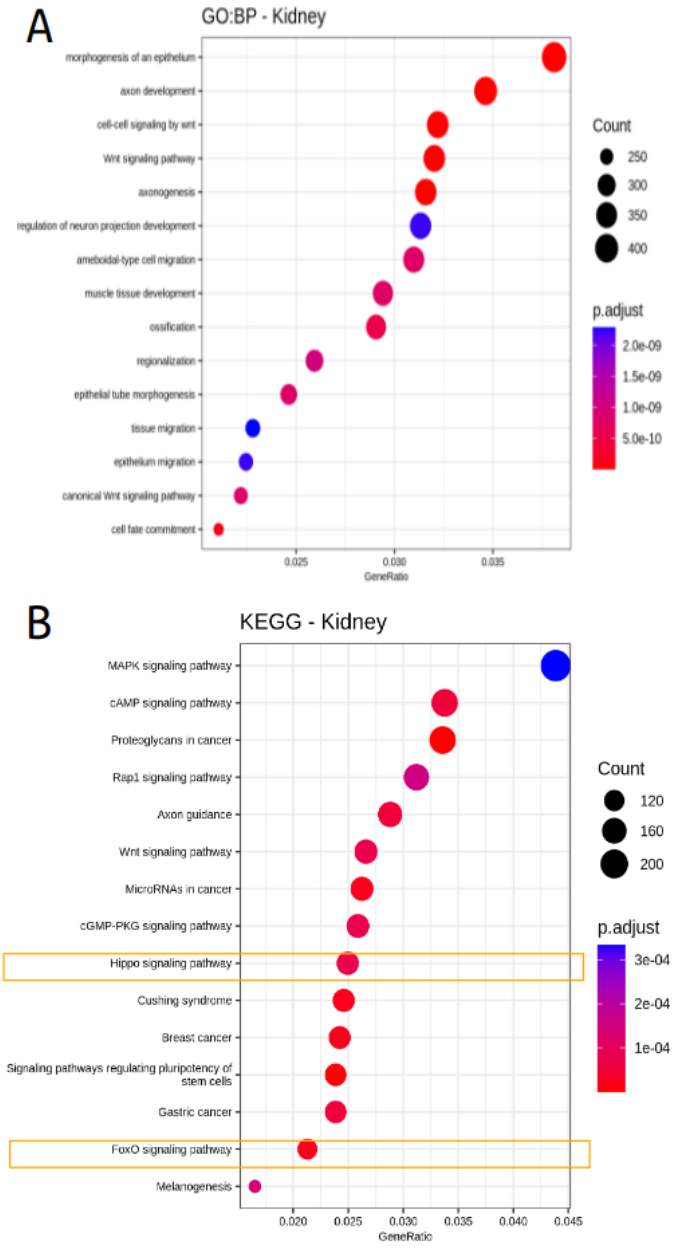
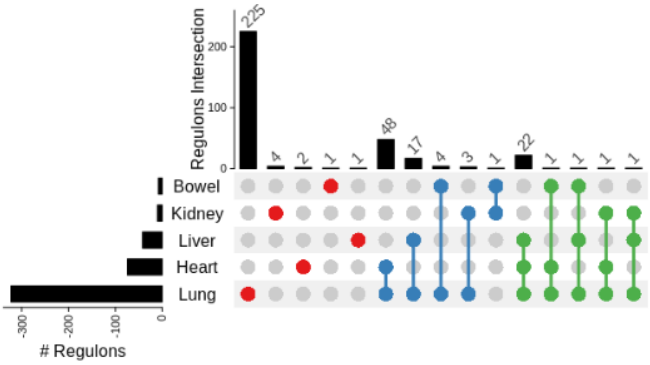


Figura 4.32: Dotplot del enriquecimiento de términos por ontología de genes, procesos biológicos (GO:BP, panel A) y por vías de señalización de KEGG (panel B) en los regulones sobre-activados del riñón con COVID-19. En el eje y está el top 15 de los términos que se encontraron más enriquecidos y en el x la proporción de genes etiquetados con ese término. El tamaño del punto indica la proporción de genes con ese término y el la escala de colores el p-valor ajustado. Los términos interesantes para el estudio presente están encerrados en cuadros.

4.4.2.2. Regulones sobre-activados y específicos por condición

Comparación de Regulones distintamente compartidos entre condiciones Como se esperaba, la respuesta ante la infección fue, a gran escala, más notoria en el pulmón, teniendo 225 regulones sobre-activados y específicos que no se compartieron en otros tejidos (véase la Figura 4.33 panel A), cuando en los otros tejidos se tuvieron 4 o menos regulones específicos. Qué regulones en concreto se comparten distintivamente se detallan, como anteriormente, en la Tabla complementaria mostrada en la Figura 4.33 panel B. De manera esperada, los tejidos que más comparten regulones sobre-activos son pulmón y corazón; pero de manera interesante, el hígado, cuyos daños durante la COVID-19 se han investigado menos frecuentemente, comparte 22 regulones sobre-activados con estos dos tejidos, lo cuál podría explicar por qué es otro de los tejidos más afectados.

A



B

	Bowel COVID19	Heart COVID19	Kidney COVID19	Liver COVID19	Lung COVID19
Bowel_COVID19	ZNF726	NA	ZNF782	NA	THRB,ZNF426,ZIC5,TBX4
Heart_COVID19	NA	HMX3,SNAI3	NA	NA	MECP2,OSR1,NKX2-4,GTF2A1,NKX3-1,EGR3,MYBL1,ZNF813,RBPJ,ESR1,HDAC2,HOXB2,ELF3,ZNF429,ZBTB33,MTHFD1,TFDP1,PAX9,CREBL2,JUNB,TEAD2,NFIL3,SNAI1,SOX4,JUND,BHLHE40,CEBPB,MYBL2,SOX11,TRPS1,ELK4,NFYA,CLOCK,EHF,ELF4,ZBTB7A,TFE3,POLR2A,SREBF2,ZNF785,ELK3,KLF3,MSX1,E2F4,CEBPZ,IRF2,BDP1,FOXJ2
Kidney_COVID19	ZNF782	NA	ZNF420,NFAT5,ZNF625,RBAK		HOXB6,SMC3,ID1
Liver_COVID19	NA	NA	NA	KLF14	HOXB4,PIR,SOX6,TFAP4,TMEM33,HES1,POLR3A,BATF,TFAP2C,CSNK2B,HIVEP3,NR1I2,MLXIP,PLG,POLR3G,BCL6B,ZNF671
Lung_COVID19	THRB,ZNF426,ZIC5,TBX4	MECP2,OSR1,NKX2-4,GTF2A1,NKX3-1,EGR3,MYBL1,ZNF813,RBPJ,ESR1,HDAC2,HOXB2,ELF3,ZNF429,ZBTB33,MTHFD1,TFDP1,PAX9,CREBL2,JUNB,TEAD2,NFIL3,SNAI1,SOX4,JUND,BHLHE40,CEBPB,MYBL2,SOX11,TRPS1,ELK4,NFYA,CLOCK,EHF,ELF4,ZBTB7A,TFE3,POLR2A,SREBF2,ZNF785,ELK3,KLF3,MSX1,E2F4,CEBPZ,IRF2,BDP1,FOXJ2	HOXB6,SMC3,ID1	HOXB4,PIR,SOX6,TFAP4,TMEM33,HES1,POLR3A,BATF,TFAP2C,CSNK2B,HIVEP3,NR1I2,MLXIP,PLG,POLR3G,BCL6B,ZNF671	HLX,EGR1,JDP2,THRA,TEF,ZNF460,MAFG,FOXO1,UBTF,ZNF358,JUN,IRF9,SIN3A,ZNF467,ZNF569,EP300,RBBP5,KLF6,MAFB,BRF2,KLF5,ETS1,ZNF91,GLIS2,CREB3L2,CHD2,NFE2L2,CREB3L1,KLF12,KLF4,NR3C1,MECOM,CREB1,SRF,MAFK,ZNF524,CREB3,ZBTB7B,ETS2,NR2F2,BCL3,KLF1,FOXN2,RFXAP,GRHPR,ATF7,SP2,STAT3,SP1,HOXA7,HIVEP2,IRF4,SMARCC2,ETV4,ATF3,STAT5B,BCLAF1,HINFP,GTF2B,HSF1,BACH1,IRF8,SPI1,GMEB1,ARID3A,SMAD5,STAT1,ATF1,RFX7,NR1D1,FOXM1,HSF4,MAFF,NHLH1,ETV3,FOXK2,MSC,ERG,ELF1,NFKB2,MXD1,NR2F6,PPARGC1A,ZNF513,FLI1,ZNF134,KDM4B,FOXF2,MAX,PKNOX1,HEYL,FOXC1,KLF17,MAZ,GATA2,RREB1,EZR,IKZF1,FOXF1,ZNF878,REST,ZBTB14,RELA,FOSL1,RUNX1,NKX1-2,NFYB,ZNF148,NR2E3,ZNF444,RARA,KLF2,HEY1,TRIM28,ESRRA,E2F1,ZNF579,BATF3,TCF7L2,AR,PRRX2,BCL6,EOMES,HNF1B,HOXC8,ERF,NR2C2,TAF1,FOXQ1,SP3,TEAD1,SIX5,NFKB1,SMAD3,GRHL2,BACH2,SREBF1,RFX2,ARX,ZNF653,ELF2,E2F3,TFEC,POU2F1,HOXB7,BHLHE41,FOXP3,SP4,GATA6,HOXB8,STAT2,SOX7,BCL11B,PATZ1,NR1D2,RFX1,CEBPA,ZFX,RUNX2,RXRB,ATF4,SOX10,SALL2,SMAD6,TBX22,PPARG,ZNF436,ATF5,NR3C2,ONECUT1,FOSB,STUB1,ZXDA,UBP1,E2F2,ZFP64,POU1F1,KLF7,ZHX1,NEUROD1,NFATC3,MEIS2,RUNX3,VSX2,ALX3,POU3F2,HAND1,MEIS1,KLF16,CCDC25,NANOG,NANOGP8,PLAGL2,SIX1,FOXA1,PSMC2,BARX2,ZNF282,VEZF1,FOXA3,NANOS1,SOX17,FOXJ3,ZNF48,OTP,DMBX1,ZNF316,EN1,SOX5,KLF11,TCF7L1,SOX9,ONECUT3,E2F7,KLF8,FOXC2,MLXIPL,SMAD1,SOX1,POU2AF1,GCM1,ATF6B,MYCN,FOXB2,ZNF35

Figura 4.33: Upset Plot y tabla complementaria de los regulones sobre-activados y más específicos distintamente compartidos entre condiciones en el análisis de tejidos COVID-19. A) A la izquierda se enlistan los tejidos con COVID-19, a su izquierda en barras se muestra la cantidad total de regulones sobre-activados en el respectivo tejido. En la parte superior en un gráfico de barras se muestra la cantidad de regulones compartidos. B) Los regulones que sólo se encuentran en un tejido dado se muestran en la diagonal de la tabla. Los regulones compartidos entre tejidos distintos están resaltados en color naranja y los tejidos que no tuvieron regulones distintamente compartidos se indican por un "NA".

La diferencia en el perfil de activación de regulones reclutados ante la infección y que fueran específicos por tejido es evidente cuando se comparan estos regulones entre tejido sano y enfermo como se muestra en la Figura 4.34. Al hacer un acercamiento y excluir los regulones sobre-activados y específicos de pulmón, podemos ver los 22 regulones que se comparten entre pulmón, corazón e hígado. Estos corresponden a los regulones guiados por los TFs NFE2L1, ZFHX3, YY2, MXD4, FOXK1, FOXN1, ZNF787, IRF1, TAF7, SMARCC1, ENO1, FOSL2, CEBPD, ETV6,YBX1, YY1, KLF9, RAD21, IRX3, KDM5A, MIOS, JAZF1. También destaca un regulon que sólo se encuentra sobre-activado y específico para hígado: KLF14. El KLF14 es conocido por su rol en la regulación negativa del receptor tipo II del TGFβ (Truty et al. (61)). De manera interesante, se ha visto que *KLF14* tiene un rol de recuperación al daño de tejido hepático mediado por sistema inmune al inhibir citocinas inflamatorias como IL-6 y TNF-α (Chen et al. (9)).

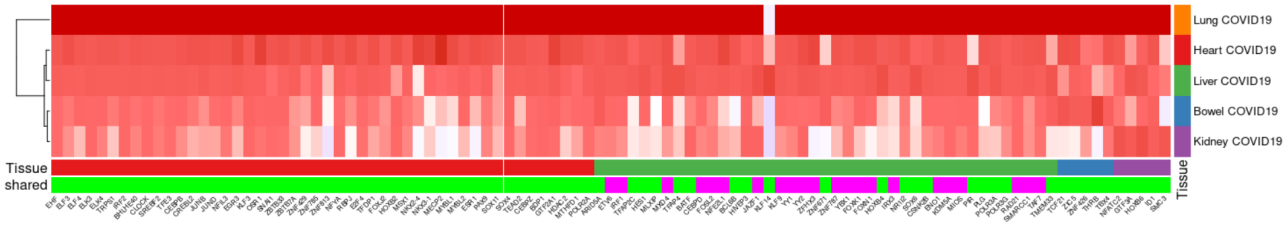


Figura 4.34: Heatmap clusterizado mostrando la especificidad de regulones por experimento o condición (RSS) de los regulones sobre-activados y que además eran más específicos por tejido excluyendo aquellos de pulmón. Los regulones son los que se indican en la Tabla 3.12. En el eje x están los regulones sobre-activados y más específicos en las muestras infectadas, indicados por el nombre de los factores transcripcionales guía de cada uno; mientras que el eje y está cada condición de tejido sano o enfermo. En la barra lateral vertical y en la primera barra horizontal se indican por colores los distintos tejidos. En la segunda barra horizontal se indica si el regulon dado se comparte entre pulmón, corazón e hígado. La métrica RSS está en el rango de valores [0,0.4].

Búsqueda de la función biológica de los regulones sobre-activados por SARS-CoV-2 y específicos por condición En las Figuras 4.35, 4.36 y 4.37 se muestran los términos biológicos enriquecidos en los regulones sobre-activados y específicos por tejido con COVID-19. No se encontraron términos para pulmón, lo cuál posiblemente es por estadística, es decir, había demasiados genes en los 225 regulones como para encontrar la sobre-representación de algún término.

En corazón, hígado y riñón se encontraron términos referentes a la migración, proliferación del tejido o incluso a la diferenciación, lo cuál puede deberse a una respuesta reparativa del tejido ante la infección. También se enriquecieron términos directamente

referentes a la infección, como la respuesta a infección de patógenos, la endocitosis y la proteólisis. Por último, en el intestino sólo se encontró un término enriquecido: la secreción de insulina.

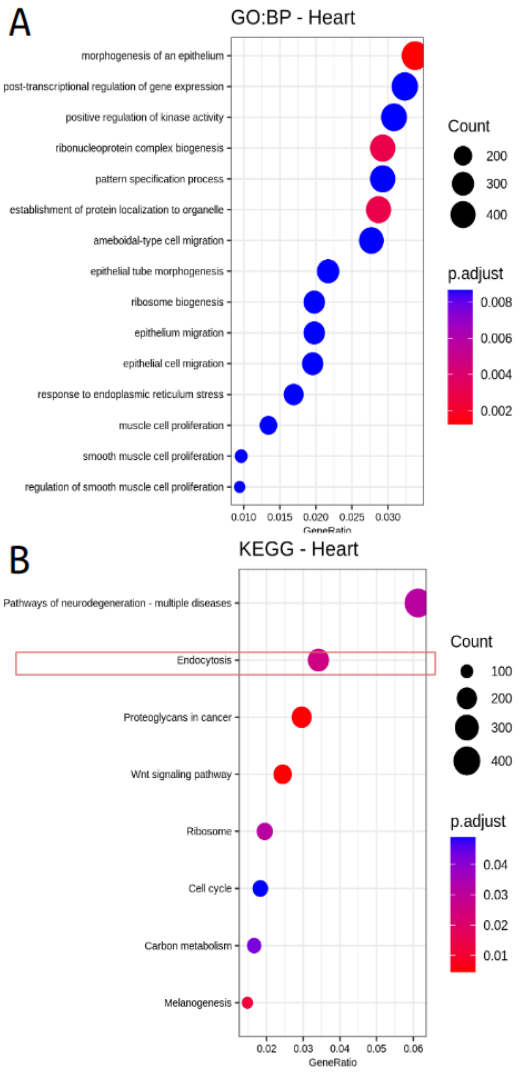


Figura 4.35: Dotplot del enriquecimiento de términos por ontología de genes, procesos biológicos (GO:BP, panel A) y por vías de señalización de KEGG (panel B) en los regulones sobre-activados y más específicos en Corazón con COVID-19. En el eje y está el top 15 de los términos que se encontraron más enriquecidos y en el x la proporción de genes etiquetados con ese término. El tamaño del punto indica la proporción de genes con ese término y el la escala de colores el p-valor ajustado.

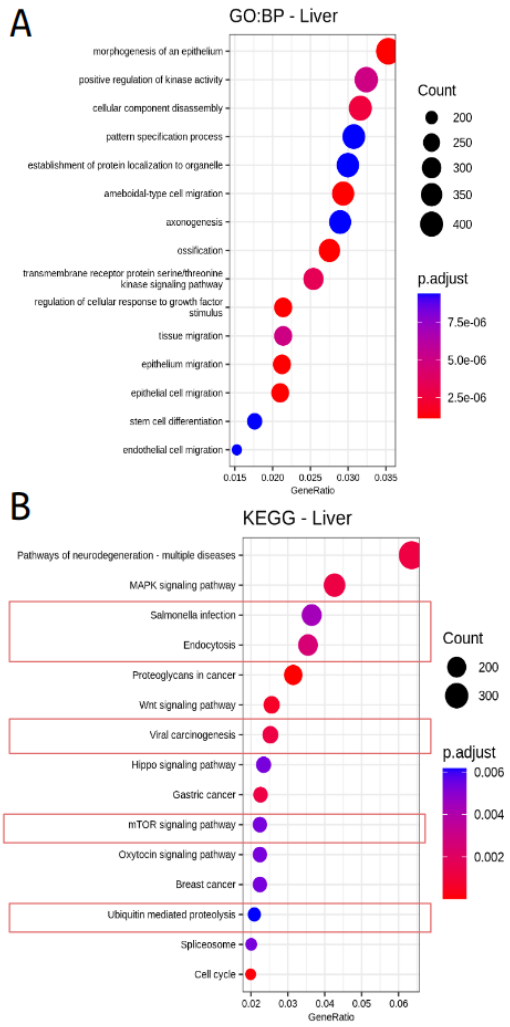


Figura 4.36: Dotplot del enriquecimiento de términos por ontología de genes, procesos biológicos (GO:BP, panel A) y por vías de señalización de KEGG (panel B) en los regulones sobre-activados y más específicos en Hígado con COVID-19. En el eje y está el top 15 de los términos que se encontraron más enriquecidos y en el x la proporción de genes etiquetados con ese término. El tamaño del punto indica la proporción de genes con ese término y la escala de colores el p-valor ajustado.

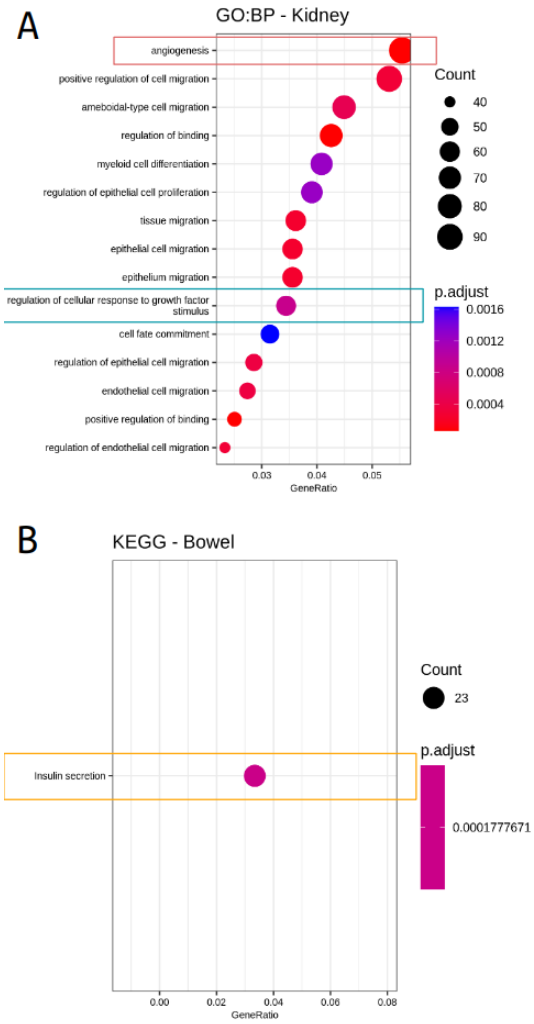


Figura 4.37: Dotplot del enriquecimiento de términos por ontología de genes, procesos biológicos (GO:BP, panel A) en Riñón COVID-19 y por vías de señalización de KEGG (panel B) en Intestino COVID-19 en sus respectivos regulones sobre-activados y más específicos. En el eje y está el top 15 de los términos que se encontraron más enriquecidos y en el x la proporción de genes etiquetados con ese término. El tamaño del punto indica la proporción de genes con ese término y la escala de colores el p-valor ajustado.

4.4.3. Análisis: Pulmón con COVID-19 en comparación con sano

4.4.3.1. Regulones sobre-activados

En el siguiente heatmap (Figura 4.38) se muestran los regulones sobre-activados usando el filtro de selección indicado en la sección 3.10, para pulmón COVID-19 con datos provenientes de 3 estudios independientes: HTATIP2, TP53, IRX6 y NANOG.

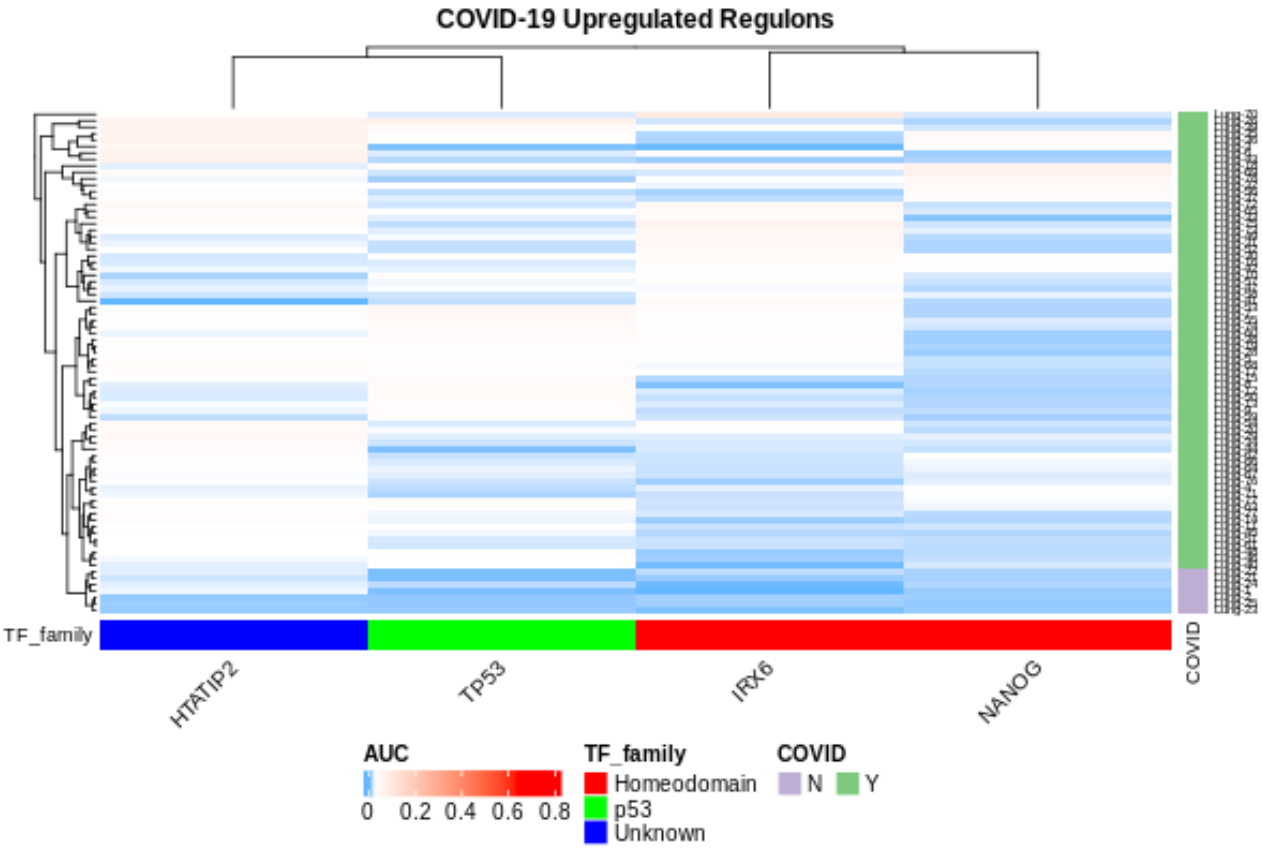


Figura 4.38: Heatmap clusterizado mostrando la activación de regulones (métrica AUC) de los regulones sobre-activados en el Pulmón con datos de tres estudios independientes. En el eje x están los regulones sobre-activados en las muestras de pulmón de pacientes con COVID-19, indicados por el nombre de los factores transcripcionales guía de cada uno; mientras que el eje y están las muestras de pulmón sano (gris en la barra lateral) o enfermo (verde). En la barra lateral horizontal se indica por colores la familia de factores transcripcionales a la que pertenece el TF guía del regulon. La leyenda indica las correspondencias de color y familia. La métrica AUC está en el rango de valores [0,0.8].

Estos regulones tuvieron los términos biológicos enriquecidos mostrados en la Figura 4.39. Vemos que uno de los términos enriquecidos es desarrollo de células epiteliales, lo cuál podría ser una respuesta de reparación del tejido dañado por COVID-19.

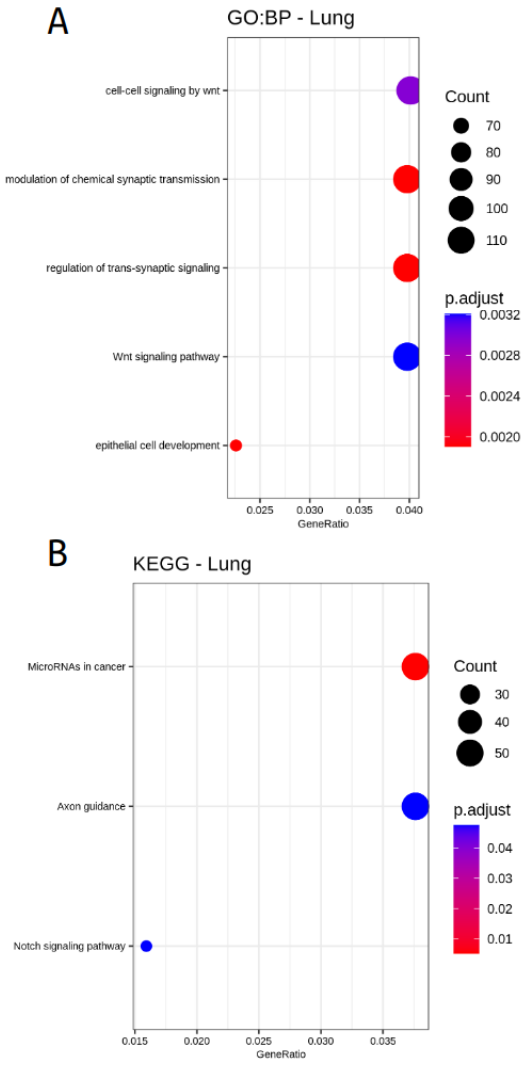


Figura 4.39: Dotplot del enriquecimiento de términos por ontología de genes, procesos biológicos (GO:BP, panel A) y por vías de señalización de KEGG (panel B) en los regulones sobre-activados y más específicos en el análisis de Pulmón con COVID-19. En el eje y están los términos que se encontraron enriquecidos y en el x la proporción de genes etiquetados con ese término. El tamaño del punto indica la proporción de genes con ese término y el la escala de colores el p-valor ajustado.

4.4.3.2. Regulones sobre-activados y específicos por condición

En la Figura 4.40 se muestran un heatmap de los regulones sobre-activados y específicos a COVID-19 utilizando un filtro de selección de genes menos estricto de FDR

(0.01), donde podemos ver la clara diferencia entre Pulmón COVID y sano, además, varios de los TFs guía de los regulones son catalogados con función inmunológica aquí (resaltados en rosa fuerte) y se comparten con los que se encontraron como relevantes en el análisis de tejidos (resaltados en rosa bajo).

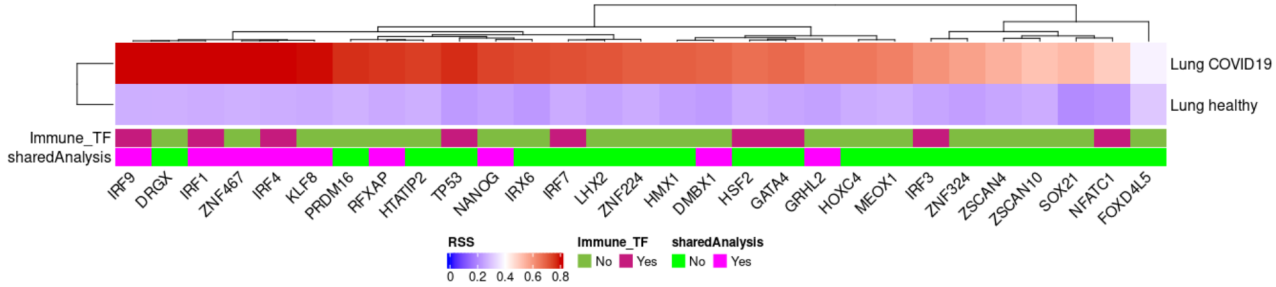


Figura 4.40: Heatmap clusterizado mostrando la especificidad de regulones por experimento o condición (RSS) de los regulones sobre-activados y más específicos en el Pulmón con datos de tres estudios independientes. En el eje x están los regulones sobre-activados y más específicos en las muestras de COVID-19, indicados por el nombre de los factores transcripcionales guía de cada uno; mientras que el eje y están las condiciones de pulmón sano o enfermo. En la primera barra horizontal se indica si el TF guía dado es catalogado aquí con función inmunológica (rosa) o no (verde). En la segunda barra horizontal se indica si el regulon dado se comparte en este análisis y en el análisis de tejidos (véase sección 4.4.2) en pulmón. La métrica RSS está en el rango de valores [0,0.8].

4.5. Contribuciones a RSAT

Como se detalló en la sección 3.8, hubo dos contribuciones principales en colaboración para la plataforma de RSAT. A continuación se encuentra anexado el pdf del artículo publicado y se menciona donde se encuentran las contribuciones generadas como trabajo paralelo en la tesis presente.

1. Se desarrolló el programa $network - interactions$ para línea de comandos y consulta web, que se presentan como resultados directos. Véase la sección *"Prediction of TF-gene interactions to build and refine gene regulatory networks"* y la última columna del diagrama de la Figura 2 del artículo 4.5.

2. Se revisó el funcionamiento y se contribuyó en el proceso de corrección de código de la REST-API web y en python. Véase la sección *"Programmatic REST API access"* del artículo 4.5.

W670–W676 *Nucleic Acids Research, 2022, Vol. 50, Web Server issue* *Published online 11 May 2022*
https://doi.org/10.1093/nar/gkac312

RSAT 2022: regulatory sequence analysis tools

Walter Santana-Garcia[1,†], Jaime A. Castro-Mondragon[2,†], Mónica Padilla-Gálvez[3], Nga Thi Thuy Nguyen[1], Ana Elizondo-Salas[3], Najla Ksouri[4], François Gerbes[5], Denis Thieffry[1], Pierre Vincens[1], Bruno Contreras-Moreira[4,*], Jacques van Helden[5,6,*], Morgane Thomas-Chollier[1,*] and Alejandra Medina-Rivera[3,*]

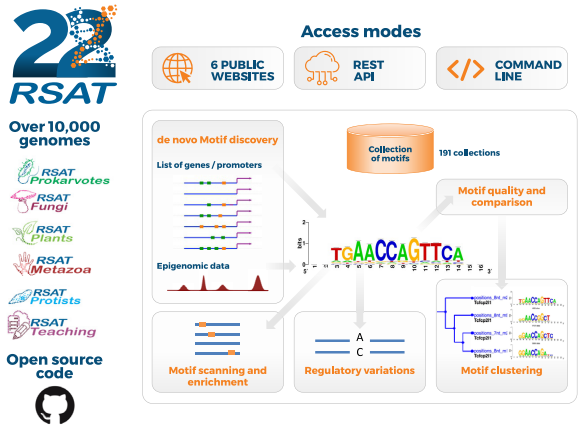
[1]Institut de biologie de l'Ecole normale supérieure (IBENS), Ecole normale supérieure, CNRS, INSERM, PSL Université Paris, 75005 Paris, France, [2]Centre for Molecular Medicine Norway (NCMM), Nordic EMBL Partnership, University of Oslo, 0318 Oslo, Norway, [3]Laboratorio Internacional de Investigación sobre el Genoma Humano, Universidad Nacional Autónoma de México, Campus Juriquilla, Blvd Juriquilla 3001, 76230 Santiago de Querétaro, México, [4]Estación Experimental de Aula Dei-CSIC, 50059 Zaragoza, Spain, [5]CNRS, Institut Français de Bioinformatique, IFB-core, UMS 3601, Evry, France and [6]Aix-Marseille Univ, INSERM UMR_S 1090, Lab Theory and Approaches of Genome Complexity (TAGC), F-13288 Marseille, France

Received March 03, 2022; Revised April 12, 2022; Editorial Decision April 13, 2022; Accepted April 20, 2022

ABSTRACT

RSAT (Regulatory Sequence Analysis Tools) enables the detection and the analysis of *cis*-regulatory elements in genomic sequences. This software suite performs (i) *de novo* motif discovery (including from genome-wide datasets like ChIP-seq/ATAC-seq) (ii) genomic sequences scanning with known motifs, (iii) motif analysis (quality assessment, comparisons and clustering), (iv) analysis of regulatory variations and (v) comparative genomics. RSAT comprises 50 tools. Six public Web servers (including a teaching server) are offered to meet the needs of different biological communities. RSAT philosophy and originality are: (i) a multi-modal access depending on the user needs, through web forms, command-line for local installation and programmatic web services, (ii) a support for virtually any genome (animals, bacteria, plants, totalizing over 10 000 genomes directly accessible). Since the 2018 NAR Web Software Issue, we have developed a large REST API, extended the support for additional genomes and external motif collections, enhanced some tools and Web forms, and developed a novel tool that builds or refine gene regulatory networks using motif scanning (network-interactions). The RSAT website provides extensive documentation, tutorials and published protocols. RSAT code is under open-source license and now hosted in GitHub. RSAT is available at http://www.rsat.eu/.

GRAPHICAL ABSTRACT



INTRODUCTION

The Regulatory Sequence Analysis Tools (RSAT) provides a wide range of bioinformatics programs enabling the analysis of genomic regulatory sequences in physiological and disease contexts. RSAT enables users to obtain genomic sequences and perform typical analyses, such as *de novo* motif discovery, or motif scanning to predict transcription factor (TF) binding sites (TFBSs). RSAT functionalities also include original analyses, such as motif quality evaluation,

*To whom correspondence should be addressed. Tel: +33 44 32 23 81; Email: mthomas@biologie.ens.fr
Correspondence may also be addressed to Alejandra Medina-Rivera. Tel: +52 55 5623 4331; Email: amedina@liigh.unam.mx
Correspondence may also be addressed to Jacques van Helden. Lab. Email: Jacques.van-Helden@univ-amu.fr
Correspondence may also be addressed to Bruno Contreras-Moreira. Tel: +34 976716089; Email: bcontreras@eead.csic.es
†The authors wish it to be known that, in their opinion, these authors should be regarded as Joint First Authors.

motif comparisons and clustering, detection and analysis of regulatory variants, building of control datasets and comparative genomics to discover motifs based on cross-species conservation. Altogether, the RSAT Web site gives access to 50 tools that can be used individually, or sequentially to perform more complex analyses. RSAT has been well-established since its initial development in 1998 (1,2). It has been regularly updated and extended with novel developments stimulated by advances in the field of regulatory genomics. We summarize here the main functionalities, and describe novelties since the previous NAR Web server issues (3–7).

RSAT FUNCTIONALITIES

RSAT tools have been individually described in the previous 2018 NAR update (3), with a historical perspective, as well as by applications (4). We summarize below the main functionalities ordered by data types to analyze, as a useful starting point for novice users (Figure 1). Pointers to the three use cases that exemplify how to combine the tools into routine analysis (3) are indicated.

Epigenomics datasets such as ChIP-seq or ATAC-seq peaks

Genome-wide datasets obtained from epigenomics experiments (e.g. ChIP-seq, ATAC-seq, ChIP-exo, DNaseI, Cut&Run, Cut&Tag) consists of genomic regions—known as peaks—that are likely bound by a given transcription factor (TF), or associated with open chromatin. The prevalent question is 'Which TF binding motifs can be detected in the peaks?'

The peaks can be analyzed with the user-friendly pipeline *peak-motifs* (5,8,9), which relies on *de novo* motif discovery to detect exceptional motifs in a set of sequences. *peak-motifs* runs multiple complementary algorithms [*oligo-analysis* (1), *dyad-analysis* (10), *position-analysis* (11) and *local-word-analysis* (8) that can all be used as independent tools], then compares the predicted motifs with annotated motif databases (*compare-matrices*), and finally predicts the positions of the putative transcription factor binding sites (TFBSs) within the peaks (*matrix-scan* (12) (Figure 2). Two datasets can be provided as input to enable differential analysis.

Alternatively, the peaks can be directly scanned with motifs (e.g. the discovered motifs, or from motif databases such as JASPAR (cf. 'Motifs represented as Position-Scoring Specific Matrices (PSSM) or consensus sequences')) to locate putative TFBSs (*dna-pattern* or *matrix-scan* (12)) or to predict potential cis-regulatory modules (*crer-scan* (3)). The tool *matrix-quality* can measure the enrichment of a specific motif within one or more peak datasets (13).

As input peaks must be provided as FASTA-formatted sequences, RSAT provides two tools to extract sequences from genome-wide peak datasets specified in BED-formated genomic coordinates (cf. 'Genomic coordinates as a BED file').

Control datasets can be built by selecting sequences at random positions from a given genome (*random-genome-fragments*), or by generating simulated sequences matching the size and composition of the peaks (*random-sequences*).

Lists of gene names or identifiers

Genome-wide datasets from transcriptomics experiments (e.g. microarrays, RNA-seq), as well as more targeted *in situ* hybridization experiments, typically results in a list of co-expressed genes. A frequent question is 'Which TFs may co-regulate the expression of these genes?' The typical analysis workflow consists in (i) retrieving sequences relative to these genes (e.g. promoter) and (ii) performing *de novo* motif discovery or motif scanning (cf. 'Epigenomics datasets such as ChIP-seq or ATAC-seq peaks'). Given a list of gene names or identifiers, *retrieve-sequences* extracts promoter sequences of locally-installed genomes, while *retrieve-ensembl-seq* (14) retrieves sequences of promoters or other specified features on-the-fly from Ensembl.

To support comparative genomics analyses, *retrieve-ensembl-seq* can also retrieve sequences from homologous genes. On the Plant server, the tool *get-orthologs-compara* additionally returns detailed information on homologous genes in a set of reference organisms, using precomputed Ensembl Compara data (15,16). On the Fungi and Prokaryotes servers, lists of orthologous genes can be obtained with *get-orthologs*. For the subsequent motif analysis step on these servers, *footprint-discovery* (17,18) and *footprint-scan* directly use cross-species conservation to detect putative regulatory signals in non-coding sequences (phylogenetic footprinting) (Figure 2).

Control datasets can be built by randomly selecting genes within a given genome with *random-gene-selection*. Use case 1 (3) combines *get-orthologs-compara, retrieve-sequences* and *matrix-scan* to predict TFBSs of VRN1 within the promoters of the FT1 gene in several plant genomes.

Motifs represented as Position-Scoring Specific Matrices (PSSM) or consensus sequences

Motifs represented as PSSMs or as consensus sequences may be obtained by *de novo* motif analysis, from databases such as JASPAR (19), or directly from the literature. Some typical questions are (i) 'Is the motif of good quality ?', (ii) ' Which sequences contain TFBS matching this motif ?', (iii) 'Does this motif resemble other motifs ?'.

First, *matrix-quality* (13) aims at assessing the quality of a PSSM on sequence datasets provided by the user, by comparing theoretical and empirical score distributions. Second, *matrix-scan* takes as input motifs to locate putative TFBSs in user-provided sequences (cf. 'Epigenomics datasets such as ChIP-seq or ATAC-seq peaks'). Third, *compare-matrices* compares two collections of matrices and returns various similarity statistics along with a PSSMs multi-pairwise alignment. *matrix-clustering* (20) regroups similar PSSMs into clusters, builds consensus PSSMs for each cluster and offers a dynamic visualization of aligned PSSMs. We applied *matrix-clustering* to regroup redundant matrices within and across motifs databases, in order to build the RSAT non-redundant motif collections for insects, plants and vertebrates (20). These collections are accessible with *retrieve-matrix* (3), which conveniently offers additional access to 187 external motifs collections, totalizing 454 524 motifs, all homogenized in TRANSFAC format (Supplementary Table S1). These collections include large databases such as JASPAR (19) and FootprintDB (21), as

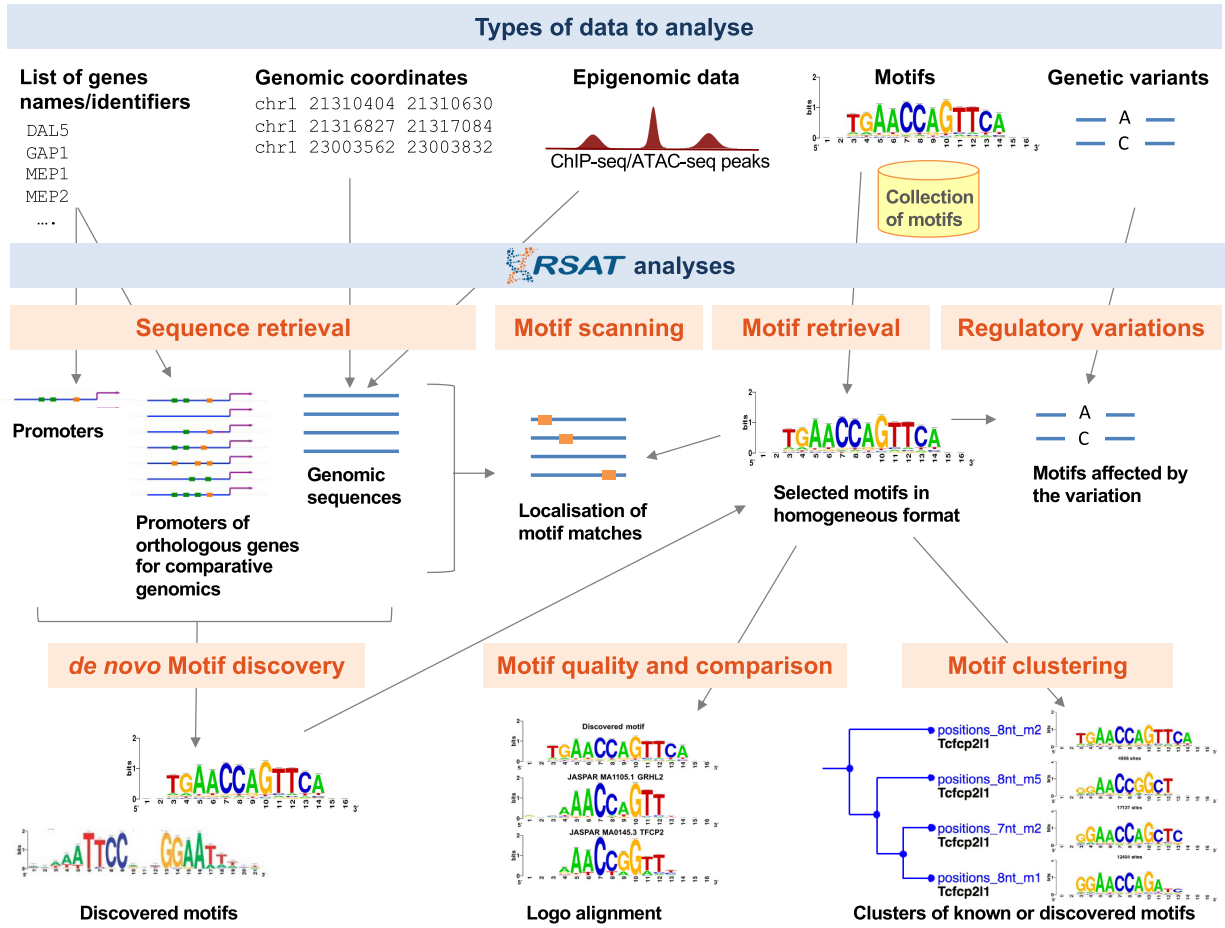


Figure 1. Overview of the main applications of RSAT, with associated input data types.

well as more specific ones such as ANISEED (22), RegulonDB (23) or RNA binding motifs, covering all kingdoms (Metazoa, Prokaryotes, Fungi, Plants). JASPAR (19) provides matrix-clustering results for each release, to provide information on the redundancy of motifs (https://jaspar.genereg.net/matrix-clusters/).

As there is no standard format for the PSSMs files, the tool *convert-matrix* performs interconversion between multiple motifs formats, and generates graphical representations of motifs in the form of logos. This allows users to focus on their scientific questions rather than formatting issues.

Control datasets can be built by generating permuted versions of PSSMs with *permute-matrix* or simulated matrix with *random-motif*.

Genomic coordinates as a BED file

Lists of features (e.g. peaks, predicted TFBSs) with their genomic coordinates are conventionally encoded in BED-formatted files (or GFF/GTF). The usual question is 'How to identify TFBSs within these regions?' The first step is to extract the corresponding genomic sequences; we provide user friendly tools with web interfaces to facilitate this task. Sequences can be automatically extracted from the UCSC genome browser with *fetch-sequences-from-UCSC* (3) or from locally -installed genomes with *sequences-from-BED/GFF/VCF*, which internally uses BEDTools and supports repeat-masking (24). Use case 2 (3) combines *retrieve-matrix*, *matrix-clustering*, *sequences-from-BED/GFF/VCF* and *matrix-scan* to generate a non-redundant AP1 motif from multiple annotated motifs, and predict TFBSs of AP1 within ChIP-seq peaks.

Lists of genetic variants as VCF files

Lists of genetic variants (SNPs, indels) can be retrieved from Genome-wide Association Studies (GWAS) and from databases such as Ensembl. A standard question is 'Which non-coding variants are affecting TF binding on cis-regulatory elements?' RSAT provides *variation-tools* (25), a series of programs to obtain information on individual variants, extract their flanking sequences, scan these flanking sequences with motif collections and predict which variants may affect TF binding.

Control datasets can be built by generating permuted versions of PSSMs with *permute-matrix*. Use case 3 (3) combines *convert-variations, retrieve-variation-seq* and *variation-*

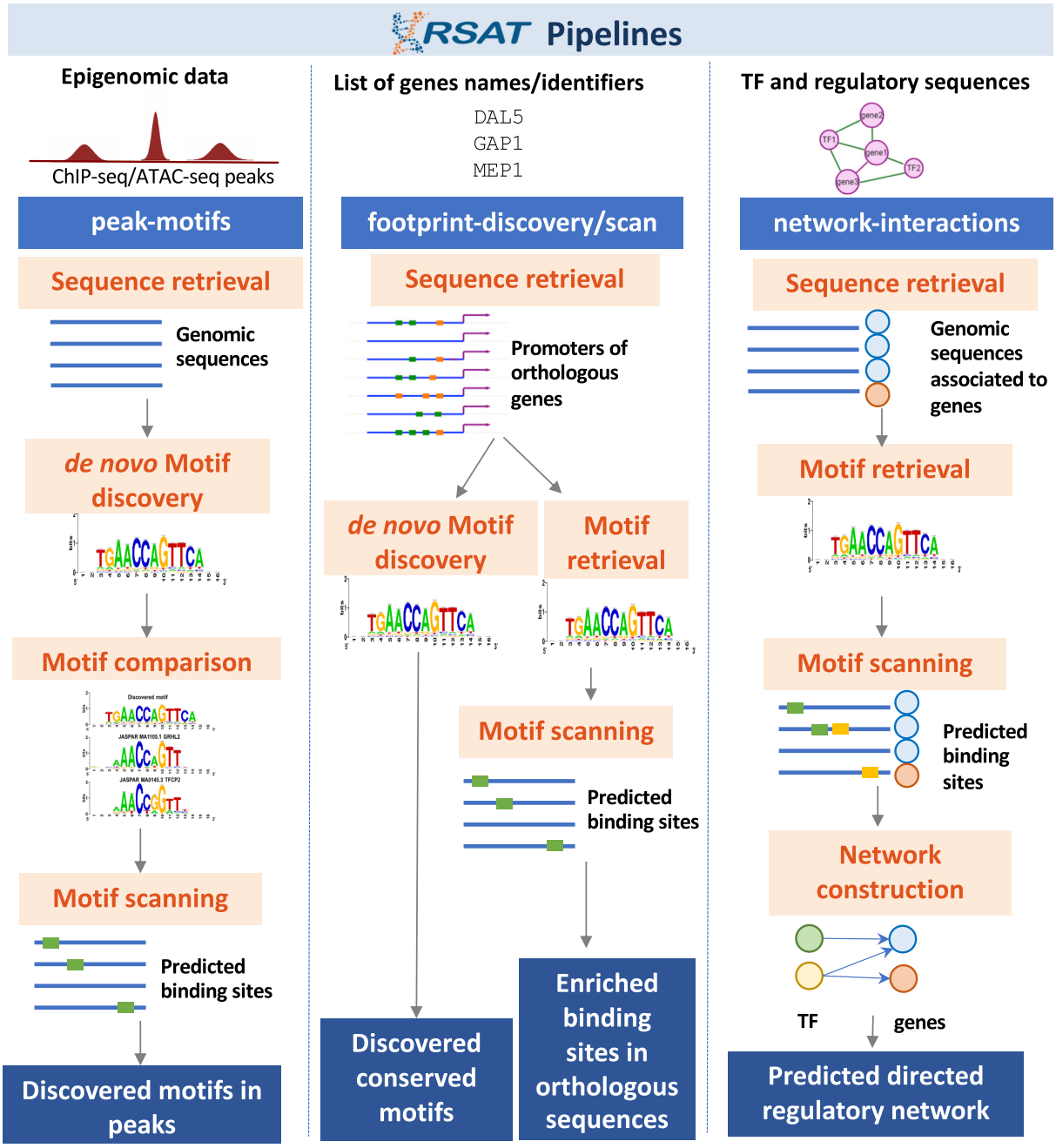


Figure 2. Three pipelines offering pre-defined combinations of RSAT tools (*peak-motifs*, *footprint-scan* and *footprint-discovery*, *network-interactions*).

scan on a VCF-formatted file specifying allelic variants detected in melanoma. It illustrates how scanning the surrounding sequences of the variants with the AP1 motif enables the identification of potential regulatory variants affecting AP1 binding.

RSAT 2022 NOVELTIES

RSAT locally installed organisms and motif collections

Since the last NAR Web server issue, we have further extended the number of supported organisms on the public servers, notably for Plants (+25 genomes) and Prokaryotes (+195 genomes). Some organisms were installed upon user request. As of February 2022, RSAT public servers support 10 076 locally installed genomes, including 9 646 Prokaryotes, 245 Fungi, 186 Protists, 91 Metazoa and 93 Plants. Besides, we have extended the number of external motif databases directly accessible in the common TRANSFAC format, from 50 to 187 external databases (cf. 'Motifs represented as Position-Scoring Specific Matrices (PSSM) or consensus sequences') (Supplementary Table S1). Some motif collections were added upon user request. Adding

new collections can now be made directly by a pull request on GitHub. All collections are freely downloadable to be used independently of RSAT (https://github.com/rsa-tools/motif_databases).

Users genome installation requests for servers are welcomed. In order to get a genome installed users have to contact the RSAT team through email 'rsat-contact@list01.biologie.ens.fr' with the information of the requested genome: organism name, genome version, source (*i.e.* NCBI, ENSEMBL) and url link to the genome data. In the case of motif collections, users can also request additions by providing: name, data, URL link and version information.

Furthermore, interested users can install genomes locally in their own RSAT instances. The documentation at https://rsa-tools.github.io/managing-RSAT contains detailed manuals to install genomes from different sources, such as RSAT servers, Ensembl, NCBI and from original FASTA and GTF data files.

Programmatic REST API access

Our programmatic SOAP/WSDL access is being replaced by the increasingly popular Web service REST API. It provides access to a large set of 49 tools of the RSAT suite. The REST API has been developed with the flask library; its documentation is generated with Swagger UI. Example clients in Python have been written to further help users using this API.

Updated web interface and tools

Some tools are highly parameterisable, thereby complexifying the corresponding Web forms. We have started to redesign these forms to simplify usage: we are now better separating the mandatory inputs/parameters from the optional ones (see *retrieve-sequence*, *matrix-clustering* and *network-interactions*). Several tools have been updated with additional functionalities or increased efficiency. This is the case of *variation-tools* (cf. 'Lists of genetic variants as VCF files'), for which haplotype scanning has been improved to assess the regulatory effect in TFBSs of haplotypes with large number of variants (SNPs and indels) in Metazoa and Plants.

Prediction of TF-gene interactions to build and refine gene regulatory networks

Many efforts have been made to infer gene regulatory networks (GRN) from transcriptomic data, with approaches based on coexpression, orthology or sequence motifs (26), but there is no consensus on a single best method. To further improve the inferred GRNs, it is common to apply motif scanning (pattern-matching) as a second step upon inferred interactions. We introduce *network-interactions*, a new user-friendly GRN reconstruction pipeline based on pattern-matching, which can help refine GRNs generated by other tools (Figure 2). It takes as input two lists: (i) the TFs of interests specified as a list of TF names and (ii) a list of genomic regions associated with gene names (typically promoter/enhancer regions of genes) provided as BED coordinates. A seed network, previously generated from other tools (i.e. based on co-expression), can optionally be provided. *network-interactions* runs *matrix-scan* using one of the motif collections available in RSAT (default is JASPAR's 2022 vertebrates motif collection (19)) to predict TF-gene interactions. *network-interactions* thereby generates several networks: (i) a complete network for all TF-gene interactions, (ii) another network focusing on TF-TF interactions, (iii) one with 3-step TFs indirect interactions (TF-TF-gene) and (iv) when provided with an input GRN, the overlap and the complements between the input network and the network generated by *network-interactions*, where the overlap includes the putative TF binding information. This novel tool extends RSAT's suite and offers a straightforward and flexible method to expand and refine GRNs.

RSAT source code on GitHub and Docker container

The RSAT source code, under AGPL-3.0 open-source license, has been transferred to GitHub, to stimulate community-wise participation in its development: https://github.com/rsa-tools. Additional RSAT documentation is available there as well. A Docker container has been built to analyze the promoters of coexpressed genes in plants (27): https://github.com/eead-csic-compbio/coexpression_motif_discovery.

Learning to use RSAT

In addition to the above-mentioned use cases, RSAT provides extensive documentation, tutorials and published protocols (4). To target non-expert users, including biologists and biomedical practitioners, the main tools are accessible through web forms with DEMO buttons and tutorials. The latest protocols (28,29) and application (27) focuses on motif discovery in plant genomes; the described approaches can generally be applied to other organisms. Most of our previously published protocols (9,12,30) are still relevant to learn about the underlying algorithms, choosing the relevant parameters and interpreting the results, despite updates in the Web interfaces. Users may also contact us via email or via our Twitter account @RSATools.

CONCLUSIONS

Compared to alternative programs, RSAT is unique for its wide range of functionalities, extensive motifs collections and >10 000 supported organisms from all kingdoms. The main alternatives are the MEME suite (31), which mainly focuses on motif analyses, and HOMER (32), which primarily focuses on motif discovery. Deep-learning methods are more focused in discovering context-specific TFBS, whereas RSAT aims at providing a complete environment for motif analysis. We aim for RSAT to be usable in combination with other programs (including MEME and HOMER); RSAT thus offers several file format conversion utility tools (*convert-matrix*, *convert-background-models*, *convert-features*, ...). After 20 years of existence, RSAT remains one of the most used tools in regulatory genomics. Looking forward, we aim at (i) continuing to enhance the suite in particular to cope with the challenges

posed by single cell technologies in terms of data analysis efficiency, and (ii) continuing to ensure long-term maintenance, with packaging in conda, a non-plant docker container and continuous integration on GitHub.

DATA AVAILABILITY

RSAT public servers are accessible from the RSAT portal at http://www.rsat.eu/. RSAT Web servers can be freely accessed by all users without login requirement. For bioinformatician users, RSAT is accessible (i) as a command-line suite for installation on a local server or on a computer cloud, from its source code https://github.com/rsa-tools, or (ii) via the REST API web programmatic access. RSAT is part of the Service Delivery Plan of the Elixir-France node (European distributed infrastructure for life-science information): https://elixir-europe.org/services/list?field_scientific_domain_tid=All&field_elixir_badge_tid=All&field_type_of_service_tid=All&field_elixir_node_target_id=981&combine=.

RSAT code and documentation is available through GitHub https://github.com/rsa-tools. The Docker container for plants is located at: https://github.com/eead-csic-compbio/coexpression_motif_discovery. Motif collections can be found at https://github.com/rsa-tools/motif_databases.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

RSAT is managed by an international team (France, Mexico, Spain, Norway). We are particularly thankful to the colleagues who help us installing and maintaining RSAT servers: Victor del Moral Chavez, Alfredo José Hernández Alvarez (Centro de Ciencias Genomicas, Cuernavaca, Mexico), Luis Alberto Aguilar Bautista and Jair Garcia Sotelo (Laboratorio Nacional de Visualización Científica Avanzada, Mexico), Aurora Martín (Estación Experimental de Aula dei, Zaragoza, Spain), along with the ABIMS platform in Roscoff, France and the mutualized task force of the Institut Français de Bioinformatique. We also thank Olivier Sand and Matthieu Defrance for regularly answering RSAT-related questions. We thank Ieva Rauluseviciute for finding and reporting bugs in RSAT programs. We thank Mauricio Guzman for designing all logos for RSAT and styling the figures. The testing squad of LIIGH trainees provided tremendous help: Ana Laura Hernández-Ledesma, Juan Manuel Martinez-Villalobos, Paula R. Reyes-Pérez. We especially acknowledge Julio Collado-Vides, who impulsed the project and supported it during the last 25 years.

FUNDING

Institut Universitaire de France (to M.T.C., N.T.T.N.); ANR [ANR-14-CE11-0006-01 to M.T.C., N.T.T.N.]; ITMO Cancer [20CM114-00 to D.T., W.S.G.]; CONACYT [11311 to A.M.R., M.P.G.]; Universidad Nacional Autónoma de México PAPIIT (UNAM) [IA203021]; Spanish State Research Agency, EUROPEAN REGIONAL DEVELOPMENT FUND (FEDER) [AGL2017-83358-R to N.K.; AEI/FEDER, UE]; Gobierno de Aragón [A08_17R, A09_20R, Phd Contract to N.K.]; PRIMA [PCI2019-103526 to B.C.M.; Programación Conjunta Internacional, Programa Estatal de l+D+i Orientada a los Retos de la Sociedad]; Erasmus+ (to N.K.); Norwegian Research Council [288404 to J.A.C.M.]; Institut Français de Bioinformatique (IFB) [ANR-11-INBS-0013]; SEP-CONACYT-ECOS-ANUIES [291235 to A.M.R., M.T.C. and D.T.] . Funding for open access charge: Institut Universitaire de France.
Conflict of interest statement. None declared.

REFERENCES

1. van Helden,J., André,B. and Collado-Vides,J. (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.*, **281**, 827–842.
2. van Helden,J., André,B. and Collado-Vides,J. (2000) A web site for the computational analysis of yeast regulatory sequences. *Yeast Chichester Engl.*, **16**, 177–187
3. Nguyen,N.T.T., Contreras-Moreira,B., Castro-Mondragon,J.A., Santana-Garcia,W., Ossio,R., Robles-Espinoza,C.D., Bahin,M., Collombet,S., Vincens,P., Thieffry,D. *et al.* (2018) RSAT 2018: regulatory sequence analysis tools 20th anniversary. *Nucleic Acids Res.*, **46**, W209–W214.
4. Medina-Rivera,A., Defrance,M., Sand,O., Herrmann,C., Castro-Mondragon,J.A., Delerce,J., Jaeger,S., Blanchet,C., Vincens,P., Caron,C. *et al.* (2015) RSAT 2015: regulatory sequence analysis tools. *Nucleic Acids Res.*, **43**, W50–W56.
5. Thomas-Chollier,M., Defrance,M., Medina-Rivera,A., Sand,O., Herrmann,C., Thieffry,D. and van Helden,J. (2011) RSAT 2011: regulatory sequence analysis tools. *Nucleic Acids Res.*, **39**, W86–W91.
6. Thomas-Chollier,M., Sand,O., Turatsinze,J.-V., Janky,R., Defrance,M., Vervisch,E., Brohée,S. and van Helden,J. (2008) RSAT: regulatory sequence analysis tools. *Nucleic Acids Res.*, **36**, W119–W127.
7. van Helden,J. (2003) Regulatory sequence analysis tools. *Nucleic Acids Res.*, **31**, 3593–3596.
8. Thomas-Chollier,M., Herrmann,C., Defrance,M., Sand,O., Thieffry,D. and van Helden,J. (2012) RSAT peak-motifs: motif analysis in full-size chip-seq datasets. *Nucleic Acids Res.*, **40**, e31.
9. Thomas-Chollier,M., Darbo,E., Herrmann,C., Defrance,M., Thieffry,D. and van Helden,J. (2012) A complete workflow for the analysis of full-size chip-seq (and similar) data sets using peak-motifs. *Nat. Protoc.*, **7**, 1551–1568.
10. van Helden,J., Rios,A.F. and Collado-Vides,J. (2000) Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res.*, **28**, 1808–1818.
11. van Helden,J., del Olmo,M. and Pérez-Ortín,J.E. (2000) Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals. *Nucleic Acids Res.*, **28**, 1000–1010.
12. Turatsinze,J.-V., Thomas-Chollier,M., Defrance,M. and van Helden,J. (2008) Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nat. Protoc.*, **3**, 1578–1588.
13. Medina-Rivera,A., Abreu-Goodger,C., Thomas-Chollier,M., Salgado,H., Collado-Vides,J. and van Helden,J. (2011) Theoretical and empirical quality assessment of transcription factor-binding motifs. *Nucleic Acids Res.*, **39**, 808–824.
14. Sand,O., Thomas-Chollier,M. and van Helden,J. (2009) Retrieve-ensembl-seq: user-friendly and large-scale retrieval of single or multi-genome sequences from ensembl. *Bioinforma. Oxf. Engl.*, **25**, 2739–2740.
15. Herrero,J., Muffato,M., Beal,K., Fitzgerald,S., Gordon,L., Pignatelli,M., Vilella,A.J., Searle,S.M.J., Amode,R., Brent,S. *et al.* (2016) Ensembl comparative genomics resources. *Database J. Biol. Databases Curation*, **2016**, baw053.

16. Yates,A.D., Allen,J., Amode,R.M., Azov,A.G., Barba,M., Becerra,A., Bhai,J., Campbell,L.I., Carbajo Martinez,M., Chakiachvili,M. *et al.* (2022) Ensembl genomes 2022: an expanding genome resource for non-vertebrates. *Nucleic Acids Res.*, **50**, D996–D1003.
17. Janky,R. and van Helden,J. (2008) Evaluation of phylogenetic footprint discovery for predicting bacterial cis-regulatory elements and revealing their evolution. *BMC Bioinf.*, **9**, 37.
18. Brohée,S., Janky,R., Abdel-Sater,F., Vanderstocken,G., André,B. and van Helden,J. (2011) Unraveling networks of co-regulated genes on the sole basis of genome sequences. *Nucleic Acids Res.*, **39**, 6340–6358.
19. Castro-Mondragon,J.A., Riudavets-Puig,R., Rauluseviciute,I., Berhanu Lemma,R., Turchi,L., Blanc-Mathieu,R., Lucas,J., Boddie,P., Khan,A., Manosalva Pérez,N. *et al.* (2022) JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **50**, D165–D173.
20. Castro-Mondragon,J.A., Jaeger,S., Thieffry,D., Thomas-Chollier,M. and van Helden,J. (2017) RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections. *Nucleic Acids Res.*, **45**, e119.
21. Contreras-Moreira,B. and Sebastian,A. (2016) FootprintDB: analysis of plant cis-regulatory elements, transcription factors, and binding interfaces. *Methods Mol. Biol. Clifton NJ*, **1482**, 259–277.
22. Brozovic,M., Dantec,C., Dardaillon,J., Dauga,D., Faure,E., Gineste,M., Louis,A., Naville,M., Nitta,K.R., Piette,J. *et al.* (2018) ANISEED 2017: extending the integrated ascidian database to the exploration and evolutionary comparison of genome-scale datasets. *Nucleic Acids Res.*, **46**, D718–D725.
23. Santos-Zavaleta,A., Salgado,H., Gama-Castro,S., Sánchez-Pérez,M., Gómez-Romero,L., Ledezma-Tejeida,D., García-Sotelo,J.S., Alquicira-Hernández,K., Muñiz-Rascado,L.J., Peña-Loredo,P. *et al.* (2019) RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in e. coli K-12. *Nucleic Acids Res.*, **47**, D212–D220.
24. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
25. 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. *Comput. Struct. Biotechnol. J.*, **17**, 1415–1428.
26. Mercatelli,D., Scalambra,L., Triboli,L., Ray,F. and Giorgi,F.M. (2020) Gene regulatory network inference resources: a practical overview. *Biochim. Biophys. Acta BBA - Gene Regul. Mech.*, **1863**, 194430.
27. Ksouri,N., Castro-Mondragón,J.A., Montardit-Tarda,F., van Helden,J., Contreras-Moreira,B. and Gogorcena,Y. (2021) Tuning promoter boundaries improves regulatory motif discovery in nonmodel plants: the peach example. *Plant Physiol.*, **185**, 1242–1258.
28. Contreras-Moreira,B., Castro-Mondragon,J.A., Rioualen,C., Cantalapiedra,C.P. and van Helden,J. (2016) RSAT::Plants: motif discovery within clusters of upstream sequences in plant genomes. *Methods Mol. Biol. Clifton NJ*, **1482**, 279–295.
29. Castro-Mondragon,J.A., Rioualen,C., Contreras-Moreira,B. and van Helden,J. (2016) RSAT::Plants: motif discovery in chip-Seq peaks of plant genomes. *Methods Mol. Biol. Clifton NJ*, **1482**, 297–322.
30. Defrance,M., Janky,R., Sand,O. and van Helden,J. (2008) Using RSAT oligo-analysis and dyad-analysis tools to discover regulatory signals in nucleic sequences. *Nat. Protoc.*, **3**, 1589–1603.
31. Bailey,T.L., Johnson,J., Grant,C.E. and Noble,W.S. (2015) The MEME suite. *Nucleic Acids Res.*, **43**, W39–W49.
32. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Mol. Cell*, **38**, 576–589.

Capítulo 5

Conclusiones

Las conclusiones de la tesis presente están divididas en dos, ya que hay dos contribuciones principales en el trabajo. La primera corresponde a los resultados obtenidos del análisis de datos y de redes de regulación de los datos de RNA-seq provenientes de infecciones virales a líneas celulares o tejidos de pacientes con COVID-19 o sanos. La segunda contribución son las herramientas bioinformáticas desarrolladas para el análisis de redes de regulación, siendo una un resultado directo de la primera contribución y otra una herramienta generada para la plataforma RSAT.

5.1. Análisis de redes de regulación bajo la infección de SARS-CoV-2

Se describen a continuación las conclusiones de los tres enfoques de análisis (Líneas celulares infectadas, tejidos con COVID-29 y Pulmón COVID-9) para entender un poco más la fisiopatología del COVID-19 desde la perspectiva molecular.

5.1.1. Infección de SARS-CoV-2 en comparación a otros virus en líneas celulares

Se utilizaron los datos de RNA-seq del estudio de Blanco-Melo et al. (4) correspondientes a experimentos de infecciones virales (RSV, IAV, HPIV3 y SARS-CoV-2) con modificaciones o tratamientos (expresión de ACE2, represión de NS1, tratamiento de IFN-β o de Ruxolitinib) en líneas celulares epiteliales (A549, NHBE y Calu3) que afectan principalmente al sistema respiratorio. El diseño experimental de estos datos nos permitió realizar una comparación sin necesidad de corregir por efectos de lote de las infecciones virales respiratorias y la particular de SARS-CoV-2, permitiéndonos así ver cuales procesos moleculares son específicamente activados por este virus.

En un primer análisis del perfil transcripcional, se hizo la observación de que las muestras infectadas por SARS-CoV-2 son heterogéneas, pero al menos tres experimentos (SARS-CoV-2 A549, SARS-CoV-2 A549 + ACE2 y SARS-CoV-2 NHBE) se agrupan. Se sugiere que uno de los elementos que forman parte de la agrupación de estas muestras es la respuesta a interferón indicado por la presencia en el mismo cúmulo del experimento de NHBE tratadon con IFN-β; así como la esperada diferencia en el perfil transcriptómico con el experimento SARS-CoV-2 Calu-3 dado que la línea celular es cancerosa a diferencia de las otras. Adicionalmente, y como era lo esperado, se encontró que en el perfil transcripcional general, la línea celular A549 con o sin ACE2 es la que más se asemeja a las células epiteliales alveolares tipo 2 (AT2), que son las que entran en contacto directo con el SARS-CoV-2 durante la infección.

Posteriormente, se generaron los regulones y se encontraron aquellos que estaban sobre-activados y que eran específicos por experimento. Para encontrar regulones novedosos, que se activaran en el proceso infeccioso de manera particular por SARS-CoV-2 y distintos de los que incluyeran la respuesta de interferón, se buscaron los regulones distintamente compartidos entre experimentos de infección a SARS-CoV-2, cuyos resultados se encuentran en las Tablas (4.13 y 4.24). Es importante notar que aunque se encontraron muchos regulones específicos para el experimento SARS-CoV-2-inf-Calu3, estos regulones son ciertamente menos confiables ya que el umbral impuesto sobre el FDR de la prueba de sobre-activación fue más laxo.

Los perfiles de activación de regulones se notaron distintos de aquellos en controles como se mostró en los heatmaps 4.16, 4.17 y 4.18, y funciones biológicas asociadas a los regulones que comprometían también hicieron sentido al involucrar funciones como vías de señalización activadas en respuesta a patógenos, vías implicadas en el sistema inmune innato, inflamación por liberación de citoquinas, procesos apoptóticos y activación del proteosoma; que de manera conjunta sugieren que los regulones sobre-activados en respuesta a la infección por SARS-CoV-2 en líneas de epitelio pulmonar dan lugar a una respuesta de ataque: se destruye el virus, se activa el sistema inmune y procesos inflamatorios y se induce la muerte celular.

Para encontrar los regulones más importantes involucrados en estos procesos, se eligieron aquellos que fueran específicos y tuvieran un log fold change positivo en comparación con los controles. Como se puede notar en la Figura 4.25, la infección de SARS-CoV-2 tiene una marca particular de activación de redes. En particular, los regulones novedosos (es decir, GRNs activadas por esta infección que no se han reportado en nuestro conocimiento) FOXP4 y ZNF730 son específicos de la infección de SARS-CoV-2 y los regulones SMARCA5, RGXR, NEUROD2 e IRF5 parecen particularmente importantes en la infección por SARS-CoV-2 en la línea celular A549 en general (estos regulones tenían una notoria firma de alto RSS en estas muestras con o sin expresar por vector ACE2 y con o sin el tratamiento de Ruxolitinib). De manera consistente, los regulones específicos SMARCA5 y RXRG se encontraron como distintamente compartidos y se encontró que tenían términos biológicos enriquecidos involucrados en la respuesta a virus y a la muerte celular (véase Fig 4.28). Cabe mencionar también la presencia de regulones guiados por TFs conocidos por su rol en el sistema inmune

innato y en procesos de inflamación de la familia STAT, ETV e IRF.

Al ver los regulones específicos para Calu3, se vio otra huella regulatoria particular para la infección de SARS-CoV-2, esto fue la activación particular de los regulones: VDR, ZNF471, ZNF485, ZNF208, PROM10, ZNF560 y NHLH2. Los términos enriquecidos para los regulones específicos de esta línea celular están involucrados en la respuesta a distintos patógenos, citoquinas como SMAD y TGF y crecimiento celular.

Para finalizar, estos resultados son consistentes entre sí al encontrar repetidamente el involucramiento de procesos inflamatorios, la activación de vías de señalización inmunológicas y de destrucción del virus, que en conjunto dan mayor confianza sobre que los regulones mencionados anteriormente sean de hecho importantes durante la infección de SARS-CoV-2.

5.1.2. COVID-19 en distintos tejidos

Se utilizaron los datos de RNA-seq provenientes de los estudios de Desai et al. (16) y Aguet et al. (3) para analizar tejidos obtenidos de pacientes con COVID-19 o sanos. Dado que esto introducía un sesgo potencial al tener muestras control y caso de dos fuentes independientes, se realizaron dos correcciones de efectos de lote, como se demostró favorable (véanse Figs. 4.4 y 4.5). En general se vio que estas muestras se segregaban principalmente tanto por el estatus de enfermedad como de un innegable sesgo proveniente del dataset, pero que las muestras se lograron integrar entre tejidos, mezclando casos de enfermedad y control.

Posteriormente se reconstruyeron los regulones por condición y se buscaron aquellos que estuvieran sobre-activados o sub-activados. Se corroboró que estos regulones tenían una firma de activación bastante notable entre casos de enfermedad o casos sanos (véase Fig. 4.31). Para conocer cuáles de ellos se prenden específicamente en algún órgano o que sólo se activan en cualquier par de órganos, estos se visualizaron y se encontraron las intersecciones como se muestra en las Figuras 4.29 y 4.30.

Como se esperaba, el pulmón fue el órgano tanto con mayor número de regulones sobre-activados como específicos, pues es donde inicia el proceso infeccioso. Pero también hubo un gran número de regulones prendidos en común en pulmón, hígado, intestino, intestino y corazón: 112 (Fig. 4.29), indicando que estos órganos lidian con la infección de una manera muy similar. Quizás más interesantes, son los regulones que no se prenden en pulmón pero sí en otros tejidos ante la infección, ya que han sido objeto de una menor atención en la comunidad científica.

De manera interesante, al buscar también cuales eran los regulones más específicos por condición se encontró que los órganos que más compartían regulones sobre-activados y específicos con pulmón eran el corazón e hígado como se muestra en la Figura 4.33 y los 22 regulones compartidos específicos se mencionan en el Heatmap 4.34. Hay un sólo regulon que no tiene un valor de especificidad RSS alto en pulmón pero sí en otro órgano: KLF14 en el hígado.

Se encontraron términos biológicos enriquecidos en los regulones relevantes de corazón, hígado y riñón relacionados a la proliferación, migración y diferenciación del

tejido; lo cuál se podría interpretar como una respuesta de reparación del tejido dañado ante la infección.

5.1.3. Pulmón con COVID-19 en comparación con sano

Se integraron datos de biopsias pulmonares provenientes de pacientes con COVID-19 de tres estudios independientes (Blanco-Melo et al. (4), Desai et al. (16) y Delorey et al. (15)). Para rescatar la variación biológica de los datos por encima de los técnicos se llevaron a cabo dos correcciones de efecto de lote, que se notó favorable al verse mezclados los datos de distintos estudios pero de una misma condición biológica (Fig. 4.6 B).

Usando un umbral de selección de regulones sobre-activados estricto (FDR 0.01), se encontraron únicamente 4 regulones sobre-activados: TP53, NANOG, IRX6 y HTATIP2 (véase Fig. 4.38), que también tiene el término de 'desarrollo de célula epitelial' enriquecido, observación que ha sido vista por otros grupos Delorey et al. (15). Y en cambio, se encontraron varios regulones sub-regulados. Utilizando un umbral menos estricto, se encontró que varios de los regulones se compartían con los regulones relevantes encontrados en el análisis de tejidos para pulmón en particular. Además, varios de los TFs guía de estos regulones están catalogados aquí como con función inmunológica.

5.2. Herramientas bioinformáticas desarrolladas

5.2.1. Pipeline de procesamiento de datos de RNA-seq para el análisis de redes de regulación de genes

Durante el desarrollo del trabajo de la tesis presente, se desarrolló una *pipeline* para el procesamiento de datos provenientes de la tecnología de secuenciación de RNAs (RNA-seq) para el análisis de redes de regulación o regulones. Fue estandarizada con los cuatro *datasets* utilizados aquí y se encuentra debidamente documentada en los reportes de análisis de datos para facilitar su reproducción o uso en futuros estudios.

El panorama general de los pasos involucrados en la *pipeline* pueden verse en la Figura 3.1 adaptada para los datos utilizados aquí; y una descripción breve de estrategias y pasos específicos se encuentra en la sección 3.2. Se considera que esta es una de las principales contribuciones.

5.2.2. network-interactions de RSAT

Se desarrolló una herramienta para la contrucción y comparación de redes de regulación llamada *network-interactions*. Para utilizar este programa solamente son necesarios dos archivos: una lista de los factores transcripcionales de interés, un archivo con las coordenadas genómicas de los genes de interés, opcionalmente, para la comparación de

redes, sólo es necesario agregar un archivo mencionando cada interacción gen-gen de la red previamente realizada. Esta función es conveniente ya que *network-interactions* puede aplicarse como un segundo paso durante la generación de redes a manera de corrección, es decir, para la depuración de interacciones no sustentadas con información de los TFBSs.

Los archivos de salida son principalmente tres: la red TF-gen generada con los archivos de entrada, la red indicando sólo las interacciones TF-TF y una red de interacciones indirectas con pasos de 3, es decir, TF-TF-gene; adicionalmente, las comparaciones de la red de entrada y la generada por el programa.

Se puede utilizar tanto en línea de comandos al tener instalado RSAT o bien, en la plataforma web del servidor metazoa de RSAT, donde además se puede generar una archivo html con toda la información de salida resumida. Para un primer paso, se puede probar el DEMO disponible.

5.2.3. Colaboración en REST-API de RSAT

Adicionalmente, se contribuyó en el servicio de interfaz REST-API para RSAT servidor Fungi y para su acceso por programas en python al dar mantenimiento a los programas de RSAT accesibles desde este servicio. Se probaron todos los programas, se discutieron los errores y posibles soluciones con los autores de los distintos códigos y se realizó el proceso de corrección de código.

Bibliografía

[1] (2020). What is a rest api? 74

[2] (2020). Who coronavirus (covid-19) dashboard. 5

[3] Aguet, F., Anand, S., Ardlie, K. G., Gabriel, S., Getz, G. A., Graubert, A., Hadley, K., Handsaker, R. E., Huang, K. H., Kashin, S., Li, X., MacArthur, D. G., Meier, S. R., Nedzel, J. L., Nguyen, D. T., Segrè, A. V., Todres, E., Balliu, B., Barbeira, A. N., Battle, A., Bonazzola, R., Brown, A., Brown, C. D., Castel, S. E., Conrad, D. F., Cotter, D. J., Cox, N., Das, S., de Goede, O. M., Dermitzakis, E. T., Einson, J., Engelhardt, B. E., Eskin, E., Eulalio, T. Y., Ferraro, N. M., Flynn, E. D., Fresard, L., Gamazon, E. R., Garrido-Martín, D., Gay, N. R., Gloudemans, M. J., Guigó, R., Hame, A. R., He, Y., Hoffman, P. J., Hormozdiari, F., Hou, L., Im, H. K., Jo, B., Kasela, S., Kellis, M., Kim-Hellmuth, S., Kwong, A., Lappalainen, T., Li, X., Liang, Y., Mangul, S., Mohammadi, P., Montgomery, S. B., Muñoz-Aguirre, M., Nachun, D. C., Nobel, A. B., Oliva, M., Park, Y., Park, Y., Parsana, P., Rao, A. S., Reverter, F., Rouhana, J. M., Sabatti, C., Saha, A., Stephens, M., Stranger, B. E., Strober, B. J., Teran, N. A., Viñuela, A., Wang, G., Wen, X., Wright, F., Wucher, V., Zou, Y., Ferreira, P. G., Li, G., Melé, M., Yeger-Lotem, E., Barcus, M. E., Bradbury, D., Krubit, T., McLean, J. A., Qi, L., Robinson, K., Roche, N. V., Smith, A. M., Sobin, L., Tabor, D. E., Undale, A., Bridge, J., Brigham, L. E., Foster, B. A., Gillard, B. M., Hasz, R., Hunter, M., Johns, C., Johnson, M., Karasik, E., Kopen, G., Leinweber, W. F., McDonald, A., Moser, M. T., Myer, K., Ramsey, K. D., Roe, B., Shad, S., Thomas, J. A., Walters, G., Washington, M., Wheeler, J., Jewell, S. D., Rohrer, D. C., Valley, D. R., Davis, D. A., Mash, D. C., Branton, P. A., Barker, L. K., Gardiner, H. M., Mosavel, M., Siminoff, L. A., Flicek, P., Haeussler, M., Juettemann, T., Kent, W. J., Lee, C. M., Powell, C. C., Rosenbloom, K. R., Ruffier, M., Sheppard, D., Taylor, K., Trevanion, S. J., Zerbino, D. R., Abell, N. S., Akey, J., Chen, L., Demanelis, K., Doherty, J. A., Feinberg, A. P., Hansen, K. D., Hickey, P. F., Jasmine, F., Jiang, L., Kaul, R., Kibriya, M. G., Li, J. B., Li, Q., Lin, S., Linder, S. E., Pierce, B. L., Rizzardi, L. F., Skol, A. D., Smith, K. S., Snyder, M., Stamatoyannopoulos, J., Tang, H., Wang, M., Carithers, L. J., Guan, P., Koester, S. E., Little, A. R., Moore, H. M., Nierras, C. R., Rao, A. K., Vaught, J. B., and

Volpi, S. (2020). The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330. 29, 33, 82, 141

[4] Blanco-Melo, D., Nilsson-Payant, B. E., Liu, W.-C., Uhl, S., Hoagland, D., Møller, R., Jordan, T. X., Oishi, K., Panis, M., Sachs, D., Wang, T. T., Schwartz, R. E., Lim, J. K., Albrecht, R. A., and tenOever, B. R. (2020). Imbalanced host response to SARS-CoV-2 drives development of COVID-19. *Cell*, 181(5):1036–1045.e9. 2, 7, 23, 25, 29, 33, 77, 139, 142

[5] Bohn, M. K., Hall, A., Sepiashvili, L., Jung, B., Steele, S., and Adeli, K. (2020). Pathophysiology of COVID-19: Mechanisms underlying disease severity and progression. *Physiology*, 35(5):288–301.

[6] Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic rna-seq quantification. *Nature Biotechnology*, 34(5):525–527.

[7] Castro-Mondragon, J. A., Jaeger, S., Thieffry, D., Thomas-Chollier, M., and van Helden, J. (2017). RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections. *Nucleic Acids Research*, 45(13):e119–e119.

[8] Castro-Mondragon, J. A., Riudavets-Puig, R., Rauluseviciute, I., Lemma, R. B., Turchi, L., Blanc-Mathieu, R., Lucas, J., Boddie, P., Khan, A., Pérez, N. M., Fornes, O., Leung, T. Y., Aguirre, A., Hammal, F., Schmelter, D., Baranasic, D., Ballester, B., Sandelin, A., Lenhard, B., Vandepoele, K., Wasserman, W. W., Parcy, F., and Mathelier, A. (2021). JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 50(D1):D165–D173.

[9] Chen, X., Tan, Q., Wang, Y., Lv, H., Wang, Z., Lin, Z., Du, Z., Xiong, S., Han, J., Tian, D., and Wang, B. (2019). Overexpression of KLF14 protects against immune-mediated hepatic injury in mice. *Laboratory Investigation*, 99(1):37–47.

[10] Chua, R. L., Lukassen, S., Trump, S., Hennig, B. P., Wendisch, D., Pott, F., Debnath, O., Thürmann, L., Kurth, F., Völker, M. T., Kazmierski, J., Timmermann, B., Twardziok, S., Schneider, S., Machleidt, F., Müller-Redetzky, H., Maier, M., Krannich, A., Schmidt, S., Balzer, F., Liebig, J., Loske, J., Suttorp, N., Eils, J., Ishaque, N., Liebert, U. G., von Kalle, C., Hocke, A., Witzenrath, M., Goffinet, C., Drosten, C., Laudi, S., Lehmann, I., Conrad, C., Sander, L.-E., and Eils, R. (2020). COVID-19 severity correlates with airway epithelium–immune cell interactions identified by single-cell analysis. *Nature Biotechnology*, 38(8):970–979.

[11] Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., and Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, 17(1).

[12] Consortium, G. O. (2004). The gene ontology (GO) database and informatics resource. *Nucleic Acids Research*, 32(90001):258D–261.

[13] Contreras-Moreira, B. and Sebastian, A. (2016). FootprintDB: Analysis of plant cis-regulatory elements, transcription factors, and binding interfaces. In *Methods in Molecular Biology*, pages 259–277. Springer New York.

[14] de Sande, B. V., Flerin, C., Davie, K., Waegeneer, M. D., Hulselmans, G., Aibar, S., Seurinck, R., Saelens, W., Cannoodt, R., Rouchon, Q., Verbeiren, T., Maeyer, D. D., Reumers, J., Saeys, Y., and Aerts, S. (2020). A scalable SCENIC workflow for single-cell gene regulatory network analysis. *Nature Protocols*, 15(7):2247–2276.

[15] Delorey, T. M., Ziegler, C. G. K., Heimberg, G., Normand, R., Yang, Y., Segerstolpe, Å., Abbondanza, D., Fleming, S. J., Subramanian, A., Montoro, D. T., Jagadeesh, K. A., Dey, K. K., Sen, P., Slyper, M., Pita-Juárez, Y. H., Phillips, D., Biermann, J., Bloom-Ackermann, Z., Barkas, N., Ganna, A., Gomez, J., Melms, J. C., Katsyv, I., Normandin, E., Naderi, P., Popov, Y. V., Raju, S. S., Niezen, S., Tsai, L. T.-Y., Siddle, K. J., Sud, M., Tran, V. M., Vellarikkal, S. K., Wang, Y., Amir-Zilberstein, L., Atri, D. S., Beechem, J., Brook, O. R., Chen, J., Divakar, P., Dorceus, P., Engreitz, J. M., Essene, A., Fitzgerald, D. M., Fropf, R., Gazal, S., Gould, J., Grzyb, J., Harvey, T., Hecht, J., Hether, T., Jané-Valbuena, J., Leney-Greene, M., Ma, H., McCabe, C., McLoughlin, D. E., Miller, E. M., Muus, C., Niemi, M., Padera, R., Pan, L., Pant, D., Pe'er, C., Pfiffner-Borges, J., Pinto, C. J., Plaisted, J., Reeves, J., Ross, M., Rudy, M., Rueckert, E. H., Siciliano, M., Sturm, A., Todres, E., Waghray, A., Warren, S., Zhang, S., Zollinger, D. R., Cosimi, L., Gupta, R. M., Hacohen, N., Hibshoosh, H., Hide, W., Price, A. L., Rajagopal, J., Tata, P. R., Riedel, S., Szabo, G., Tickle, T. L., Ellinor, P. T., Hung, D., Sabeti, P. C., Novak, R., Rogers, R., Ingber, D. E., Jiang, Z. G., Juric, D., Babadi, M., Farhi, S. L., Izar, B., Stone, J. R., Vlachos, I. S., Solomon, I. H., Ashenberg, O., Porter, C. B. M., Li, B., Shalek, A. K., Villani, A.-C., Rozenblatt-Rosen, O., and Regev, A. (2021). COVID-19 tissue atlases reveal SARS-CoV-2 pathology and cellular targets. *Nature*, 595(7865):107–113.

[16] Desai, N., Neyaz, A., Szabolcs, A., Shih, A. R., Chen, J. H., Thapar, V., Nieman, L. T., Solovyov, A., Mehta, A., Lieb, D. J., Kulkarni, A. S., Jaicks, C., Xu, K. H., Raabe, M. J., Pinto, C. J., Juric, D., Chebib, I., Colvin, R. B., Kim, A. Y., Monroe, R., Warren, S. E., Danaher, P., Reeves, J. W., Gong, J., Rueckert, E. H., Greenbaum, B. D., Hacohen, N., Lagana, S. M., Rivera, M. N., Sholl, L. M., Stone, J. R., Ting, D. T., and Deshpande, V. (2020). Temporal and spatial heterogeneity of host response to SARS-CoV-2 pulmonary infection. *Nature Communications*, 11(1).

[17] Dogan, B., Kailasam, S., Corchado, A. H., Nikpoor, N., and Najafabadi, H. S. (2019). A domain-resolution map of in vivo DNA binding reveals the regulatory consequences of somatic mutations in zinc finger transcription factors.

[18] Drost, H.-G. and Paszkowski, J. (2017). Biomartr: genomic data retrieval with R. *Bioinformatics*, 33(8):1216–1217.

[19] Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19):3047–3048.

[20] Frith, M. C. (2003). Cluster-buster: finding dense clusters of motifs in DNA sequences. *Nucleic Acids Research*, 31(13):3666–3668.

[21] Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D., and Saez-Rodriguez, J. (2019). Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Research*, 29(8):1363–1375.

[22] Griffiths, A. J. F., Wessler, S. R., Carroll, S. B., and Doebley, J. (2011a). *INTRODUCTION TO GENETIC ANALYSIS*. W. H. Freeman and Company, New Yprk, United States of America.

[23] Griffiths, A. J. F., Wessler, S. R., Carroll, S. B., and Doebley, J. (2011b). *INTRODUCTION TO GENETIC ANALYSIS*. W. H. Freeman and Company, New Yprk, United States of America.

[24] Gu, Z. (2022). Complex heatmap visualization. *iMeta*, 1(3).

[25] Hernandez-Corchado, A. and Najafabadi, H. S. (2021). A base-resolution panorama of the in vivo impact of cytosine methylation on transcription factor binding.

[26] Hrdlickova, R., Toloue, M., and Tian, B. (2016). Rna-seq methods for transcriptome analysis. *WIREs RNA*, 8(1).

[27] Hu, Z., Huang, X., Zhang, J., Fu, S., Ding, D., and Tao, Z. (2022). Differences in clinical characteristics between delta variant and wild-type SARS-CoV-2 infected patients. *Frontiers in Medicine*, 8.

[28] Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE*, 5(9):e12776.

[29] Ibrahim, D. M. and Mundlos, S. (2020). The role of 3d chromatin domains in gene regulation: a multi-facetted view on genome organization. *Current Opinion in Genetics and Development*, 61:1–8.

[30] Imrichová, H., Hulselmans, G., Atak, Z. K., Potier, D., and Aerts, S. (2015). i-cisTarget 2015 update: generalized cis-regulatory enrichment analysis in human, mouse and fly. *Nucleic Acids Research*, 43(W1):W57–W64.

[31] Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K. R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., Palin, K., Vaquerizas, J. M., Vincentelli, R., Luscombe, N. M., Hughes, T. R., Lemaire, P., Ukkonen, E., Kivioja, T., and Taipale, J. (2013). DNA-binding specificities of human transcription factors. *Cell*, 152(1-2):327–339.

[32] Jung, S., Potapov, I., Chillara, S., and del Sol, A. (2021). Leveraging systems biology for predicting modulators of inflammation in patients with COVID-19. *Science Advances*, 7(6).

[33] Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., and Yamanishi, Y. (2007). KEGG for linking genomes to life and the environment. *Nucleic Acids Research*, 36(Database):D480–D484.

[34] Kang, Y., Thieffry, D., and Cantini, L. (2021). Evaluating the reproducibility of single-cell gene regulatory network inference algorithms. *Frontiers in Genetics*, 12.

[35] Kellis, M. (2021). 18.1: Introduction to regulatory genomics.

[36] Khan, P., Fytianos, K., Tamò, L., Roth, M., Tamm, M., Geiser, T., Gazdhar, A., and Hostettler, K. E. (2018). Culture of human alveolar epithelial type II cells by sprouting. *Respiratory Research*, 19(1).

[37] Kheradpour, P. and Kellis, M. (2013). Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Research*, 42(5):2976–2987.

[38] Kuhn, R. M., Haussler, D., and Kent, W. J. (2012). The UCSC genome browser and associated tools. *Briefings in Bioinformatics*, 14(2):144–161.

[39] Kulakovskiy, I. V., Vorontsov, I. E., Yevshin, I. S., Sharipov, R. N., Fedorova, A. D., Rumynskiy, E. I., Medvedeva, Y. A., Magana-Mora, A., Bajic, V. B., Papatsenko, D. A., Kolpakov, F. A., and Makeev, V. J. (2017). HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-seq analysis. *Nucleic Acids Research*, 46(D1):D252–D259.

[40] Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R., and Weirauch, M. T. (2018). The human transcription factors. *Cell*, 172(4):650–665.

[41] Lamers, M. M. and Haagmans, B. L. (2022). SARS-CoV-2 pathogenesis. *Nature Reviews Microbiology*, 20(5):270–284.

[42] Liao, M., Liu, Y., Yuan, J., Wen, Y., Xu, G., Zhao, J., Cheng, L., Li, J., Wang, X., Wang, F., Liu, L., Amit, I., Zhang, S., and Zhang, Z. (2020). Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nature Medicine*, 26(6):842–844.

[43] LU, Y., WAMBACH, M., KATZE, M. G., and KRUG, R. M. (1995). Binding of the influenza virus NS1 protein to double-stranded RNA inhibits the activation of the protein kinase that phosphorylates the eIF-2 translation initiation factor. *Virology*, 214(1):222–228.

[44] Luscombe, N. M., Greenbaum, D., and Gerstein, M. (2001). What is bioinformatics? a proposed definition and overview of the field. *Methods of information in medicine*, 40(04):346–358.

[45] Marstrand, T. T. and Storey, J. D. (2014). Identifying and mapping cell-type-specific chromatin programming of gene expression. *Proceedings of the National Academy of Sciences*, 111(6).

[46] McCarthy, D. J., Campbell, K. R., Lun, A. T. L., and Wills, Q. F. (2017). Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in r. *Bioinformatics*, page btw777.

[47] Mercatelli, D., Scalambra, L., Triboli, L., Ray, F., and Giorgi, F. M. (2020). Gene regulatory network inference resources: A practical overview. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1863(6):194430.

[48] Minkoff, J. M. and tenOever, B. (2023). Innate immune evasion strategies of SARS-CoV-2. *Nature Reviews Microbiology*.

[49] Moerman, T., Santos, S. A., González-Blas, C. B., Simm, J., Moreau, Y., Aerts, J., and Aerts, S. (2018). GRNBoost2 and arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics*, 35(12):2159–2161.

[50] Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A., and Bejerano, G. (2013). Enhancers: five essential questions. *Nature Reviews Genetics*, 14(4):288–295.

[51] Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A., and Murali, T. M. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods*, 17(2):147–154.

[52] Quinlan, A. R. (2014). BEDTools: The swiss-army tool for genome feature analysis. *Current Protocols in Bioinformatics*, 47(1).

[53] Reimers, M. and Carey, V. J. (2006). [8] bioconductor: An open source framework for bioinformatics and computational biology. In *Methods in Enzymology*, pages 119–134. Elsevier.

[54] Safran, M., Dalah, I., Alexander, J., Rosen, N., Stein, T. I., Shmoish, M., Nativ, N., Bahir, I., Doniger, T., Krug, H., Sirota-Madi, A., Olender, T., Golan, Y., Stelzer, G., Harel, A., and Lancet, D. (2010). GeneCards version 3: the human gene integrator. *Database*, 2010(0):baq020–baq020.

[55] Santana-Garcia, W., Castro-Mondragon, J. A., Padilla-Gálvez, M., Nguyen, N. T. T., Elizondo-Salas, A., Ksouri, N., Gerbes, F., Thieffry, D., Vincens, P., Contreras-Moreira, B., van Helden, J., Thomas-Chollier, M., and Medina-Rivera, A. (2022). RSAT 2022: regulatory sequence analysis tools. *Nucleic Acids Research*, 50(W1):W670–W676.

[56] Siddiqi, H. K. and Mehra, M. R. (2020). COVID-19 illness in native and immunosuppressed states: A clinical–therapeutic staging proposal. *The Journal of Heart and Lung Transplantation*, 39(5):405–407.

[57] Singh, A. J., Ramsey, S. A., Filtz, T. M., and Kioussi, C. (2017). Differential gene regulatory networks in development and disease. *Cellular and Molecular Life Sciences*, 75(6):1013–1025.

[58] Suo, S., Zhu, Q., Saadatpour, A., Fei, L., Guo, G., and Yuan, G.-C. (2018). Revealing the critical regulators of cell identity in the mouse cell atlas. *Cell Reports*, 25(6):1436–1445.e3.

[59] Tanaka, Y., Higashihara, K., Nakazawa, M. A., Yamashita, F., Tamada, Y., and Okuno, Y. (2021). Dynamic changes in gene-to-gene regulatory networks in response to SARS-CoV-2 infection. *Scientific Reports*, 11(1).

[60] Tay, M. Z., Poh, C. M., Rénia, L., MacAry, P. A., and Ng, L. F. P. (2020). The trinity of COVID-19: immunity, inflammation and intervention. *Nature Reviews Immunology*, 20(6):363–374.

[61] Truty, M. J., Lomberk, G., Fernandez-Zapico, M. E., and Urrutia, R. (2009). Silencing of the transforming growth factor- (TGF) receptor II by krüppel-like factor 14 underscores the importance of a negative feedback mechanism in TGF signaling. *Journal of Biological Chemistry*, 284(10):6291–6300.

[62] Turatsinze, J.-V., Thomas-Chollier, M., Defrance, M., and van Helden, J. (2008). Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nature Protocols*, 3(10):1578–1588.

[63] Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63.

[64] Wanner, N., Andrieux, G., i Mompel, P. B., Edler, C., Pfefferle, S., Lindenmeyer, M. T., Schmidt-Lauber, C., Czogalla, J., Wong, M. N., Okabayashi, Y., Braun, F., Lütgehetmann, M., Meister, E., Lu, S., Noriega, M. L. M., Günther, T., Grundhoff, A., Fischer, N., Bräuninger, H., Lindner, D., Westermann, D., Haas, F., Roedl, K., Kluge, S., Addo, M. M., Huber, S., Lohse, A. W., Reiser, J., Ondruschka, B., Sperhake, J. P., Saez-Rodriguez, J., Boerries, M., Hayek, S. S., Aepfelbacher, M., Scaturro, P., Puelles, V. G., and Huber, T. B. (2022). Molecular consequences of SARS-CoV-2 liver tropism. *Nature Metabolism*, 4(3):310–319.

[65] Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., Najafabadi, H. S., Lambert, S. A., Mann, I., Cook, K., Zheng, H., Goity, A., van Bakel, H., Lozano, J.-C., Galli, M., Lewsey, M. G., Huang, E., Mukherjee, T., Chen, X., Reece-Hoyes, J. S., Govindarajan, S., Shaulsky, G., Walhout, A. J., Bouget, F.-Y., Ratsch, G., Larrondo, L. F., Ecker, J. R., and Hughes, T. R. (2014).

Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, 158(6):1431–1443.

[66] Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.

[67] Williams, C. R., Baccarella, A., Parrish, J. Z., and Kim, C. C. (2016). Trimming of sequence reads alters RNA-seq gene expression estimates. *BMC Bioinformatics*, 17(1).

[68] Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., Fu, X., Liu, S., Bo, X., and Yu, G. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation*, 2(3):100141.

[69] Xie, Z., Hu, S., Blackshaw, S., Zhu, H., and Qian, J. (2009). hPDI: a database of experimental human protein-DNA interactions. *Bioinformatics*, 26(2):287–289.

[70] Yang, L., Liu, S., Liu, J., Zhang, Z., Wan, X., Huang, B., Chen, Y., and Zhang, Y. (2020). COVID-19: immunopathogenesis and immunotherapeutics. *Signal Transduction and Targeted Therapy*, 5(1).

[71] Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., Das, P. K., Kivioja, T., Dave, K., Zhong, F., Nitta, K. R., Taipale, M., Popov, A., Ginno, P. A., Domcke, S., Yan, J., Schübeler, D., Vinson, C., and Taipale, J. (2017). Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science*, 356(6337).

[72] Yu, G., Wang, L.-G., Yan, G.-R., and He, Q.-Y. (2014). DOSE: an r/bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics*, 31(4):608–609.

[73] Yuki, K., Fujiogi, M., and Koutsogiannaki, S. (2020). COVID-19 pathophysiology: A review. *Clinical Immunology*, 215:108427.

[74] Zhang, Y., Parmigiani, G., and Johnson, W. E. (2020). ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genomics and Bioinformatics*, 2(3).

[75] Zhu, F., Farnung, L., Kaasinen, E., Sahu, B., Yin, Y., Wei, B., Dodonova, S. O., Nitta, K. R., Morgunova, E., Taipale, M., Cramer, P., and Taipale, J. (2018). The interaction landscape between transcription factors and the nucleosome. *Nature*, 562(7725):76–81.