UNIVERSIDAD DE GRANADA

PROGRAMA DE DOCTORADO EN ESTADÍSTICA MATEMÁTICA Y APLICADA

Departamento de Estadística e Investigación Operativa – Universidad de Granada

Centro Pfizer – Universidad de Granada – Junta de Andalucía de Genómica e Investigación Oncológica (GENYO)

TESIS DOCTORAL

Análisis integrado y meta-análisis de datos biomédicos

Juan Antonio Villatoro García

Dirigida por:

Pedro Carmona Sáez

Granada, 2024

Editor: Universidad de Granada. Tesis Doctorales
Autor: Juan Antonio Villatoro García
ISBN: 978-84-1195-673-4
URI: https://hdl.handle.net/10481/102172

"*No diré no lloréis, pues no todas las lágrimas son amargas*"

(Gandalf - El Señor de los Anillos: El Retorno del Rey)

AGRADECIMIENTOS

Algunos describen la vida como una carrera llena de obstáculos que debemos superar para pasar por las diferentes etapas que la componen. Y en este caso, una etapa llega a su fin. Sin embargo, me voy a permitir el lujo de definir también la vida como la define mi padre: "La vida son combinaciones". Defínción que me gusta especialmente ya que se ajusta mucho a mi campo de estudio que es la Estadística. Y en esta serie de combinaciones y probabilidades me he encontrado con personas que me han ayudado y empujado a ir superando los diferentes obstáculos. Dichosa mi suerte que necesitaría escribir otra tesis y mucho más para expresar lo agradecido que estoy.

Muchas gracias en primer lugar, a mi director de tesis, Pedro, por la confianza puesta en mí. Has sido mi orientador y guía en este ciclo. Sólo puedo agradercete todo el apoyo y enseñanzas que me has dado para afrontar este entorno tan complejo que es el mundo de la investigación.

Además, en este mundo de la investigación no he estado sólo, sino que he estado rodeado de un equipo de profesionales y compañeros que, hoy en día, puedo llamar amigos. Sin vosotros mi camino en la ciencia habría sido mucho más complicado. Gracias Dani por tu creatividad e imaginación. Con todas las ideas que tienes en la mente, podríamos escribir miles de artículos y proyectos. Has sido un verdadero mentor para muchos de nosotros cuando dábamos nuestros primeros pasos en la ciencia. Gracias Jordi por ser otro maestro y un referente para mí. Tus conocimientos y consejos son de vital importancia y somos muy afortunados de contar contigo como compañero. Gracias Raúl por tu constancia y disposición. Tu forma de trabajar y de adaptación a la hora afrontar los retos te convierte en el compañero ideal que cualquiera querría tener en su equipo. Soy el fan número uno de tus gráficos de *ggplot2*. Gracias Adri por tu explosividad y espontaneidad. Aunque te digamos que matas moscas a cañonazos, te aseguro que eres la persona que ha movilizado y unido al grupo desde el principio. Gracias Alba por ser una integrante más del grupo. Sin ti los momentos de desconexión del trabajo no serían lo mismo para nosotros. Gracias Iván por tu tranquilidad y buen humor contagiosos. Aunque no estés tanto por GENyO siempre estás cuando el resto de personas te necesitamos. Gracias Marina por tu simpatía y alegría. Tu sonrisa anima al grupo y nos ayuda a ver las cosas con una perspectiva más positiva. Gracias Samu por tu actitud tan positiva. Tus anécdotas y experiencias sabes que nos encantan, pero nos gustan porque las cuentas tú. Gracias, Pablo, por tu impresionante capacidad de trabajo. Tus habilidades para sintetizar, organizar y apoyar a los demás hacen que la carga del grupo sea mucho más ligera. Y gracias al resto de personas que de una forma u otra han formado y forman parte de este grupo: Inés, Jose, Sergio, Robledillo y muchos más. Gracias porque sin vosotros las tapas, las videollamadas del COVID-19 y las quedadas no habrían sido lo mismo.

Agradecer también al resto de gente que forma GENyO. Aunque no mecione nombres específicos para asegurarme de inlcuir a todo el mundo, después de aproximadamente seis años en el centro, puedo afirmar con certeza que me llevo un sinfín de valiosos recuerdos y experiencias inolvidables.

También deseo extender mi agradecimiento a los profesores y compañeros del Departamento de Estadística e Investigación Operativa. Comencé mi camino como

alumno, y ahora me siento parte del mismo. Sería injusto destacar a alguien en particular, pero quisiera hacer una mención especial a Yolanda, mi tutora del TFG y TFM, quien fue la que me recomendó empezar a hacer las prácticas en GENyO. Y, por otro lado, quiero dar las gracias también a la Unidad Docente de Bioestadística de Medicina, donde he impartido la mayor parte de mis clases, y donde siempre he sentido el respaldo y la confianza necesarios para desarrollarme como docente.

Gracias a mis amigos de toda la vida Ramón y Andrés, aunque estemos a distancia no hemos perdido el contacto. Esas videollamadas y esos ratos que teníamos para vernos eran para mí un gran apoyo en momentos difíciles.

También quiero dar las gracias de una forma muy especial a la que más me ha aguantado y soportado este tiempo. Soy cosciente de lo aprensivo, intranquilo y pesado que soy muchas veces. Gracias por apoyarme, aconsejarme y sostenerme para que no me derrumbe. Gracias por redondear tantas veces mi mente tan matemática y cuadriculada. Gracias Rebeca por ser parte de mi vida.

Agredecer de corazón a mi familia. Gracias a mis tíos y a mis primos por su apoyo. Gracias de forma especial a mí tía Inés y a mi abuela porque sin vuestro cariño incondicional yo no sería el mismo. Y gracias por supuesto a mi madre, a mi padre y a mi hermano por todos esos momentos de sacrificio para ayudarme en todo lo que he necesitado. Gracias por mostrarme y guiarme a través de los diversos caminos que he recorrido, y por estar a mi lado en cada uno de ellos. Gracias por vuestro amor, confianza y fe en mí. Gracias porque no tengo palabras suficientes para expresar lo mucho que valoro todo lo que hacéis y habéis hecho por mí.

Por último, no quiero terminar los agradecimientos sin recordar a quienes ya no están, mis abuelos. Ellos también fueron una parte esencial de mi vida, y sin ellos esto no habría sido posible.

Solo espero que os sintáis orgullosos de mí. Muchas gracias a todos.

CRITERIOS DE CALIDAD DE LOS ARTÍCULOS QUE AVALAN LA TESIS

Los siguientes artículos avalan esta tesis doctoral

Toro-Domínguez D*, **Villatoro-García JA***, Martorell-Marugán J, Román-Montoya Y, Alarcón-Riquelme ME, Carmona-Sáez P. A survey of gene expression meta-analysis: methods and applications. *Briefing Bioinformatics*. Published online February 25, 2020. doi:10.1093/bib/bbaa019.
* Los autores comparten la primera autoría

Mathematics, Mathematical & Computational Biology			
Año JCR	Factor de Impacto	Rango	Cuantiles
2020	11.622	3/78	Q1,D1

Villatoro-García JA, Martorell-Marugán J, Toro-Domínguez D, Román-Montoya Y, Femia P, Carmona-Sáez P. DExMA: An R Package for Performing Gene Expression Meta-Analysis with Missing Genes. Mathematics. 2022;10(18):3376. doi:10.3390/math10183376.

Mathematics, Mathematics			
Año JCR	Factor de Impacto	Rango	Cuantiles
2022	2.4	23/330	Q1,D1

Villatoro-García JA, López-Domínguez R, Martorell-Marugán J, Luna J de D, Lorente JA, Carmona-Sáez P. Exploring the interplay between climate, population immunity and SARS-CoV-2 transmission dynamics in Mediterranean countries. Sci Total Environ. 2023;897:165487. doi:10.1016/j.scitotenv.2023.165487.

Biology & Biochemistry, Environmental Sciences			
Año JCR	Factor de Impacto	Rango	Cuantiles
2023	8.2	31/358	Q1,D1

Martorell-Marugán J*, **Villatoro-García JA***, García-Moreno A, López-Domínguez R, Requena F, Merelo JJ, Lacasaña M, Luna JD, Díaz-Mochón JJ, Lorente JA, Carmona-Sáez P. DatAC: A visual analytics platform to explore climate and air quality indicators associated with the COVID-19 pandemic in Spain. Sci Total Environ. 2021;750:141424. doi:10.1016/j.scitotenv.2020.141424.
* Los autores comparten la primera autoría

Biology & Biochemistry, Environmental Sciences			
Año JCR	Factor de Impacto	Rango	Cuantiles
2021	10.754	26/279	Q1,D1

Dado que se cumplen los criterios para redactar la tesis como compedio de artículos, este documento ha sido redactado siguiendo esta modalidad

FINANCIACIÓN:

Esta tesis ha sido posible gracias a la ayuda obtenida por el candidato:

- Ayudas para la Formación del Profesorado Universitario 2019. Ministerio de Ciencias, Innovación y Universidades (Referencia: FPU19/01999).

Además, esta tesis ha recibido soporte de los siguientes proyectos de investigación:

- Identificación de Biomarcadores en Lupus Eritematoso Sistémico Mediante Análisis Integrado de Transcriptoma y Metiloma
  Nombres investigadores principales: Pedro Carmona Saez
  Entidad/es financiadora/s: Consejería de Salud. Junta de Andalucía
  Nombre del programa: Ayudas para Investigación, Desarrollo e Innovación Biomédica y en Ciencias de la Salud en Andalucía.
  Referencia: PI-0173-2017
  Fecha de inicio-fin: 2017 - 2020
  Cuantía total: 59.364 €
- DatAC (Data Against COVID-19): Herramienta de integración de datos sobre COVID-19 y análisis de factores asociados a focos de contagio y propagación de la enfermedad
  Nombres investigadores principales: Pedro Carmona Sáez
  Entidad/es financiadora/s: Junta de Andalucía
  Nombre del programa: Proyectos de investigación sobre el SARS-COV-2 y la enfermedad COVID-19.
  Referencia: CV20-36723
  Fecha de inicio-fin: 09/09/2020 - 09/09/2021
  Cuantía total: 48.000 €
- Medicina genómica de precisión en enfermedades autoinmunes: Inferencia de redes de regulación y búsqueda de tratamientos
  Nombres investigadores principales: Pedro Carmona Sáez
  Entidad/es financiadora/s: Consejería de Transformación Economía, Industria, Conocimiento y Universidades, Junta de Andalucía
  Nombre del programa: Proyectos I+D+i 2020.
  Referencia: P20_00335
  Fecha de inicio-fin: 04/10/2021 - 30/06/2023
  Cuantía total: 100.350 €
- Integración de datos ómicos para descubrimiento de biomarcadores e inferencia de redes
  Nombres investigadores principales: Pedro Carmona Sáez
  Entidad/es financiadora/s: Ministerio de ciencia e innovación
  Nombre del programa: Proyectos del Plan Nacional 2020.
  Referencia: PID2020-119032RB-I00
  Fecha de inicio-fin: 01/09/2021 - 30/11/2024
  Cuantía total: 145.200 €

- AUTOIMMOMICS: Desarrollo de una plataforma centralizada de datos ómicos en enfermedades autoinmunes para el descubrimiento de nuevos tratamientos y biomarcadores
  Nombres investigadores principales: Pedro Carmona Sáez; Marta Alarcón Riquelme
  Entidad/es financiadora/s: Junta de Andalucía
  Nombre del programa: Proyectos I+D+i del Programa Operativo FEDER 2020.
  Referencia: B-CTS-40-UGR20
  Fecha de inicio-fin: 01/07/2021 - 30/06/2023
  Cuantía total: 30.000 €

ÍNDICE:

RESUMEN

En los últimos tiempos, el desarrollo y expansión de tecnologías experimentales de alto rendimiento ha provocado un notable incremento en la cantidad de datos generados y almacenados en bases de datos, marcando el inicio de la era del Big Data. En este nuevo contexto, no solo es vital obtener muestras representativas, sino también analizar la vasta cantidad de datos disponibles para descubrir nuevos conocimientos y formular nuevas hipótesis. Este panorama plantea desafíos y oportunidades significativos para la Estadística, cuyo rol es esencial para proporcionar los métodos y herramientas necesarios para examinar y analizar esta gran cantidad de información.

La integración de datos es un proceso crucial en este entorno, ya que busca combinar la información proveniente de múltiples fuentes para que pueda ser utilizada de manera coherente y eficiente. En los últimos años, las técnicas de integración de datos han evolucionado para poder almacenar y procesar datos en tiempo real, no estructurados y heterogéneos. Además, dentro de estos procesos, no solo es importante la combinación de diferentes niveles de información, sino también la obtención de resultados significativos a partir de ellos. Este es el objetivo del *Data Fusion* o Fusión de Datos, la última etapa de la integración de datos.

En el ámbito de los métodos de integración y fusión de datos, las técnicas de meta-análisis han ganado gran popularidad. Estas técnicas se centran en la combinación de los resultados obtenidos en diferentes estudios para obtener un resultado común y más confiable. No obstante, el uso extensivo de estas técnicas ha llevado en algunos casos a su aplicación inadecuada, generando problemas de fiabilidad y reproducibilidad de los resultados. Debido a esto, es crucial contar con flujos de trabajo definidos e implementados en software abierto que permitan aplicar correctamente estos métodos y asegurar la obtención de resultados fiables.

Este ha sido objetivo principal de esta tesis doctoral, en la cual se han desarrollado métodos y herramientas de software para integrar y aplicar técnicas de meta-análisis a datos biomédicos. En la era del *Big Data* los datos biomédicos han experimentado también una gran expansión. Entre ellos, algunos de los que han ganado más relevancia han sido los datos ómicos, y concretamente los datos de transcriptómica o expresión génica, gracias a los avances en las técnicas de secuenciación y, por otro lado, los datos clínicos y epidemiológicos, que están siendo un foco importante de investigación gracias a su creciente disponibilidad a través de los historiales clínicos electrónicos, y que por otra parte, su uso ha tenido un gran impacto en la investigación biomédica durante la pandemia de COVID-19. .

En cuanto a los datos de expresión génica, el trabajo realizado se ha centrado en el desarrollo de técnicas de meta-análisis para la integración de estudios independientes. En primer lugar, se llevó a cabo una revisión exhaustiva de las técnicas y del software disponibles, elaborándose un flujo de trabajo que cubre los principales pasos y análisis que se tienen que aplicar. Posteriormente, el estudio se enfocó en el problema de los genes faltantes, que ocurre cuando el número de genes no es el mismo en todos los estudios que se desean combinar, lo que puede resultar en la pérdida de información si solo se trabaja con los genes comunes en todos ellos. Esto llevó al desarrollo de una nueva herramienta

denominada DExMA, un paquete de R que implementa las funciones necesarias para aplicar todos los pasos del meta-análisis a datos de expresión génica. DExMA aborda el problema de los genes faltantes desde dos enfoques diferentes: considerando los genes presentes en al menos un porcentaje mínimo de estudios y mediante la imputación de genes a partir de los presentes en muestras de otros estudios con valores de expresión similares. Su aplicación a datos reales demostró como estos dos enfoques conservaban más información y proporcionaban mejores resultados que el procedimiento habitual de trabajar solo con los genes comunes.

Además, se exploró una nueva alternativa a estos enfoques: la combinación basada en la meta-datos de los genes, es decir, la integración de las rutas biológicas asociadas a ellos. Este enfoque pertenece al ámbito de las técnicas de meta-análisis de enriquecimiento anotaciones funcionales. Al estudiar estas técnicas y detectar varias debilidades en los métodos actuales, desarrollamos una nueva metodología llamada GSEMA. Esta metodología combina técnicas de enriquecimiento funcional de una sola muestra con técnicas de meta-análisis. Los resultados obtenidos al aplicar GSEMA a datos reales y simulados demostraron ser más consistentes que los obtenidos con métodos anteriores en este campo. En especial, GSEMA demostró ser altamente útil para combinar estudios con un alto número de genes faltantes, ya que las rutas biológicas significativas identificadas eran mucho más relevantes en las condiciones estudiadas en comparación con los resultados obtenidos con otros métodos.

En el área de análisis de datos clínicos epidemiológicos, la investigación doctoral se centró en el análisis de datos relacionados con COVID-19 y su posible estacionalidad. En primer lugar, se realizó una recopilación e integración de datos provenientes de fuentes heterogéneas para estandarizar y unificar información sobre la incidencia y evolución de la patología y su asociación con los factores ambientales. Esta integración de datos permitió el desarrollo de una aplicación web denominada DatAC, que proporciona información sobre el cambio de las variables relacionadas con el COVID-19 y factores ambientales para diferentes comunidades y provincias de España. La aplicación no solo ofrece el acceso público y centralizado a esta gran cantidad de datos, sino que facilita a los investigadores el análisis y visualización de los mismos gracias a los mapas y modelos implementados en ella.

Finalmente, aprovechando la información recopilada, se llevó a cabo un análisis sobre la relación entre los factores ambientales y la dispersión del virus, considerando el estado de inmunidad poblacional como un factor determinante en su posible estacionalidad. Mediante la aplicación de modelos específicos a los datos y un meta-análisis subsecuente para obtener un efecto global de las regiones, se concluyó que solo se observa cierta estacionalidad del virus en presencia de un alto porcentaje de población inmunizada. Este fenómeno resulta en un leve aumento en la transmisión del virus durante los períodos fríos y secos y a una reducción progresiva durante los períodos más cálidos.

Estos resultados. enfatizan la importancia de no solo integrar, sino también reutilizar de manera efectiva los datos generados, evitando que se conviertan en simples acúmulos de información sin valor analítico. Más allá de la simple recopilación, es necesario transformar estos datos en conocimiento útil, capaz de impulsar avances significativos en la investigación biomédica. Además, se resalta la necesidad de desarrollar y aplicar

técnicas estadísticas rigurosas, junto con software de código abierto que garantice la transparencia y la reproducibilidad de los estudios. De esta manera, se asegura que los resultados obtenidos no solo sean de alta calidad y fiables, sino también replicables y accesibles a la comunidad científica, contribuyendo así a un progreso sostenido y colaborativo en los campos diferentes campos de la investigación científica.

ABREVIATURAS

ACAT: *Aggregated Cauchy Association Test*

ADN: Ácido Desoxirribonucleico

ARN: Ácido Ribonucleico

ARNm: Ácido Ribonucleico mensajero

COVID-19: Enfermedad por Coronavirus 2019

DNA: *Deoxyribonucleic Acid*

FDAE: Función de Distribución Acumulativa Empírica

GEO: *Gene Expression Omnibus*

GSA: *Gene Set Analysis*

GSEA: *Gene Set Enrichment Analysis*

GSVA: *Gene Set Variation Analysis*

iGSEA: *Integrative Gene Set Enrichment Analysis*

INE: Instituto Nacional de Estadística

KNN: *K-Nearest Neighbors*

LES: Lupus Eritematoso Sistémico

MAPE: Meta-Analysis for Pathway Enrichment

MEA: Modelos de Efectos Aleatorios

MEF: Modelos de Efectos Fijos

NCBI: *National Center for Biotechnology Information*

NGS: *Next Generation Sequencing*

OR: *Odds Ratio*

ORA: Over-Representation Analysis

QuSAGE: Quantitative Set Analysis for Gene Expression

RENAVE: Red Nacional de Vigilancia Epidemiológica

RNA: *Ribonucleic Acid*

RNA-Seq: Secuenciación del RNA

RR: Riesgo de Relativo

scRNA-Seq: *Sincle cell RNA-Seq*

ssGSEA: *single sample Gene Set Enrichment Analysis*

VEN: Valor de Enriquecimiento Normalizado

Z-score: *Z-score Gene Set Enrichment Analysis*

1. INTRODUCCIÓN

1.1. La Estadística en la era del *Big Data*

Tradicionalmente, debido a la dificultad y al coste para acceder a todos elementos de una población, uno de los procesos fundamentales del campo de la Estadística es el muestreo poblacional, es decir, obtener muestras representativas de la población y realizar inferencias sobre ellas para obtener conclusiones y validar o rechazar las hipótesis iniciales sobre la población. Este proceso lejos de perder importancia con el paso del tiempo, sigue siendo un proceso relevante y de gran importancia en nuestra sociedad.

Sin embargo, en los últimos años, el auge y la expansión de las tecnologías ha generado un notable aumento en la cantidad de datos producidos y almacenados en bases de datos. Esta explosión de datos, junto con la creciente capacidad de procesamiento, presenta una oportunidad sin precedentes para el análisis y la adquisición de nuevos conocimientos en diversos campos del saber. Según un informe de Petroc Taylo[1] publicado en Statista[2], en el año 2020 se estimaba que el volumen de datos globales almacenados era de 64.2 zettabytes y con una predicción de crecimiento hasta los 181 zettabyes para el año 2025. Este momento histórico en el que nos encontramos actualmente es lo que llamamos era de los datos masivos o era del *Big Data*, término acuñado por Roger Mougalas, de O'Reilly Media en una entrevista en el *New York Times*[3].

En este nuevo escenario, surge un paradigma diferente: ya no solo es esencial obtener muestras representativas para corroborar nuestras hipótesis iniciales, sino también analizar la inmensa cantidad de datos disponibles para descubrir nuevos conocimientos y formular nuevas hipótesis. Esto plantea nuevos desafíos y oportunidades para la Estadística como ciencia y disciplina matemática, ya que su papel es fundamental para proporcionar los métodos y herramientas necesarios para examinar y analizar esta gran cantidad de información.

En este marco, algunas de los campos del conocimiento que ha experimentado grandes avances son las ciencias biológicas y de la salud. Por lo que la Bioestadística es una de las ramas de las Estadística que ha sufrido un mayor desarrollo en los últimos años. Al igual que en otras áreas de la Estadística, la investigación en bioestadística ha evolucionado para incluir métodos que permitan analizar e integrar conjuntos masivos de datos como el aprendizaje automático (o *machine learning*) o la minería de textos. Las técnicas bioestadísticas han sido cruciales en los últimos años, permitiendo por ejemplo la identificación de biomarcadores para diversas enfermedades al analizar grandes conjuntos de datos genómicos[4] o para gestionar la crisis sanitaria global producida por la pandemia producida por la enfermedad por Coronavirus 2019 (COVID-19)[5].

1.2. Integración de datos y *Data Fusion*

En este contexto de crecimiento masivo en la generación y acumulación de datos disponibles para su análisis, surge la necesidad de combinar diferentes fuentes de información para obtener una visión más completa y precisa. Aquí, es donde la integración de datos o *data integration* cobra relevancia. La integración de datos puede definirse como el proceso de combinar datos de múltiples fuentes heterogéneas en una estructura unificada[6,7]. Tradicionalmente, la integración de datos se centraba

principalmente en la creación de sistemas de almacenamiento de datos estructurados, conocidos como data warehouses[8], en los cuales se aplicaban métodos ETL (*Extract, Transform, Load*). En este proceso, los datos se extraían de diversas fuentes, se transformaban para cumplir con los estándares de calidad y se cargaban en almacenes de datos centralizados, donde eran consolidados y gestionados[8].

Sin embargo, estos enfoques presentan numerosas limitaciones al integrar los datos generados actualmente, debido al gran volumen de información en tiempo real y a la predominancia de datos no estructurados y muy heterogéneos, como textos libres e imágenes. Por este motivo, numerosos investigadores han propuesto nuevos métodos y herramientas para enfrentar estos desafíos. Algunos ejemplos de estos son la transformación de datos después de ser cargados, conocida como métodos ELT (*Extract, Load, Transform*), gracias a la capacidad de almacenamiento de datos en su forma original para análisis posteriores en sistemas de almacenamientos conocidos como *data lakes* o la aplicación de modelos conceptuales (*conceptual modeling*) y técnicas de aprendizaje automático (*machine learning*) para mejorar la integración y análisis de datos[9].

No obstante, en la actualidad, no solo es crucial almacenar y procesar datos para hacerlos más accesibles, sino también combinarlos para extraer nueva información. Generalmente, el proceso de integración de datos puede dividirse a grandes rangos en tres pasos[10]: en primer lugar, la coincidencia de esquemas o *schema matching*, en la que se afronta la búsqueda de correspondencias entre elementos o estructuras de las bases de datos. El segundo paso es la resolución de entidades o *entity resolution* donde se identifican los registros duplicados que representan la misma entidad. Y por último la fusión de datos o *data fusion,* donde se aplican técnicas estadísticas y algoritmos informáticos que persiguen la combinación y el análisis de los datos con el objetivo de obtener información a partir de los mismos[11].

En lo que respecta a la fusión de datos, existen tres tipos principales en función del tipo de combinación[12]:

1) Fusión a nivel de observación, la cual implica la combinación directa de los datos crudos. Un ejemplo sería la unificación de los datos recogidos por distintas estaciones meteorológicas como la temperatura, la humedad o la velocidad del viento en una sola base de datos conjunta.

2) Fusión a nivel de características que combina características o atributos derivados de datos crudos. Por ejemplo, a partir del historial de compras de los clientes de una empresa elaborar diferentes comportamientos como el número y categorías de los productos de compran y fusionar la información para obtener un perfil propio de cada cliente.

3) Fusión a nivel de decisión, la cual persigue combinar los resultados o decisiones obtenidas a partir de los datos para obtener un resultado final mucho más robusto. Un ejemplo de este tipo de fusión sería la combinación de los resultados de diferentes pruebas médicas para obtener una conclusión final acerca de la enfermedad de un paciente.

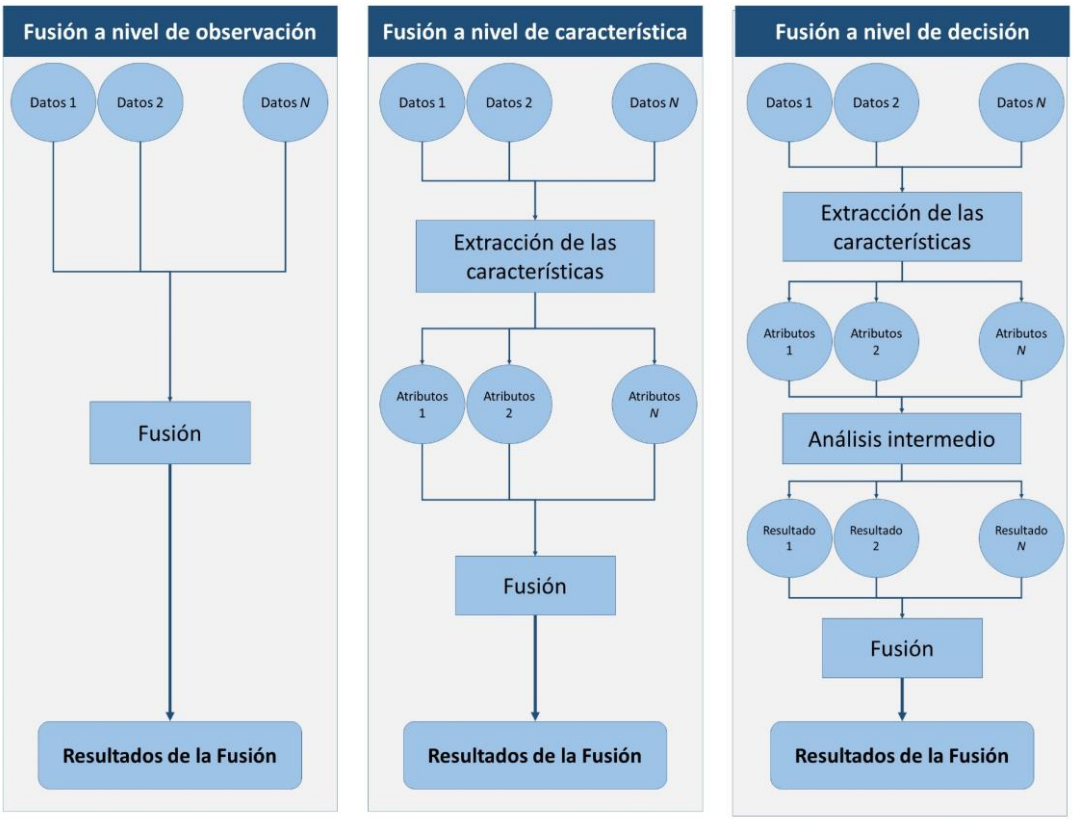


Figura 1. Tipos principales de la Fusión de Datos. La figura muestra los tipos principales de Fusión de Datos mostrando el nivel al que se realiza la fusión Adaptada de *Michael Schmitt y Xiao Xiang Zhu 2016*[12].

Además de esta clasificación general, se han propuesto numerosas clasificaciones más específicas, como la clasificación JDL[13] (*Joint directors of laboratories classification*) que divide las técnicas en cinco niveles en función del nivel de procesamiento de los datos o la clasificación de Dasarathy[14] que agrupa las técnicas en función de los datos de entrada y salida. En este contexto, se han propuesto numerosos métodos estadísticos y algoritmos informáticos que permitan combinar los distintos tipos de datos en cada uno de los niveles de información como modelos de aprendizaje automático o métodos de estadística bayesiana[15,16].

En conclusión, el objetivo final de la fusión de datos y, por ende, de la integración de datos es la obtención de nuevos resultados y conocimientos a partir de toda la información recopilada y almacenada.

Una técnica que ha ganado gran relevancia en este ámbito es el meta-análisis, cuyo objetivo es la integración de resultados de diversos estudios para obtener un resultado conjunto a ellos. Originalmente, se consideraba como un conjunto de métodos que tenía como finalidad la combinación de magnitudes resultantes de los análisis de estudios individuales. Sin embargo, en la actualidad, gracias a la disponibilidad de numerosas investigaciones en diferentes fuentes de información, los métodos del meta-análisis se han convertido además en unas técnicas de integración de datos enmarcadas dentro de los métodos de fusión de datos a nivel de decisión al combinar los resultados y decisiones obtenidas de los diferentes estudios.

1.3. El meta-análisis como técnica de integración de datos.

Aunque en las décadas de 1920 y 1930, algunos estadísticos como Ronald Fisher o Egon Pearson ya habían desarrollado métodos que permitían la combinación de los p-valores de diferentes test estadísticos[17–19], el término de meta-análisis fue acuñado en sus inicios por el estadístico americano Gene V. Glass en su artículo *"Primary, Secondary, and Meta-Analysis of Research"* publicado en 1976 en la revista *American Educational Research Association*[20]. En este artículo, Glass describe el meta-análisis como un análisis de los análisis, es decir, un método que permite *sintetizar los resultados de múltiples estudios de investigación*. Desde ese momento, estas técnicas comenzaron a ganar popularidad, principalmente debido a las mejoras en las computadoras y la capacidad de realizar un mayor número de análisis.

No obstante, fue a partir de finales de los años 90, con la definición de sus estándares junto con los de las revisiones sistemáticas, cuando el uso de estas técnicas inició un gran crecimiento en el campo de la investigación biomédica. En la actualidad, solo en la base de datos de PubMed[21], se han publicado más de 200,000 artículos que contienen la palabra meta-análisis en su título (Figura 2)

.

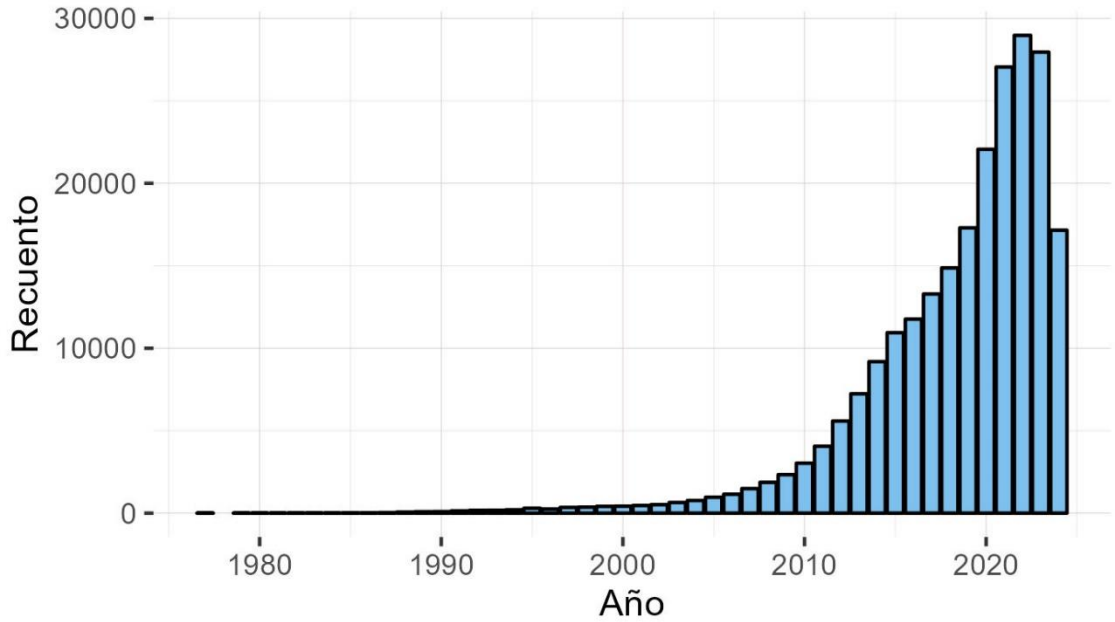


Figura 2. Número de artículos que contienen la *"meta-analysis"*. Diagrama de barras del número de artículos publicados cada año que contiene la palabra *"meta-analysis"*. Fuente de los datos PubMed (https://pubmed.ncbi.nlm.nih.gov/).

Como se ha indicado, uno de los motivos del amplio uso de estas técnicas es la aplicación conjunta del meta-análisis y las revisiones sistemáticas. Una revisión sistemática es un tipo de investigación que recopila y sintetiza de manera estructurada y metódica toda la información disponible sobre una pregunta de investigación científica. Se lleva a cabo siguiendo un protocolo preestablecido que define los criterios de selección, minimiza el sesgo y maximiza la reproducibilidad y la fiabilidad de los resultados[22]**.** Una vez realizada la revisión sistemática, el meta-análisis permite combinar los diferentes estudios y obtener una nueva evidencia científica a partir de ellos[23].

Este nuevo enfoque del meta-análisis para generar nuevos conocimientos a partir de la información contenida en diversas fuentes de datos ha transformado los métodos del meta-análisis, como ya hemos comentado anteriormente, en una técnica de integración de datos. Con una metodología estadística bien definida, esta aproximación facilita la generación de resultados nuevos, fiables y robustos.

1.3.1. Métodos de meta-análisis

Dentro del meta-análisis, existen diversos métodos para combinar los resultados y datos provenientes de distintos estudios. Cada uno de estos métodos tiene características específicas y requiere ciertas condiciones para ser aplicado de manera adecuada. Los diferentes métodos se pueden dividir en tres categorías principales: métodos basados en la combinación del tamaño del efecto, métodos basados en la combinación de p-valores, y los métodos basados en la combinación de rangos.

Métodos basados en la combinación de tamaños de efecto: estas técnicas tratan de explicar la magnitud de un fenómeno a lo largo de los diferentes estudios[24]. Para ello, en cada estudio se calcula una medida (denominada tamaño de efecto), cuya obtención depende de la naturaleza de los datos y del análisis que se está llevando a cabo. Por ejemplo, para el caso de estudios de tablas binarias se pueden calcular medidas de asociación como razones de productos cruzados (odds ratio) o riesgos relativos[24] mientras para datos continuos se pueden calcular correlaciones o estimadores de la diferencia de medias. Posteriormente, las diferentes medidas son combinadas mediante el uso de diferentes modelos.

Una técnica más avanza y complementaria a este tipo de método es la **meta-regresión.** Esta técnica extiende el enfoque de los métodos basados en la combinación de tamaño de efecto permitiendo también investigar cómo diferentes variables de los estudios (llamadas covariables) influyen en los tamaños de efecto individuales[24].

Métodos basados en la combinación de p-valores: este enfoque tiene como objetivo integrar los p-valores provenientes de los contrastes de hipótesis en los estudios individuales en un único p-valor que resuma la información general[25]. Para esta combinación se han desarrollado numerosas técnicas que utilizan diversos estadísticos y distribuciones basados en los p-valores de los estudios[25].

Métodos basados en la combinación de rangos: Estos métodos están diseñados únicamente para estudiar varias variables de la misma naturaleza, como por ejemplo el caso de los estudios de expresión génica en los que los genes son considerados cada uno como una variable. Primero, se ordenan los valores de cada variable en cada uno de los estudios. Luego, para cada variable, se calcula un estadístico mediante la combinación de los diferentes rangos, y finalmente, se obtiene un p-valor a partir de este estadístico[26].

Cada uno de estos métodos serán tratados en mayor profundidad y adaptados de manera específica para cada uno de los datos en estudio a lo largo de las diferentes secciones de la tesis doctoral.

1.3.2. Retos actuales de las técnicas del meta-análisis

A pesar de que la metodología del meta-análisis está claramente definida, su amplio uso ha dado lugar a diversos inconvenientes derivados de una aplicación inadecuada de la

misma. Algunos ejemplos de estas situaciones han sido descritos por el grupo del *Dr. Ioannidis*, quienes hacen referencia al aumento de estudios de meta-análisis con posibles limitaciones en cuanto a fiabilidad y reproducibilidad[27,28]. Asimismo, *Park et al*. han identificado errores analíticos frecuentes en un gran número de meta-análisis publicados en el campo del estudio de datos genéticos[29].

Además de estos problemas, los métodos de meta-análisis pueden enfrentarse a desafíos intrínsecos a la naturaleza de los datos en estudio, lo que complica la obtención de resultados fiables. Por ejemplo, la presencia de numerosos valores faltantes o la heterogeneidad significativa entre estudios[24], que puede surgir debido a diferencias en los diseños de estudio, las poblaciones analizadas, o las técnicas de medición utilizadas, lo cual puede llevar a resultados inconsistentes o sesgados.

Adicionalmente, la presencia de sesgos de publicación y de selección[24], donde solo los estudios con resultados positivos o significativos tienden a ser publicados, puede distorsionar los resultados del meta-análisis. Estos sesgos subrayan la necesidad de una evaluación crítica y sistemática de la literatura disponible para asegurar la integridad de los resultados obtenidos.

Esto evidencia la necesidad de contar con científicos expertos en estas técnicas que desarrollen herramientas y metodologías adecuadas para los diferentes tipos de datos y que ayuden a los investigadores a aplicar los métodos de una manera adecuada y a obtener resultados fiables y robustos. Además, es de gran importancia el desarrollo de guías que puedan ayudar a mitigar estos problemas, asegurando que los meta-análisis se realicen y reporten de manera transparente y reproducible.

En esta tesis doctoral se han aplicado de manera exhaustiva los diversos métodos del meta-análisis, adaptándolos específicamente a cada tipo de dato en estudio y considerando los problemas particulares asociados a cada uno. Todas estas características serán tratadas a lo largo de las diferentes secciones de la tesis doctoral.

2. OBJETIVOS

El objetivo principal de esta tesis doctoral es el desarrollo e implementación de técnicas de integración y meta-análisis de datos biomédicos. Para el cumplimiento de este objetivo, se proponen los siguientes objetivos específicos:

1. Revisión del estado de arte de las diferentes técnicas de meta-análisis aplicadas a datos de expresión génica, así como los paquetes de software disponibles en este contexto. Elaboración posterior de un flujo de trabajo y procedimientos necesarios para aplicar correctamente los diferentes pasos del meta-análisis que sirva como de guía de referencia para este tipo de estudios.
2. Desarrollo de un paquete de R que permita aplicar de una forma integral los diferentes pasos del meta-análisis de datos de expresión génica.
3. Caracterización del impacto de los valores faltantes en estudios de meta-análisis de expresión génica e implementación de técnicas que permitan su tratamiento.
4. Desarrollo de una nueva estrategia de meta-análisis basado en metadatos de los genes (rutas biológicas, conjuntos de genes relacionados entre sí…) como técnica alternativa al meta-análisis de datos de expresión génica para controlar la presencia de los valores faltantes.
5. Desarrollo de una aplicación de integración de datos para estudios epidemiológicos en la enfermedad de COVID-19 y aplicación de técnicas de meta-análisis para el análisis de la influencia de factores ambientales en la dispersión del virus

3. META-ANÁLISIS PARA INTEGRACIÓN DE ESTUDIOS DE EXPRESION

3.1. Conceptos previos

3.1.1. Datos de expresión génica

La investigación en Biomedicina ha experimentado un gran avance gracias al desarrollo de las técnicas de secuenciación masiva, también llamadas en inglés *Next generation sequencing* (NGS). Estas técnicas han reducido significativamente los costes y el tiempo de secuenciación de ácidos nucleicos con un gran número de aplicaciones. Un ejemplo destacado es la secuenciación del genoma humano, que originalmente tomó 13 años y requirió una inversión de 4800 millones de dólares[30,31]. En la actualidad este proceso cuesta significativamente menos y puede completarse en tan sólo unas pocas horas[32]. Este tipo de tecnologías tienen diversas aplicaciones para llevar a cabo estudios de sistemas biológicos a gran escala[33], las denominadas ciencias -omicas.

Los análisis ómicos abarcan distintas áreas, como la genómica que estudia la secuencia de ácidos desoxirribonucleicos (ADN o DNA por sus siglas en inglés), la transcriptómica, la cual estudia el ácido ribonucleico (ARN o RNA por sus siglas en inglés) o la proteómica que estudia las proteínas. Cada una de estas áreas proporciona una perspectiva multidimensional y global de los procesos biológicos, permitiendo a los investigadores obtener una compresión más completa de los diferentes sistemas biológicos (Figura 3). Dentro de las diversas disciplinas ómicas, la transcriptómica emerge como una de las más relevantes. La transcriptómica se especializa en el análisis de la expresión génica mediante la cuantificación de moléculas de ARN de una o varias muestras biológicas. La expresión génica es el proceso por el cual a partir de la información contenida en un gen se genera un producto funcional, como una proteína. La regulación de este proceso es fundamental para el funcionamiento de las células y por consiguiente del propio organismo. Una desregulación de la expresión génica, puede tener graves consecuencias para el organismo, llegando a producir desde trastornos en la biología celular hasta al desarrollo de enfermedades como el cáncer. La desregulación de la expresión génica puede manifestarse como sobreexpresión (el gen se expresa a un nivel excesivo respecto a un estado base) o infraexpresión (el gen se expresa a un nivel inferior respecto a un estado base).

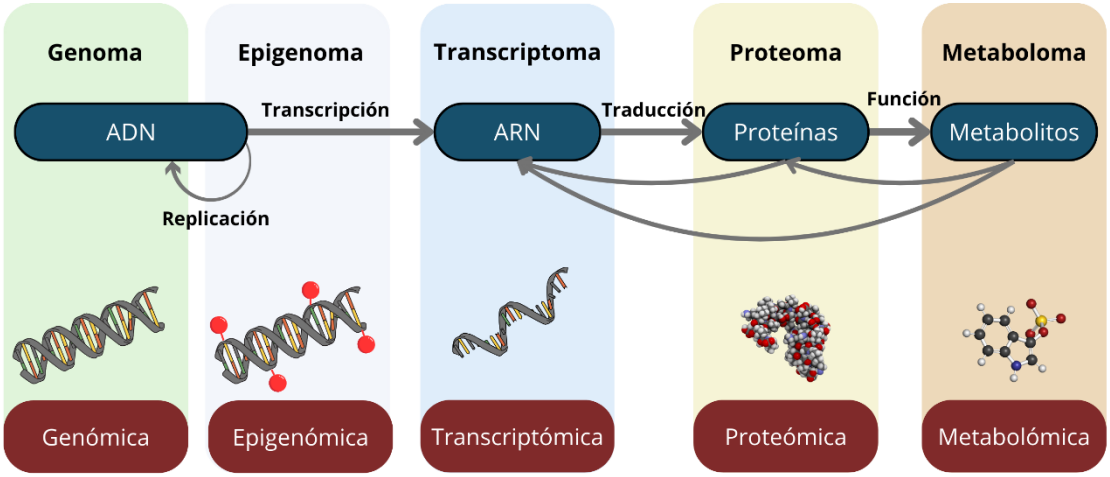


Figura 3: Ciencias -ómicas. Esquema de las los elementos y áreas que estudia cada una de las ciencias ómicas. Adaptado de Fundación Instituto Roche (2019). Informe "Anticipando Ciencias Ómicas". https://www.institutoroche.es/static/archivos/Informes_anticipando_CIENCIAS_OMICAS.pdf

Las tecnologías que obtienen los datos transcriptómicos y miden la expresión génica han ido evolucionando a lo largo de los años. En un comienzo, se hacía uso de las tecnologías de microarrays o DNA chip. En estas técnicas, la expresión de un gen se medía en base a la cantidad ADN copia (ADNc) que hibrida en sondas específicas en el chip. Posteriormente, estas serían reemplazadas por las tecnologías de NGS que con su aplicación para la secuenciación de RNA (RNA-Seq) permiten cuantificar la expresión del gen mediante secuenciación de las secuencias de ARN mensajero. Finalmente, la evolución de estas tecnologías ha permitido desarrollar la secuenciación de ARN de una sola célula o single-cell RNA-Seq (scRNA-Seq), que permite analizar la expresión génica a nivel de células individuales. Esto proporciona una resolución mucho más fina y permite detectar la heterogeneidad celular dentro de una muestra. En scRNA-Seq, se aíslan células individuales, se extrae el ARN de cada célula y se realiza la secuenciación.

A partir del procesamiento de los resultados obtenidos por estas tecnologías se obtienen los datos de expresión génica, los cuales indican el valor de expresión de gen en las diferentes muestras biológicas. Matemáticamente, estos datos se representan en una matriz de datos $M_{G\times N}$, siendo G el número de genes (variables) y N el número de muestras (o pacientes). Cada elemento de la matriz, m_{ij}, es el valor de expresión del gen i en la muestra j.

$$M_{G\times N} = \begin{pmatrix} m_{11} & \cdots & m_{1N} \\ \vdots & \ddots & \vdots \\ m_{G1} & \cdots & m_{GN} \end{pmatrix}$$

Los valores de las matrices de expresión varían en función de la tecnología aplicada para obtenerlos. En el caso de las tecnologías de *microarrays* los valores de expresión consisten en valores en escala continua. No obstante, para las tecnologías RNA-Seq al realizar un conteo de las secuencias del ARNm, los valores consisten datos cuantitativos discretos. Respecto a la técnica scRNA-Seq, los valores obtenidos son también cuantitativos discretos., pero a nivel de célula única, es decir, en las matrices de expresión existe una dimensión más que corresponde a la célula individual (matriz de 3

dimensiones). En la práctica, como parte del análisis se suelen procesar los datos para obtener matrices de expresión por cada tipo celular, de manera que las matrices sean similares a las obtenidas en RNA-Seq. A este procedimiento se le denomina pseudo-bulk RNA-Seq. La existencia de diferentes tecnologías y, por lo tanto, de diferentes matrices de expresión lleva a que se requiera de un pre-procesamiento y flujo de trabajo específico en función de la tecnología empleada para la obtención de los datos. En esta tesis doctoral se han analizado principalmente matrices de expresión tanto de microarrays como de RNA-Seq.

Además, desde el punto de vista de la Estadística, los datos de expresión génica poseen una particularidad única: a diferencia de otros contextos, el número de variables (genes) es considerablemente mayor que el número de muestras (o pacientes). Esta disparidad se debe a la propia naturaleza de los datos, ya que el número de genes (miles) supera ampliamente el número de muestras secuenciadas (cientos). A parte de este atributo estos datos también se caracterizan por:

- Presentar algunas muestras con una alta correlación entre sí.
- La gran parte de la variabilidad de los estudios puede estar explicada por un pequeño conjunto de genes
- Los genes presentan varias interacciones entre sí, como por ejemplo las agrupaciones en conjuntos funcionales, lo que puede producir que la expresión de gen esté influencia por otros genes e influencie a su vez la expresión de otros.
- En algunas de las matrices la mayoría de los elementos pueden ser cero o carecer de valor. Este concepto se denomina esparsidad (*sparsity* en inglés) y en estos datos es debido principalmente a la propia tecnología de secuenciación o a la baja cantidad de material genético a la hora de procesar las muestras.

3.1.2. Bases de datos Públicas de datos de expresión

Gran parte de los datos de expresión génica generados con las tecnologías de NGS o microarrays son almacenados en repositorios públicos debido que la mayor parte de las revistas científicas requieren el acceso abierto de los datos originales como parte del proceso de publicación. Esto ha provocado un crecimiento exponencial en la cantidad de estudios e información almacenados en estas bases de datos (ver Figura 4). Por lo tanto, estos datos almacenados se convierten en una fuente de información sin precedentes para poder llevar a cabo nuevos análisis y generar resultados más allá de los obtenidos en los estudios originales. Su integración con otras fuentes o el análisis integrado con otros estudios nos permite generar nuevo conocimiento u obtener resultados más consistentes y robustos.

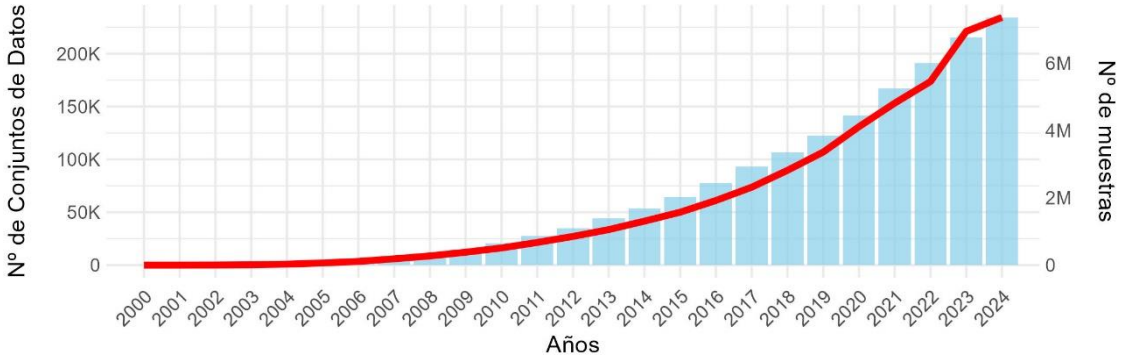


Figura 4: Evolución del número de conjuntos de datos y de muestras almacenados en la base de datos de GEO desde el año 2000 hasta junio de 2024. El diagrama de barras azul indica el número de conjuntos de datos almacenados (eje Y izquierdo) y la línea roja el número de muestras almacenadas (eje Y derecho). Los datos usados para generar el gráfico fueron extraídos de la página web de GEO *summary* (https://www.ncbi.nlm.nih.gov/geo/summary/?type=history).

Actualmente, existen varios repositorios de datos públicos como, por ejemplo, *ArraysExpress*[34] mantenida por el European Bioinformatics Institute (EMBL_EBI) y que proporciona acceso a estudios de microarrays y RNA-Seq de diferentes organismos o *Sequence Read Archive*[35] (SRA) desarrollada por el *National Center for Biotechnology Information* (NCBI) y que contiene datos crudos de secuenciación de RNA-Seq.

Una de las bases de datos públicas más populares es la Gene Expression Omnibus (GEO)[35–37], que al igual que SRA fue desarrollada y es mantenida por el NCBI. Aunque en sus inicios GEO fue creado con la intención de almacenar datos procedentes de estudios desarrollados mediante tecnologías de *microarrays*, aunque en los últimos años se adaptó para que contuviera además datos de expresión generados mediante tecnologías de RNA-Seq. Además, a diferencia de SRA contiene los datos transcriptómicos ya procesados (datos de expresión), teniendo un enlace a la base de datos de SRA para los datos crudos del estudio. Con el objetivo de organizar esta gran cantidad de información y permitir a los investigadores un mejor acceso a la misma, en GEO existen una serie de códigos identificadores. En concreto, se asignan códigos para cada conjunto de datos (código GSE), para cada muestra (código GSM) y por cada plataforma de secuenciación (código GPL). Gracias a esto sabemos que a principios de junio de 2024 había almacenados 229,343 conjuntos de datos, con 7,219,580 muestras y que habían sido generados por 26,157 plataformas de secuenciación diferentes. Además, el equipo de GEO procesa, cura y anota algunos de los conjuntos de datos GSE y con ellos generan otros conjuntos de datos que identifican con el código GDS.

3.1.3. El análisis de expresión diferencial

Como se mencionó anteriormente, una alteración en la actividad de un gen puede tener consecuencias significativas en las células y, en última instancia, en los organismos. Estas alteraciones pueden provocar desde disfunciones celulares hasta enfermedades[38]. Por lo tanto, es fundamental conocer los niveles de expresión de los genes para desarrollar nuevas estrategias de diagnóstico, tratamiento y prevención de enfermedades.

El análisis de expresión diferencial es la técnica usada para identificar los genes que se expresan de una manera diferente entre dos o más grupos de muestra[39]. Habitualmente, la expresión diferencial suele aplicarse entre un grupo de pacientes (grupo experimental)

y un grupo de control (sanos o con una condición distinta)[40]. Estadísticamente consiste en aplicar un test de diferencias de medias entre dos grupos. Para ello se hacen uso de test estadísticos de diferencias de medias y de software informático desarrollado para tales fines. Además, es habitual realizar una corrección por comparaciones múltiples a los p-valores obtenidos para cada uno de los genes, utilizando métodos como la corrección de *Bonferroni*[41] o el método de *Benjamini & Hochberg*[42].

El primer paso de este análisis es el preprocesamiento y la normalización de los datos, con el objetivo de reducir la variabilidad inherente a las propias tecnologías[43] o a la propia influencia de los investigadores que realizaron el experimento (*llamado efecto de lote o Batch effect* en inglés)[44].

Una vez tratados los datos se aplica el test estadístico para analizar la expresión diferencial entre los distintos grupos. Para ello, desde la aparición de los datos de microarrays se desarrollaron numerosas metodologías, estando entre las más utilizadas los modelos lineales y métodos de Bayes empíricos implementados en paquete de R *limma*[45].

Posteriormente, con la llegada de los datos de expresión procedentes de los estudios de RNA-Seq, fue necesario tratar estos datos desde una perspectiva diferente. Al ser datos cuantitativos discretos algunos autores propusieron el uso de test basados en distribuciones de datos discretos como distribuciones binomiales negativas[46,47] o de Poisson[48]. Estos métodos fueron implementados en algunos de los paquetes de R más usados para tratar este tipo de datos como son *DESeq2*[49] o *edgeR*[47,50]. En el caso de *limma* también fue adaptado para tratar este de datos, haciendo uso de normalizaciones para adaptar los datos de cuantitativos discretos a los datos cuantitativos continuos admitidos por su metodología[45,51]. Además de estos métodos se desarrollaron otras metodologías no paramétricas como el método NOISeq[52], la cual tiene su metodología implementada en el paquete de R *NOISeq*[52,53] o el método SAMseq[54]. En términos generales, los métodos no paramétricos son más adecuados cuando en los datos no se puede asumir que provengan de una distribución específica o que siguen distribuciones mucho más complejas[40]. Por otro lado, algunos autores puntualizan que los métodos no paramétricos solo obtienen resultados fiables cuando los tamaños muestrales son grandes (algo menos habitual en estudios de RNA-Seq), siendo los métodos paramétricos mucho más eficientes para tamaños muestrales más pequeños[40,55].

3.1.4. Meta-análisis de expresión génica y aplicaciones

En el contexto del análisis de expresión diferencial, las ventajas significativas que presenta el meta-análisis para combinar resultados de diferentes investigaciones han llevado a un aumento considerable en el número de publicaciones que aplican estas técnicas a este tipo de datos en los últimos años. En este campo, a este tipo de meta-análisis se le denominan meta-análisis de expresión génica (*gene expression meta-analysis en inglés*) o meta-análisis de expresión diferencial (*differential expression meta-analysis*). El objetivo principal es la combinación de los resultados de los estudios individuales de expresión diferencial para obtener los genes diferencialmente expresados a lo largo de los diferentes estudios.

Específicamente, el meta-análisis de expresión génica se puede aplicar en un amplio espectro de situaciones, las cuales se pueden resumir en tres tipos de aplicaciones principales (ver Figura 5):

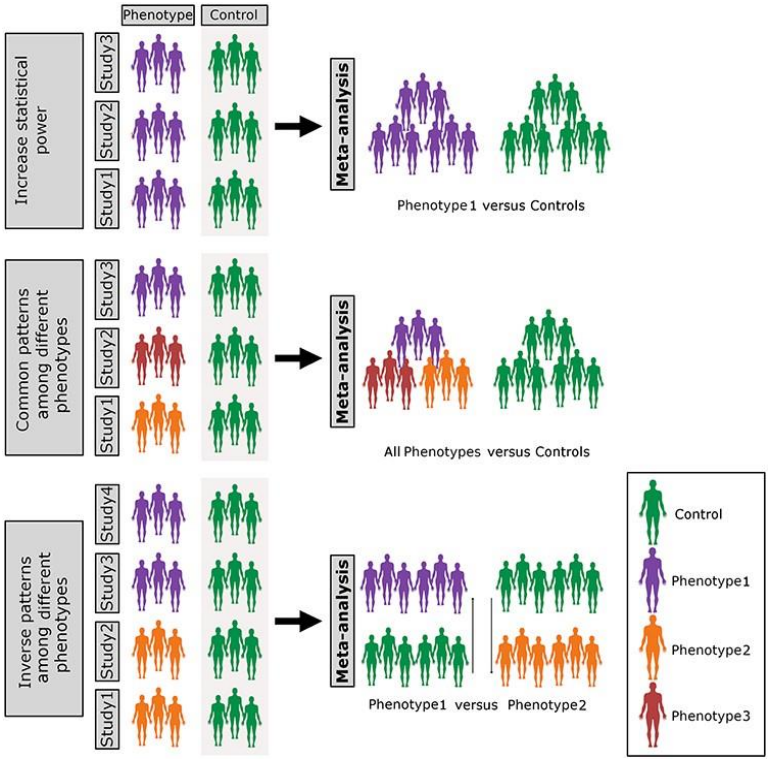


Figure 5. Aplicaciones del meta-análisis de expresión génica. La figura muestra las tres principales aplicaciones u objetivos del meta-análisis de expresión génica. El aumento de la potencia estadística, la búsqueda de patrones comunes o inversos entre distintas condiciones. Figura extraída del artículo de *Toro-Domínguez, Villatoro-García et al.*[56]

Incremento del tamaño muestral y la potencia estadística. La integración de estudios o cohortes que comparten una misma condición o fenotipo permite aumentar significativamente el tamaño muestral y, en consecuencia, la potencia estadística para detectar genes con diferencias relevantes entre grupos de casos y controles. Esta estrategia es especialmente útil para identificar biomarcadores consistentes y robustos asociados con una determinada condición. Ejemplos de esta aplicación pueden observarse en estudios de cáncer[57,58], enfermedades autoinmunes[59,60] o trastornos mentales[61].

Búsqueda de patrones comunes entre diferentes condiciones. Esta aplicación consiste en buscar patrones comunes de expresión génica que compartan diferentes enfermedades o fenotipos, como conjuntos genes diferencialmente expresados entre distintas enfermedades respecto a muestras sanas o diferentes tratamientos farmacológicos, lo que resulta útil para identificar rutas biológicas o mecanismos moleculares compartidos entre diferentes condiciones. Algunos ejemplos de esta aplicación pueden encontrarse en estudios de enfermedades autoinmunes[62–64] o trastornos neuronales[65] en los cuales se desean buscar patrones comunes de expresión génica en enfermedades con manifestaciones clínicas similares.

Búsqueda de patrones inversos entre diferentes condiciones. Esta última aplicación se basa en integrar estudios de expresión génica de diferentes enfermedades o fenotipos con el objetivo de identificar patrones de expresión inversos, es decir, identificar genes que en una condición se encuentren sobreexpresados respecto a controles y simultáneamente

infraexpresados en la otra respecto al grupo de control. Ejemplos de esto es su uso en enfermedades con comorbilidades inversas, como en el caso del Alzheimer y el cáncer[66], para buscar estos patrones inversos también a nivel de expresión génica. Además, también podría aplicarse a la hora de reutilizar fármacos, debido a la hipótesis de que, si un fármaco produce patrones de expresión genética inversos a los de la enfermedad, esto podría usarse como tratamiento de la misma[67,68].

Aunque el objetivo de las tres aplicaciones sea distinto, no quiere decir que cualquiera de los métodos descritos sea específico de alguna de ellas. Un análisis más profundo de estas aplicaciones, así como las consideraciones específicas en cada una de ellas son tratadas en *Toro-Domínguez y Villatoro-García et al*[56] (proporcionado en el Anexo 9.1).

3.1.5. Tratamiento de datos faltantes en estudios de meta-análisis de expresión génica

A pesar de la gran utilidad de las técnicas de meta-análisis de expresión génica, su amplio uso ha llevado al surgimiento de determinados problemas y controversias, puestas de manifiesto por diferentes investigadores[28,29]. Uno de los problemas detectados en este tipo de análisis es el impacto de los genes no medidos o genes faltantes (*missing genes*)[69,70]. En este caso, este problema es más complejo que el problema de los valores faltantes (*missing values*) tradicional, ya que aparece cuando se desea combinar estudios con distinto número de genes, es decir, en los distintos estudios no se han medido los mismos genes. En el problema de los valores faltantes tradicional, en determinadas variables existen elementos concretos ausentes y que habitualmente son estimados mediante técnica de imputación. Esto ocurre, por ejemplo, en un análisis de expresión diferencial de un estudio cuando para algún gen no se tiene el valor de expresión diferencial de una de las muestras y el cual se suele imputar a partir del resto de valores[71].

Sin embargo, en el caso de los genes faltantes del meta-análisis esto implica que para algunos de los estudios existen variables completas que no contienen ningún valor de expresión para ninguna de las muestras, por lo que hay una gran falta de información. (Figura 6). Esto ocurre con mucha frecuencia a la hora de analizar estudios procedentes de bases de datos públicas en los que se han usado diferentes plataformas de secuenciación, por lo que los genes medidos difieren de uno a otros estudios. Este problema ha sido poco caracterizado a lo largo de la literatura, analizándose en la mayoría de casos sólo los genes comunes a todos los estudios y descartando el resto, lo que puede producir una pérdida de información relevante[69].

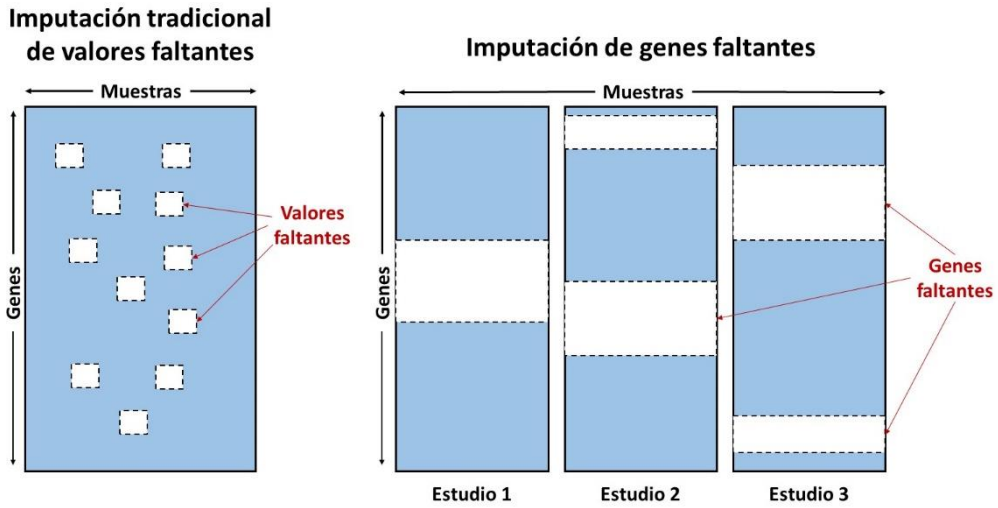


Figura 6: Esquema de la diferencia entre Imputación tradicional de valores faltantes y la Imputación de genes faltantes. En la imputación tradicional de valores faltantes se imputan los valores de expresión génica de un único conjunto de datos. En el problema de la imputación de genes faltantes, se imputan los genes no medidos y que no están presentes en ninguna muestra del conjunto de datos. Figura del artículo de *Mancuso et al., 2020*[70].

No obstante, en los últimos años, se han comenzado a aplicar técnicas de imputación de genes no medidos a partir de la información de estos estudios que si los contienen[70,72]. Estas técnicas tal y como describe *Bobak et al.*[69] pueden ser usadas en el meta-análisis de expresión génica para imputar los genes faltantes de un estudio a partir de los otros estudios incluidos en el meta-análisis.

3.2. Metodología

El meta-análisis de expresión génica aplica las técnicas propias del meta-análisis a los datos de expresión génica. Por lo tanto, como ya se ha mencionado anteriormente, es una extensión del análisis de expresión diferencial con el propósito de identificar los genes diferencialmente expresados a lo largo de los estudios. En esta sección vamos a describir los diferentes métodos del meta-análisis de expresión génica que se han implementado, los cuales han sido detallados y aplicados en dos de los artículos de esta tesis doctoral, *Toro-Domínguez y Villatoro-García et al.*[56] y *Villatoro-García et al.*[73], los cuales están proporcionados en el Anexo 9.1 y Anexo 9.2 respectivamente.

3.2.1. Meta-análisis basado en tamaño de efectos

El meta-análisis basado en tamaños de efecto tiene como objetivo la obtención de una magnitud combinada a partir de las diferentes medidas obtenidas en los distintos conjuntos de datos. A esta medida es lo que denominamos tamaño de efecto y su forma de cálculo varía en función del tipo de análisis que se esté llevando a cabo[24]. En el meta-análisis de expresión génica este tamaño de efecto es calculado para cada uno de los genes en estudio.

3.2.1.1. Cálculo del tamaño de efecto

El tamaño de efecto calculado en cada estudio depende del tipo de datos (continuos o discretos) y del análisis que se esté realizando. Por ejemplo, para tablas binarias o datos categóricos se suele trabajar con efectos basados en medidas de asociación como la razón

de producto cruzado u *odds ratio* en inglés (*OR*) y el riesgo relativo o en inglés *risk ratio* (*RR*).

En el caso específico de los datos de expresión génica, tanto los datos microarrays como los datos de RNA-Seq tras su normalización y estandarización de la matriz de expresión, son datos cuantitativos continuos en los que se desea comparar dos grupos de diferentes tamaños entre sí (grupo experimental y grupo de referencia o control). Por lo tanto, para estimar el tamaño de efecto, lo más adecuado es calcular estimadores de los **tamaños de efectos basados en la diferencia de medias**.

Para estimar la diferencia la diferencia de medias hay varios estimadores que se han usado en el contexto de estudios de meta-análisis que se han implementado para esta tesis.

Uno de los más ampliamente recomendados es la *g de Hedges*[74,75], la deriva a su vez del estimador de la *d de Cohen*[76]. En el caso particular de los datos de expresión génica, dados *K* estudios, la *d de Cohen* para el gen *i* en el estudio *j*, d_{ij}, se calcularía como:

$$d_{ij} = \frac{\overline{X_{ij_E}} - \overline{X_{ij_C}}}{S_{ij_{conjunta}}} \tag{3.1}$$

donde:

- $\overline{X_{ij_E}}$ y $\overline{X_{ij_C}}$ son las medias muestrales del gen *i* en el estudio *j* de cada uno de los grupos respectivamente
- $S_{ij_{conjunta}} = \sqrt{\frac{\left(n_{j_E}-1\right)S_{ij_E}^2 + \left(n_{j_C}-1\right)S_{ij_C}^2}{\left(n_{j_E}+n_{j_C}-2\right)}}$, es la desviación típica conjunta del gen *i* del estudio *j* de ambos grupos en el gen , siendo n_{J_E} y n_{j_C} los tamaños muestrales del cada uno de los grupos en el estudio *j* y $S_{ij_E}^2$ y $S_{ij_C}^2$ las varianzas muestrales de cada uno de los grupos en el estudio *j* para el gen *i*.

Además, la varianza de d_{ij} se puede aproximar mediante:

$$V_{d_{ij}} = \frac{n_{j_E}+n_{j_C}}{n_{j_E}\times n_{j_C}} + \frac{d_{ij}^2}{2\left(n_{j_E}+n_{j_C}\right)} \tag{3.2}$$

El estimador de la *g de Hedges* aplica una corrección del sesgo que tiene asociado la *d de Cohen*, debido a que tiende a sobreestimar la diferencia de medias estandariza poblacional en muestras con tamaños pequeños[24,75]. Aunque *Hedges* proporcionó la fórmula exacta de este factor de corrección, en la práctica se suele aproximar[24].

$$J = \left(1 - \frac{3}{4\times df - 1}\right) \tag{3.3}$$

Por lo tanto, *la g de Hedges* se puede definir como:

$$g_{ij} = J_j \times d_{ij} \tag{3.4}$$

Siendo J_j el factor de corrección para el estudio *j* y *df* los grados de libertad asociados a la $S_{conjunta}$, que en este caso son $n_{j_E} + n_{j_C} - 2$. Teniendo en cuenta este factor de corrección podemos aproximar también la varianza de la *g de Hedges*:

$$V_{g_{ij}} = J_j^{\,2} \times V_{d_{ij}} \tag{3.5}$$

Por definición el factor de corrección es siempre menor o igual a 1. Para tamaños muestrales grandes tiende a aproximarse a 1, lo que hace que, en estos casos, la diferencia entre la *g de Hedges* y *d de Cohen* sea generalmente insignificante.

Sin embargo, a lo largo de la literatura, varios autores han señalado que el uso de la *g de Hedges* puede introducir un sesgo que tiende a estimar de manera inexacta la varianza del efecto combinado[77,78]. En un meta-análisis estándar, donde sólo se combina el tamaño de efecto de varios estudios, tales sesgos pueden no afectar de forma significativa a los resultados. No obstante, en el contexto del meta-análisis de expresión génica, en el que se evalúan múltiples variables en cada estudio, este sesgo puede contribuir a aumentar las tasas de falsos positivos y negativos. Esto es debido en gran medida a que tamaños de efectos pequeños generan a su vez varianzas pequeñas lo que puede dar lugar a rutas significativas. De manera inversa, tamaños de efectos grandes, dan lugar a varianzas grandes y por consiguiente a una falta de significación estadística. Para corregir este sesgo se han propuesto otros cálculos alternativos de la estimación de la varianza, como el propuesto por *Lin et al*[77] que se basa en la media de los diferentes estimadores de la varianza. En el caso del meta-análisis de expresión génica se adaptaría como:

$$V_{g_{ij}} = \frac{1}{n_{j_E}} + \frac{1}{n_{j_C}} + \frac{\overline{g_i}^{\,2}}{2\left(n_{j_E} \times n_{j_C}\right)} \tag{3.6}$$

Donde $\overline{g_i}$ es la media de los estimadores de la *g de Hedges* para el gen *i*.

Aunque *Lin et al*[77] analizan otros estimadores en su estudio, el resto produce resultados similares a (3.6), por lo que en esta tesis doctoral ésta ha sido la corrección de la varianza considerada.

Por otra parte, en el análisis de expresión diferencial se han desarrollado numerosos métodos que permiten un análisis más preciso de la diferencia de medias con el objetivo de reducir el número de falsos positivos debido a diferencias pequeñas con varianzas reducidas. En particular, uno de los métodos más ampliamente aplicados es el test de la *t de Student moderada*[79], el cual se encuentra implementado en el paquete de R de *limma*[45,80]. A diferencia del estadístico *t de Student* tradicional, el estadístico *t de Student* moderada calcula la varianza de las diferentes variables (genes) considerando la información del resto de las variables, lo que permite aumentar y corregir las varianzas más pequeñas.

En nuestro caso, decidimos implementar la *g de Hedges* basada en este estadístico *t de Student moderada.* Para ello, en primer lugar, se obtienen los valores de la *t de Student moderada* y sus grados de libertad para cada uno de los genes en cada uno de los estudios. Posteriormente, se obtiene la *d de Cohen* correspondiente a cada uno de los valores de la

t de Student moderada y sus grados de libertad a partir de la demostración de *Rosenthal and Rosnow*[81]:

$$d_{ij} = \frac{\left(n_{j_E} + n_{j_C}\right) \times t_{ij}}{\sqrt{n_{j_E} \times n_{j_C}} \times \sqrt{df_{ij}}} \tag{3.7}$$

Siendo t_{ij} el estadístico de la *t de Student* moderada obtenida en el estudio *j* para el gen *i*, df_{ij} los correspondientes grados de libertad de t_{ij}, n_{j_E} y n_{j_C} los tamaños muestrales de los grupos experimental y de control respectivamente . Finalmente, a partir de las diferentes d_{ij} obtenemos los estimadores de la *g de Hedges* para cada gen en cada estudio. Este tamaño de efecto es comparable al propuesto por *Marot et al.*[82], con la diferencia de que, en este caso, se consideran los grados de libertad ajustados por *limma*, mientras que en su estudio solo se calcula la *d de Cohen* a partir del estadístico de la *t de Student moderada* y los correspondientes tamaños muestrales.

Al trabajar con *limma*, a la hora de trabajar con datos de RNA-Seq se deben procesar los datos previamente para aplicar este paquete a este tipo de datos tal y como recomiendan sus autores[83,84]. Esto implica que los recuentos de las matrices de expresión se convierten a $\log_2$ recuentos por millón (logCPM) y la relación media-varianza se modela con pesos de precisión (denominado *voom*)[51] o con una tendencia empírica Bayes a priori (denominado *lima-trend)*[51].

Una vez que se han calculado en cada estudio los diversos tamaños de efecto para gen, es posible determinar **el efecto combinado** de cada uno de ellos. Este valor se obtiene como la media ponderada de los distintos tamaños de efecto de cada gen en cada uno de los estudios incluidos en el análisis. Específicamente, existen dos modelos para asignar los pesos a cada estudio y luego calcular el tamaño de efecto combinado: el modelo de efectos fijos y el modelo de efectos aleatorios, los cuales se basan en diferentes suposiciones.

3.2.1.2. El modelo de efectos fijos (MEF)

El modelo de efectos fijos (MEF) o en inglés *Fixed-effect model* supone que todos los estudios comparten un tamaño efecto verdadero o común (θ) a todos ellos, es decir, que todos los factores que influyen en el tamaño de efecto son iguales en todos los estudios. Esto implica que estudios con mucha información suelen recibir un mayor pesos respecto a los estudios que contiene una menor cantidad de información[24].

En este modelo, asumimos que dado *K* estudios, los diferentes tamaños de efectos observados se distribuyen según una normal de media θ (el tamaño de efecto común) y varianza σ^2. Por lo tanto, el tamaño de efecto (θ_j) del *j*-ésimo estudio está determinado por:

$$\theta_j = \theta + \varepsilon_j \tag{3.8}$$

Donde ε_j es el error de muestreo asociado al *j*-ésimo estudio. Se asumen que $\varepsilon_j \sim N(0, V_{\theta_j})$, siendo V_{θ_j} la varianza del tamaño de efecto del *j*-ésimo estudio.

Esto implica que en el modelo MEF, el cálculo efecto combinado es una estimación del tamaño de efecto común, θ. Para la obtención del efecto combinado a cada estudio se asigna como peso (ω) el inverso de la varianza del tamaño de efecto estimado de cada uno de los estudios $\left(V_{\theta_j}\right)$.

En el caso específico del meta-análisis de expresión génica si consideramos que el tamaño de efecto del gen *i* en el estudio *j*, θ_{ij}, como el estimador de la *g de Hedges*, entonces:

$$\theta_{ji} = g_{ij} \tag{3.9}$$

$$V_{\theta_{ij}} = V_{g_{ij}} \tag{3.10}$$

Por lo tanto, el peso asignado al estudio *j* para el gen *i* se calcula como:

$$\omega_{ij} = \frac{1}{V_{\theta_{ij}}} \tag{3.11}$$

En consecuencia, el efecto combinado del gen *i* $\left(\theta_i\right)$ para un número *K* de estudios se calcula como:

$$\theta_i = \frac{\sum_{j=1}^{K} \omega_{ij}\theta_{ij}}{\sum_{j=1}^{K} \omega_{ij}} \tag{3.12}$$

El inverso de la varianza de θ_i se obtiene como el inverso de la suma de los pesos, es decir:

$$V_{\theta_i} = \frac{1}{\sum_{j=1}^{K} \omega_{ij}} \tag{3.13}$$

A partir de esta varianza podemos calcular además los intervalos de confianza para el efecto para el efecto combinado. Los límites inferior (LI_i) y superior (LS_i) se calcularían como

$$LI_i = \theta_i - z_\alpha \times \sqrt{V_{\theta_i}} \tag{3.14}$$

$$LS_i = \theta_i + z_\alpha \times \sqrt{V_{\theta_i}} \tag{3.15}$$

Donde z_α es el valor de una Normal estándar, $N(0,1)$, a un nivel de confianza α.

Por otro lado, si consideramos que un gen estará diferencialmente expresado cuando su efecto combinado sea igual a 0, entonces tendríamos el siguiente contraste de hipótesis:

$$\begin{cases} H_0: \theta_i = 0 \\ H_1: \theta_i \neq 0 \end{cases}$$

El valor del estadístico Z_i para el contraste de hipótesis se calcula como:

$$Z_i = \frac{\theta_i}{\sqrt{V_{\theta_i}}} \tag{3.16}$$

Por lo que el p-valor a dos colas se obtiene como:

$$p_i = 2\left[1 - \Phi\left(|Z_i|\right)\right] \tag{3.17}$$

Siendo $\Phi\left(Z_i\right)$ la función de distribución de una normal estándar.

3.2.1.3. El modelo de efectos aleatorios (MEA)

El modelo de efectos aleatorios (MEA) o en inglés Random-Effect Model a diferencia del modelo MEF, no asume la existencia de un tamaño efecto común a todos los estudios, sino que supone que el efecto verdadero varía de unos estudios a otros y, por ende, una distribución de los tamaños de efecto verdadero. Esto implica que hay una nueva fuente error en los tamaños de efectos de los estudios, es decir, para el tamaño de efecto del *j*-ésimo estudio está determinado por:

$$\theta_j = \theta + \zeta_j + \varepsilon_j \tag{3.18}$$

donde $\zeta_j \sim N\left(0, \mathrm{T}^2\right)$ es la verdadera variación de los tamaños de efecto, siendo T^2 la varianza entre los estudios (inter-estudio).

Esta nueva fuente de variación implica que en el modelo MEA el cálculo del efecto combinado es una estimación de la media de los tamaños efectos verdaderos. El peso asignado a cada estudio es la varianza de cada uno de los estudios (V_{θ_j}), la cual se calcula como la suma de la varianza intra-estudio y que coincide con la varianza del tamaño del efecto del estudio, y la varianza inter-estudio que denominamos T^2.

Para estimar T^2 uno de los métodos más comunes es el método de los momentos o también llamado método de *DerSimonian and Laird*[85]. Este método obtiene un estimador τ^2 a partir de la varianza total de los *K* estudios considerados. En el caso concreto del meta-análisis de datos de expresión génica para el gen *i* su estimador de la varianza inter-estudio, τ_i^2 se obtendría como:

$$\tau_i^2 = \frac{Q_i - df}{C_i} \tag{3.19}$$

donde:

Q es la varianza total de los estudios que se obtiene como[24]:

$$Q_i = \sum_{j=1}^{K} \omega_{ij} \left(\theta_{ij} - \theta_i \right)^2 \tag{3.20}$$

Y simplificando:

$$Q_i = \sum_{j=1}^{K} \omega_{ij} \theta_{ij}^2 - \frac{\left(\sum_{j=1}^{K} \omega_{ij} \theta_j \right)^2}{\sum_{j=1}^{K} \omega_{ij}} \tag{3.21}$$

df (*degrees of freedom)* son los grados de libertad siendo *K* el número de estudios:

$$df = K - 1 \tag{3.22}$$

Y *C* es un factor de ajuste que se calcula como:

$$C = \sum_{j=1}^{K} \omega_{ij} - \frac{\sum_{j=1}^{K} \omega_{ij}^2}{\sum_{j=1}^{K} \omega_j} \tag{3.23}$$

En el caso de que en la ecuación (3.19) se obtenga un valor negativo, τ_i^2 toma el valor de 0. Aunque el método de *DerSimonian and Laird* es uno de los más comúnmente aplicados para estimar la varianza inter-estudio existen otros métodos de gran relevancia como el estimador de Bayes empírico[86] o el estimador restringido de máxima verosimilitud[87].

Una vez calculada la estimación de la varianza inter-estudio, se asignan los pesos como:

$$\omega_{ij}^* = \frac{1}{V_{\theta_{ij}} + \tau_i^2} \tag{3.24}$$

Por lo tanto, del mismo modo que en el modelo MEF, se calcula el efecto combinado como:

$$\theta_i^* = \frac{\sum_{j=1}^{K} \omega_{ij}^* \theta_{ij}}{\sum_{j=1}^{K} \omega_{ij}^*} \tag{3.25}$$

La varianza del efecto combinado, sus intervalos de confianza, así como el valores del estadístico Z_i y su p-valor asociado al contraste de hipótesis se calculan de forma análoga a las ecuaciones (3.13), (3.14), (3.15), (3.16) y (3.17) del modelo MEF.

Tanto en el MEF como en el MEA, una vez obtenido los diferentes efectos combinados y sus correspondientes p-valores asociados, los p-valores se deben ajustar por algún método como el método de *Benjamini & Hochberg* para reducir el número de falsos positivos.

3.2.2. Meta-análisis basado en la combinación de p-valores

El propósito de los métodos basados en la combinación de p-valores es fusionar los p-valores obtenidos de manera individual en cada uno de los estudios para obtener un p-valor común. Muchos de estos métodos existían antes de la concepción del meta-análisis[19,88], ya que inicialmente se empleaban para combinar distintos p-valores. Sin embargo, debido a sus características, sus técnicas fueron extrapoladas a la integración de estudios realizada por el meta-análisis.

Estos métodos presentan una diferencia con respecto a los métodos basados en la combinación de tamaños de efectos. En el caso de los métodos basados en la combinación de tamaños de efecto el rechazo de la hipótesis nula ($H_0: \theta = 0$) implica que los verdaderos tamaños de efectos ($\theta_1 = \cdots \theta_i = \cdots = \theta_K$) son distintos de 0 en todos los estudios[24]. Por otro lado, en el caso de la combinación de p-valores, un p-valor combinado significativo (menor de 0,05) no necesariamente implica que todos los p-valores individuales sean significativos en cada estudio, ya que las hipótesis pueden variar de un método a otro. En la mayoría de los métodos, la hipótesis alternativa es que en al menos uno de los estudios el efecto o test tratado es significativo, es decir:

$$\begin{cases} H_0: \theta_1 = \cdots \theta_i = \cdots = \theta_K = 0 \\ \quad H_1: Al\ menos\ un\ \theta_j \neq 0 \end{cases}$$

Por lo tanto, en el caso específico de los datos de expresión génica, que un gen sea obtenido como diferencialmente expresado tras realizar estos métodos, no necesariamente implica que esté se pueda considerar diferencialmente expresado en todos los estudios.

Para estos datos los p-valores de los estudios individuales se obtienen a partir de los análisis de expresión diferencial en cada uno de los estudios haciendo uso de la metodología y software adecuados para cada tipo de datos, tal y como se ha descrito en la sección 3.1.3. Posteriormente, una vez que se han obtenido los diferentes p-valores de cada gen en cada uno de los conjuntos de datos, estos se combinan por alguna técnica de combinación de p-valores. Existen numerosos métodos para la combinación de p-valores, pero en esta sección vamos a desarrollar sólo los llevados a cabo durante la tesis doctoral.

3.2.2.1. Método de Fisher

El método de Fisher o test de probabilidad combinada de Fisher, llamado así en honor del estadístico *Ronald Fisher* fue el primer método de combinación de p-valores desarrollado.

Este test usa como estadístico la suma de los logaritmos de los p-valores[19,25]:

$$-2 \times \sum_{j=1}^{K} \ln\left(p_{ij}\right) \tag{3.26}$$

Donde p_{ij} es cada uno de los p-valores individuales del gen *i* en cada uno de los estudios.

En este caso, bajo la hipótesis nula se distribuye como una χ^2 con $2 \times K$ grados de libertad[25].Este método es sensible a p-valores pequeños de modo que un p-valor individual pequeño da lugar a un p-valor combinado pequeño.

3.2.2.2. Método de Pearson

Desarrollado por el estadístico Karl Pearson[89], el método de Pearson aplica un estadístico si similar al método de Fisher[25]:

$$-2\times\sum_{i=1}^{K}\ln\left(1-p_{ij}\right) \tag{3.27}$$

Bajo la hipótesis nula este estadístico también se distribuye también como una χ^2 con $2\times K$ grados de libertad[25]. A diferencia del método de Fisher, este método es más sensible a p-valores grandes.

3.2.2.3. Método de Stouffer

Formulado por el sociólogo Samuel A. Stouffer, el método de Stouffer asume que los p-valores se pueden transformar a valores z_{ij} a partir de la inversa de la función de distribución de una normal $\left(\Phi^{-1}\right)$:

$$z_{ij}=\Phi^{-1}\left(1-p_{ij}\right) \tag{3.28}$$

A partir de los valores z_i se construye en estadístico:

$$\frac{\sum_{i=1}^{K}z_{ij}}{\sqrt{K}} \tag{3.29}$$

El cual bajo la hipótesis nula se distribuye bajo una distribución normal tipificada $N\left(0,1\right)$.

Este método tiene como ventaja que permite la inclusión de pesos, ω_i, en los diferentes estudios, siendo el estadístico en ese caso de forma (método de Stouffer ponderado):

$$\frac{\sum_{j=1}^{K}\omega_{ij}z_{ij}}{\sqrt{\sum_{j=1}^{K}\omega_{ij}^2}} \tag{3.30}$$

Es recomendable que los pesos asignados sean la varianza de los estadísticos empleados para obtener los diferentes p-valores, aunque en el caso de que no se disponga de esta información, la raíz cuadrada de los diferentes tamaños muestrales de los estudios se ha demostrado que proporciona también buenos resultados. [90,91]. Además, esta asignación de pesos suele proporcionar resultados más fiables que el método de Fisher a reducir la influencia de los p-valores más pequeños.[91].

3.2.2.4. El método de Tippett

El método de Tippett o método del mínimo de los p-valores, fue formulado por estadístico Leonard Henry Caleb Tippett considera como estadístico el mínimo de los p-valores[92]:

$$\min(p_{i1},...p_{ij},...p_{iK}) \tag{3.31}$$

El mínimo de los p-valores se distribuye bajo la hipótesis nula como una distribución *Beta* de parámetros $\alpha = 1$ y $\beta = K$, $Beta(1,K)$ [25,93]. Este método sólo es recomendable cuando el número de estudios es muy pequeño, ya que solo en la combinación solo se considera un p-valor[94].

3.2.2.5. El método de Wilkinson

El método del máximo de los p-valores o método de Wilkinson, fue desarrollado por B. C.S. Wilkinson y adapta el método de Tippett para considerar el máximo de los p-valores[95]:

$$\max(p_{i1},...p_{ij},...p_{iK}) \tag{3.32}$$

En este caso bajo la hipótesis nula se distribuye bajo una $Beta(K,1)$ [94]. Este método tiene la ventaja que en este caso si se rechaza la hipótesis nula, se puede concluir que el efecto es significativo en todos los estudios, sin embargo, al considerar un solo p-valor, pueden obtenerse muchos falsos negativos, es decir el error tipo II del test puede ser elevado[94].

3.2.2.6. Método de la prueba de asociación de Cauchy agregada (ACAT)

El método de la prueba de asociación de Cauchy agregada o como es conocido en inglés, Aggregated Cauchy Association Test method (ACAT)[96] fue desarrollo para combinar p-valores de estudios de asociaciones genéticas que requieren de resultados robustos en estudios con alta heterogeneidad. Al igual que el método de Stouffer transforma los p-valores para obtener el estadístico:

$$\sum_{j=1}^{K} \tan\left(\left(0.5 - p_{ij}\right)\pi\right) \tag{3.33}$$

Bajo la hipótesis nula se distribuye bajo una distribución de Cauchy estándar[96,97]. Al aproximarse por una distribución de Cauchy permite dar más peso a los p-valores extremos ya que las colas de Cauchy tienen mayor peso de probabilidad que una distribución normal[96]. Además, al igual que el método de Stouffer permite la asignación de pesos, ω_i, a los estudios:

$$\sum_{j=1}^{K} \omega_{ij} \tan\left(\left(0.5 - p_{ij}\right)\pi\right) \tag{3.34}$$

Donde $\omega_{ij} > 0$ y $\sum_{j=1}^{K} \omega_{ij} = 1$. Bajo estas condiciones el estadístico se sigue aproximando bajo la hipótesis nula a una distribución de Cauchy estándar.

3.2.3. Métodos de combinación de rangos

Los métodos basados en la combinación de rangos se diferencian de otros métodos de meta-análisis porque solo se pueden aplicar cuando hay numerosas variables de la misma

naturaleza en estudio y se pueden distinguir dos grupos de muestras. Esto se debe a que estos métodos fueron desarrollados específicamente para los datos de expresión génica[26].

En este caso, consideramos los dos grupos en estudio, es decir, el grupo experimental *(E)* y el grupo de Control (*C).* Por lo que la matriz de expresión se podría dividir en dos matrices:

$$M_{G\times N}=\begin{pmatrix} e_{11} & \cdots & m_{1N} \\ \vdots & \ddots & \vdots \\ e_{G1} & \cdots & e_{GN_e} \end{pmatrix} \rightarrow E=\begin{pmatrix} E_{11} & \cdots & e_{1N_e} \\ \vdots & \ddots & \vdots \\ E_{G1} & \cdots & R_{Ge} \end{pmatrix}; C=\begin{pmatrix} c_{11} & \cdots & c_{1N_c} \\ \vdots & \ddots & \vdots \\ c_{G1} & \cdots & c_{GN_c} \end{pmatrix}$$

Donde N_e y N_c son los tamaños muestrales de los grupos experimental y control respectivamente.

Posteriormente, en cada estudio se calcula por pares una magnitud que permita ser ordenada o un tamaño de efecto[26]. Por ejemplo, se puede calcular el llamado *fold-change* (FC) o su logaritmo. El FC es la razón entre el valor de expresión de un grupo o muestra frente a otro grupo o muestra. De esta forma, obtendríamos un matriz por estudio de todas las comparaciones posibles. En el caso del FC se obtendría:

$$X_{G\times(N_e\times N_c)}=\begin{pmatrix} e_{11}/c_{11} & \cdots & e_{11}/c_{1N_c} & t_{12}/c_{11} & \cdots & e_{1N_e}/c_{1N_c} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ e_{G1}/c_{G1} & \cdots & e_{G1}/c_{GN_c} & t_{G2}/c_{G1} & \cdots & e_{GN_e}/c_{GN_c} \end{pmatrix}$$

Posteriormente, por se calcula el rango de cada gen por columnas, asignando 1 al efecto más pequeño y *R* al rango más grande, obteniéndose una matriz *R* de rangos[26]:

$$R_{R\times(N_e\times N_c)}=\begin{pmatrix} r_{11} & \cdots & r_{1(N_e\times N_c)} \\ \vdots & \ddots & \vdots \\ r_{G1} & \cdots & r_{G(N_e\times N_c)} \end{pmatrix}$$

Este proceso se repite en todos los estudios. Una vez obtenidas todas las matrices de rangos, se forma un vector para cada gen que incluye todos los rangos de ese gen en cada estudio. A partir de estos vectores se combinan los rangos mediante dos métodos: productos de rangos (*Rank Product* en inglés) o la suma de los rangos (*Rank sum* en inglés).

El **producto de rangos** para cada variable calcula como estadístico la media geométrica de los rangos[26]:

$$RP_i=\left(\prod_{n=1}^{N} r_{in}\right)^{1/N} \tag{3.35}$$

donde *N* es el tamaño del vector de rangos del gen *i.* Cuando este método fue desarrollado se obtenía el p-valor para la variable mediante un p-valor empírico[26]. Sin embargo, en la actualidad se conoce la distribución exacta del estadístico[98], aunque en la práctica se

suelen aplicar diferentes aproximaciones basadas en la distribución Gamma para obtener un p-valor aproximado[99,100].

Por otro lado, la **suma de rangos** implica calcular como estadístico la suma de los rangos para cada gen:

$$RS_i = \sum_{n=1}^{N} r_{in} \tag{3.36}$$

En estos métodos, tanto el cálculo de los p-valores empíricos como de las aproximaciones de los p-valores requieren de un gran coste computacional, por lo que solo se recomiendan cuando existen tamaños muestrales pequeños o pocas variables en estudio. Además, el coste computacional del producto de rangos es superior al de la **suma de rangos**.

Debido a este coste computacional elevado algunos investigadores sugieren aplicar estos métodos desde un enfoque distinto[101,102], sin realizar las comparaciones por pares en cada estudio. En su lugar, proponen calcular el FC u otra medida entre los grupos de cada uno de los estudios, obteniéndose una magnitud por gen en cada conjunto de datos. Posteriormente, obtener los rangos de cada gen en cada estudio a partir de las medidas calculadas. En último lugar, a partir de estos rangos se aplica el método del producto o la suma de rangos[101]. En la literatura estos métodos también se conocen como **producto de los rangos** (en inglés *product of the ranks*) o **suma de los rangos** (*sum of the ranks en inglés*)[101].

3.2.4. Cuantificación de la heterogeneidad

Para evaluar la fiabilidad y precisión del efecto combinado en un meta-análisis, es crucial conocer la heterogeneidad entre los estudios. Esto se debe a que una elevada heterogeneidad puede disminuir la consistencia y confiabilidad de los resultados. Además, la heterogeneidad es un factor determinante al elegir el modelo estadístico adecuado, ya que el MEF solo es fiable cuando los estudios son homogéneos. A lo largo de nuestro trabajo consideramos dos técnicas para medir la heterogeneidad en meta-análisis: el *test de la Q de Cochran*[103] y el estadístico I^2 de inconsistencia[104].

El ***test de la Q de Cochran*** está basado en la varianza total calculada en (3.20) y (3.21). Por lo tanto, el estadístico que aplica este test es la estimación Q de la varianza total de los tamaños de efectos:

$$Q_i = \sum_{j=1}^{K} \omega_{ij} \left(\theta_{ij} - \theta_i \right)^2 = \sum_{j=1}^{K} \omega_{ij} \theta_{ij}^2 - \frac{\left(\sum_{j=1}^{K} \omega_{ij} \theta_j \right)^2}{\sum_{j=1}^{K} \omega_{ij}} \tag{3.37}$$

Donde Q_i es la estimación de la varianza total del gen i. Este estadístico bajo la hipótesis nula de homogeneidad se distribuye bajo una χ^2 con $K-1$ grados de libertad[103]. Cabe señalar que esta medida para medir la heterogeneidad no es adecuada cuando existen tamaños muestrales pequeños[103]. Por ese motivo, implementamos además el estadístico I^2. En concreto, el estadístico I^2 mide la proporción de varianza observada que refleja

diferencias reales en los tamaños de efectos[24]. El I^2 para cada se obtiene a partir de la siguiente fórmula[104]:

$$I_i^2 = \left(\frac{Q_i - df}{Q_i} \right) \times 100\% \quad (3.38)$$

Donde I_i^2 es el valor del estadístico de inconsistencia I^2 para el gen *i. Higgins et al.*[104] establecen un valor del 25% para considerar que los estudios tienen heterogeneidad baja.

El *test de la Q de Cochran* y la I^2 son medidas complementarias para medir la heterogeneidad en los estudios de meta-análisis, ya que mientras la I^2 no depende de los tamaños muestrales, el *test de la Q de Cochran* nos proporciona un p-valor con el que tomar una decisión directa acerca de la homogeneidad[24].

3.2.5. Flujo de trabajo del meta-análisis de expresión diferencial

En el trabajo de *Toro-Domínguez y Villatoro-García et al*[56] (proporcionado en el Anexo 9.1), realizamos una descripción de los diferentes procedimientos que un investigador debe llevar a cabo para aplicar correctamente un meta-análisis de datos de expresión génica. Este flujo de trabajo puede resumirse en los siguientes pasos:

1. Hipótesis y selección de casos

En función del objetivo del estudio, se definen criterios de elegibilidad para buscar y seleccionar los conjuntos de datos a incluir, los cuales pueden ser tanto biológicos (enfermedad, tejido, edad, etc.) como técnicos (plataforma, tamaño de la muestra, controles de calidad, etc.). Los repositorios públicos como GEO permiten realizar búsquedas rápidas y automáticas utilizando palabras clave u ontologías. Es crucial considerar diferentes parámetros al seleccionar los datos, ya que pueden influir significativamente en los resultados. Por ejemplo, la heterogeneidad entre estudios se refiere a la diversidad técnica o biológica de los mismos, y las distintas cohortes de pacientes generadas en diferentes plataformas y laboratorios a menudo revelan conclusiones contradictorias para una misma pregunta. La selección de un nivel alto o bajo de heterogeneidad entre estudios puede servir a distintos propósitos; una alta heterogeneidad reduce la potencia estadística, pero incrementa la generalidad de los resultados[105,106]. Además, es importante trabajar con datos equilibrados en cuanto al número de muestras de cada clase y al tamaño muestral de cada estudio para homogeneizar el peso que cada estudio ejerce sobre los resultados. Como regla general, el efecto aislado de un estudio se minimiza con un mayor número de estudios, priorizando el significado general de los resultados. Por lo tanto, la selección adecuada de datos y la consideración de la heterogeneidad son fundamentales para obtener resultados robustos y generalizables en los análisis de expresión génica[107].

2. Procesamiento previo y normalización de los datos

A la hora de analizar estudios procedentes de diferentes plataformas es necesario aplicar un procesamiento previo de los datos. En el caso de que los datos sean crudos, en primer lugar, hay que obtener las matrices de expresión de cada uno de ellos, para lo cual es necesario aplicar flujos de análisis estándar en función de cada plataforma[108].

Posteriormente, es recomendable en muchos casos la normalización de los datos para minimizar las variaciones no biológicas[109]. Esto procesos de procesamiento previo y de normalización deben ser consistentes entre los diferentes estudios para minimizar lo máximo posible la heterogeneidad técnica. Existen numerosas revisiones y artículos que detallan cómo llevar a cabo este procesamiento y normalización según las plataformas de secuenciación utilizadas para obtener los datos[110,111].

Además, debido a la utilización de diferentes identificadores de genes, los diferentes conjuntos de datos pueden tener identificadores distintos, como Entrez Gene ID[112], Ensembl Gene ID[113], Official Gene Symbol[114] o identificadores propios de plataforma. En el caso de que los estudios provengan de diferentes plataformas es necesario anotar los diferentes conjuntos de sondas o genes con un mismo identificador común. En el caso de que varias sondas o marcadores pertenezcan al mismo identificador común es necesario agrupar sus valores de expresión para obtener la expresión del gen final[115]. Algunos enfoques se basan en resumir cada gen utilizando la sonda o marcador con la media más alta o más baja, la media absoluta o la varianza, o promediando los valores de todas sus sondas o marcadores.

Finalmente, es posible que los datos descargados ya hayan sido previamente procesados y normalizados. Sin embargo, si estos procesos no se realizaron de manera adecuada, existe el riesgo de introducir errores que podrían conducir a conclusiones incorrectas. Por ello, es crucial implementar rigurosos controles de calidad al trabajar con estos datos y considerar la posibilidad de problemas inherentes al procesamiento previo.

3. Controles de calidad

Es esencial realizar controles de calidad para detectar valores atípicos, mediciones inconsistentes o problemas técnicos en las muestras. Por un lado, los valores atípicos en los datos pueden introducir sesgos en los análisis posteriores, lo que lleva a estimaciones erróneas de los valores resultantes. Además, es necesario manejar los datos faltantes en las muestras, el cual es un problema común que puede producir resultados poco fiables cuando se hacen inferencias sobre una población basada en dichas muestras. Por lo tanto, los resultados dependen considerablemente de los métodos utilizados para procesar valores faltantes y atípicos[116]. En el caso de estudios de expresión génica, el problema de los valores faltantes en genes dentro de un conjunto de datos puede abordarse mediante imputación. Hay numerosas técnicas de imputación, entre los más frecuentemente utilizados está la sustitución del valor faltante por el promedio de expresión de ese gen o el uso de modelos derivados de la similitud entre genes, como imputación basada en la descomposición en valores singulares o en modelos de regresión. En el campo existen varias revisiones que abordan este problema y evalúan las diferentes técnicas de imputación[117–119]. En lo que respecta a la identificación de muestras atípicas generalmente se basa en la distribución anormal de los valores respecto a la normalidad o en medidas de similitud entre muestras, tales como correlaciones, distancias de Mahalanobis o análisis de componentes principales[120–122].

4. Corrección del efecto de lote o *batch effect*

El efecto de lote es la inter-heterogeneidad técnica producida por factores externos al estudio y por sesgos técnicos. Esto puede producir que la expresión génica producida por

la condición de estudio se vea eclipsada por estas fuentes de error. Por este motivo la corrección del efecto de lote se usa para eliminar estas fuentes de variación y controlar estos sesgos. Algunos de los enfoques más usados son el ajuste mediante el uso de covariables[45], la eliminación de la variación producida por variables ocultas[123] o el método de Bayes empírico[44].

5. Selección del método de meta-análisis

A hora de integrar los estudios de expresión génica mediante las técnicas de meta-análisis, el método empleado juega un papel crucial en los resultados finales del análisis, por lo que su elección es un paso fundamental.

En primer lugar, es importante considerar que las hipótesis alternativas de los diferentes contrates difieren de un método a otro.

En el caso de los métodos de basados en combinación de los tamaños de efectos, asumen como hipótesis alternativa que el tamaño de efecto (θ) es distinto de 0 en todos los estudios *(HSA)*, es decir, se puede considerar que le gen es significativo en todos los estudios si el resultado del meta-análisis es significativo[24]:

$$HSA = \begin{cases} H_0: \theta_i = 0 \\ H_1: \theta_i \neq 0 \end{cases}$$

Sin embargo, en el caso de los métodos de p-valor combinado (con la excepción del método de Wilkinson y el método de Pearson que si consideran el contraste *HSA)* y los métodos de combinación de rangos asumen como hipótesis alternativa que en al menos uno de los estudios el tamaño de efecto o la medida de referencia es distinta de 0 en al menos uno de los estudios *(HSB)*[25,93,99], es decir, se considera que el gen es significativo en al menos uno de los estudios si el resultado del meta-análisis es significativo:

$$HSB = \begin{cases} H_0: \theta_1 = \cdots \theta_i = \cdots = \theta_K = 0 \\ H_1: Al\ menos\ un\ \theta_j \neq 0 \end{cases}$$

Por otro lado, la heterogeneidad producida a la hora de combinar estudios de diferentes naturalezas con altos niveles de heterogeneidad puede influir en los resultados de los diferentes métodos. En este caso, los métodos basados en la combinación de tamaños de efecto se ven mucho más afectados negativamente que el resto (incluso aunque se use el MEA). Por lo tanto, a la hora de combinar estudios procedentes de diferentes plataformas de secuenciación o de diferentes condiciones biológicas puede ser más recomendable aplicar una combinación de p-valores o de rangos.

Dado el caso anterior de combinar estudios de diferentes naturalezas o plataformas, a la hora de seleccionar entre combinación de p-valores o de combinación de rangos, hay que destacar que los métodos de combinación de rangos requieren de mucha más capacidad de cómputo que los métodos de p-valores por lo que no son recomendables en el caso de disponer un gran número de estudios (en el caso del producto de rangos no se recomienda incluir más de 10 estudios), pero tienen la ventaja de que se ven menos afectados por valores extremos (en inglés *outliers*). Otra alternativa para aplicar métodos basados en la combinación de rangos cuando existe un gran número de estudios o muestras es aplicar los métodos de suma de los rangos *(sum of the ranks)* o el producto de los rangos *(product of the ranks)* pero sus resultados son menos robustos.

Por último, los métodos de p-valores destacan por su simplicidad y facilidad a la hora de combinar estudios muy heterogéneos. Entre ellos hay que destacar que los métodos de Tippet y Wilkinson son muchos más extremos al considerar como estadísticos el mínimo y el máximo de los p-valores respectivamente, por lo que omiten información del resto de estudios y puede llevar a la obtención de un gran número de falsos positivos o falsos negativos[25,93]. Por otro lado, los métodos de Fisher y Pearson, a pesar de considerar más información que los métodos de Tippet y Wilkinson, se ven muy afectados por los valores extremos, por lo que también pueden obtener una alta tasa de falsos positivos y negativos[25,93]. Por último, los métodos de Stouffer y ACAT resultan de gran utilidad cuando se disponen de diferentes pesos para cada uno de los estudios[90,97], aunque requieren de suposiciones sobre las distribuciones de los p-valores que no siempre pueden considerarse.

En la Figura 7 se muestra de forma resumida la adecuación de cada uno de los métodos en función de las características de los análisis:

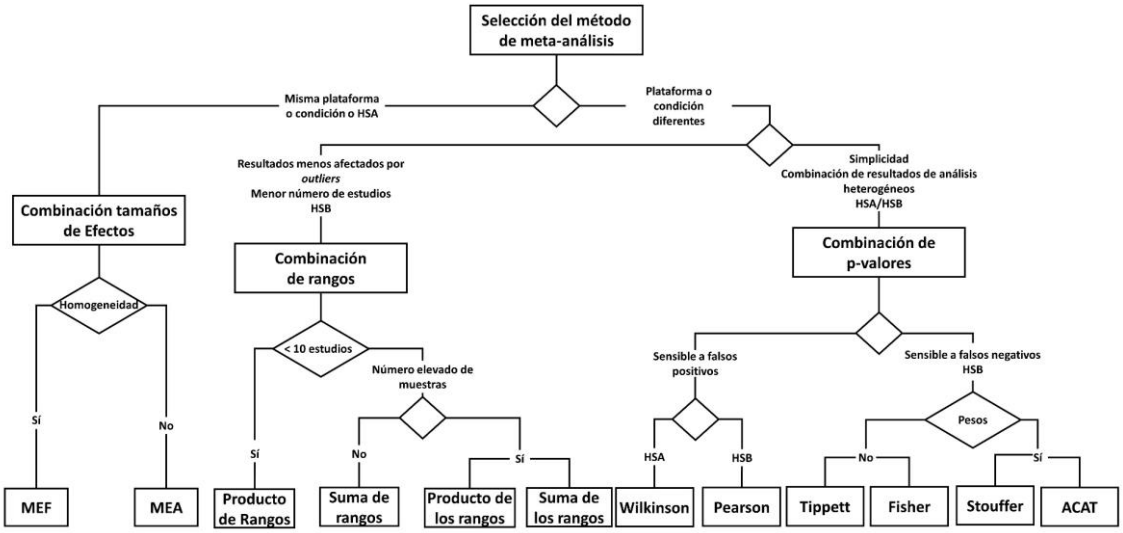


Figura 7. Esquema de decisión del método de meta-análisis. La figura resume las principales recomendaciones para la selección del método de meta-análisis más adecuado en función de las características de los datos. HSA: cuando el gen es significativo en el meta-análisis, lo es en todos los estudios. HSB: cuando el gen es significativo en el meta-análisis, lo es en al menos un estudio. Se trata de recomendaciones especialmente relacionadas con el tipo de datos. En la selección final del método, también es importante tener en cuenta el objetivo del meta-análisis. Figura adaptada del artículo *Toro-Domínguez et al 2021*[56].

6. Aplicación del meta-análisis e interpretación de los resultados

Tras realizar el meta-análisis se obtiene una lista de los genes diferencialmente expresados. Para interpretar estos resultados de manera adecuada, se utilizan herramientas de visualización como mapas de calor y gráficos de redes de interacción, que facilitan un análisis exploratorio y permiten observar cómo se expresan los genes en cada uno de los estudios incluidos. Además, con esta lista de genes es posible realizar análisis de enriquecimiento funcional para caracterizar las principales funciones biológicas que se encuentran alteradas entre las diferentes condiciones estudiadas.

3.2.6. Control de los genes faltantes

En nuestra tesis doctoral y a la hora de implementar software para aplicar las técnicas de meta-análisis a datos de expresión génica, controlamos la posible existencia de genes faltantes desde dos enfoques distintos: *Mínimo número de estudios que contiene al gen* y la imputación *sampleKNN*.

El primer enfoque, denominado enfoque de la ***Mínima proporción de estudios que contiene al gen,*** considera en el meta-análisis de expresión génica aquellos genes que estén presentes en al menos una proporción determinada de estudios fijados por el usuario. Por ejemplo, dado 6 estudios, si fijamos la proporción de estudios en 0.5, se considerarán aquellos genes que estén presentes en al menos 3 estudios descartando aquellos genes que no cumplan el requisito. Los resultados finales para cada gen hacen referencia sólo al número de estudios en los que están contenidos. Además, los resultados del paquete mostrarán la proporción de estudios en los que el gen está contenido.

El segundo enfoque aplica el método ***sampleKNN*** descrito por *Mancuso et al.*[70]. Este método está basado en el método de imputación de los K-vecinos más cercanos (*K-Nearest Neighbors*, KNN), el cual imputa el valor faltante a partir de las K observaciones más similares. Estas observaciones más similares se obtienen a partir de las distancias entre las observaciones sin valores faltantes y el valor faltante que se desea estimar. Posteriormente, se imputa el valor utilizando alguna medida como la media o la mediana de las K observaciones. El método KNN ya se ha usado anteriormente junto con otros métodos de imputación para estimar los genes faltantes en el meta-análisis[69]. Tradicionalmente, este y otros métodos estimaban la expresión de los genes faltantes a partir de la expresión los genes comunes a todos los estudios. Sin embargo, *Mancuso et al.*[70] propusieron la estimación en el espacio de las muestras para imputar estos valores, es decir, imputar las muestras con genes faltantes a partir de muestras similares de otros estudios que si contienen a los genes faltantes. *Mancuso et al.*[70] demostraron en su artículo que estimar imputar los genes de un estudio a partir de otro en el espacio de las muestras ofrecía mejores resultados que imputando a partir de los genes no faltantes. En nuestro trabajo adaptamos el método de *sampleKNN (*Método KNN para imputar en el espacio de las muestras) para usarlo en meta-análisis. El procedimiento que llevamos a cabo es el siguiente:

En primer lugar, se aplica una tipificación de todos los estudios que se van a incluir en el meta-análisis, de tal modo que todos los valores de expresión tengan media 0 y desviación típica 1. Posteriormente, se concatenan todas las matrices de expresión en una sola matriz y se aplica el método *sampleKNN* para imputar las diferentes muestras. Finalmente se deshace el escalado y se separan de nuevo las matrices de cada uno de los estudios.

3.3. Resultados

3.3.1. Revisión y análisis comparativo de software disponible para meta-análisis de expresión génica

Existe un gran número de herramientas y software para aplicar diversos métodos de meta-análisis en datos de expresión génica. A continuación, se proporciona una breve

descripción de algunas de las herramientas disponibles públicamente, enfocándonos en herramientas web y paquetes de R. Un análisis más en profundidad puede encontrarse en el artículo que avala esta tesis doctoral[56] y que proporcionamos en el Anexo 9.1.

3.3.3.1. Herramientas web

- *NetworkAnalyst*[124]*:* una de las aplicaciones web más populares. Su principal ventaja es que contiene implementados los principales pasos del meta-análisis de expresión génica como controles de calidad, los métodos de meta-análisis (basados en tamaños de efectos y en combinación de p-valores) y el análisis de enriquecimiento. No permite el uso de datos públicos, por lo que los usuarios tienen que cargar y preparar sus propios datos para poder usarla. La aplicación está disponible en: https://www.networkanalyst.ca/.
- *ImaGEO*[125]*:* es una aplicación web principalmente pensada para analizar directamente datos de expresión de la base de datos pública de GEO, aunque también permite al usuario introducir sus propios datos. Al igual que *NetworkAnalyst* tiene implementados una serie de controles de calidad, los principales métodos del meta-análisis basado en la combinación de tamaños de efectos y p-valores y análisis de enriquecimiento funcional. Está disponible en http://bioinfo.genyo.es/imageo/.
- *Gemma*[126]: es una aplicación web diseñada para analizar conjuntos de datos curados y precargados de GEO, aunque también permite al usuario introducir sus propios datos. Permite realizar meta-análisis de expresión génica basado en la combinación de p-valores entre otros análisis de este tipo de datos. Está disponible en https://gemma.msl.ubc.ca.
- *ExAtlas*[127]: al igual que *Gemma* es otra herramienta web diseñada para analizar datos precargados de GEO, así como datos introducidos por el usuario. Además de otros análisis permite aplicar meta-análisis basados en la combinación de tamaños de efectos y p-valores. Está disponible en https://lgsun.grc.nia.nih.gov/exatlas/.
- *ShinyMDE*[128]*:* es una herramienta web diseñada en R y Shiny que permite al usuario aplicar métodos de meta-análisis basado en la combinación de p-valores. El usuario tiene que introducir conjuntos de datos de expresión génica procedentes de microarrays de plataformas de Affymetrix e Illumina con los p-valores precalculados previamente. Se encuentra disponible en https://hussain.shinyapps.io/App-1/.

3.3.3.2. Paquetes de R

- *MetaOmics*[129]: es una aplicación web diseñada en Shiny que se carga en el entorno de R. Contiene diferentes módulos que permite aplicar los diferentes pasos del meta-análisis de expresión génica. Tiene implementados métodos basados en la combinación de tamaños de efectos, de p-valores y de rangos.
- *MetaIntegrator*[130] es un paquete de R que implementa métodos de meta-análisis y visualización de sus resultados. Permite aplicar el MEA y el método de Fisher de combinación de p-valores.
- *metap*[131]: es un paquete de R que implementa numerosos métodos de meta-análisis basado en la combinación de p-valores.

- *GeneMeta*[132] es otro paquete de R que implementa métodos del meta-análisis basados en la combinación de tamaños de efecto a partir de las matrices de expresión incluidas por el usuario.
- *metaMA*[49]*:* es un paquete de R que contiene funciones para realizar meta-análisis basados en la combinación de tamaños de efectos y p-valores a partir de matrices de expresión de datos de microarrays.
- *metaRNASeq*[133]: es un paquete de R similar a *metaMA* pero que aplica meta-análisis de combinación de p-valores (métodos de Fisher o Stouffer) para integrar estudios de expresión de RNA-Seq.
- *RankProd* [26,134]*:* es un paquete de R especializado en análisis de expresión diferencial y meta-análisis de expresión génica a partir de métodos basados en la combinación de rangos (producto de rangos y suma de rangos).
- *RankAggreg*[135]: es un paquete de R que permite aplicar métodos de combinación de suma de rangos a partir de lista de genes ordenadas.
- *OrderedList*[136]*:* Este paquete R combina listas de genes ordenadas a partir de algoritmos de combinación de rangos.
- *metahdep*[137]: es un paquete de R que permite aplicar meta-análisis basados en la combinación de tamaños de efecto a partir de datos crudos de microarrays.
- *metaSeq*[138]: este paquete de R permite aplicar los métodos de combinación de p-valores de Fisher y Stouffer para analizar datos de RNA-Seq
- MetaVolcanoR[139]: es un paquete de R que permite aplicar el MEA de combinación de tamaños de efectos y el método de Fisher y posteriormente mostrar diferentes visualizaciones como el diagrama de volcán o *volcano plot* en inglés.
- *crossmeta*[140]: es un paquete de R que permite la descarga de datos de microarrays de GEO y posteriormente aplica un meta-análisis basado en la combinación de tamaños de efecto.

3.3.2. DExMA: Un paquete de R para aplicar meta-análisis de expresión génica con genes faltantes

Como se ha introducido en la sección 3.1.5, uno de los problemas que nos podemos encontrar a la hora de combinar estudios de expresión génica es la posible existencia de genes no medidos en algunos de los estudios, también llamados genes faltantes o *missing genes*[69]. Este problema aparece principalmente a la hora de combinar estudios procedentes de bases de datos públicas, los cuales se han secuenciado mediante diferentes plataformas. Para afrontar este problema tanto en el contexto del meta-análisis de expresión génica como en el contexto de otros análisis de este tipo de datos, algunos autores han propuesto diferentes métodos que permitan imputar estos valores faltantes a partir de la expresión de otros genes[69] o a partir de la expresión de los genes faltantes en otras muestras que si lo contienen[70]. Sin embargo, la mayoría de los paquetes de R desarrollados para realizar meta-análisis de expresión génica solo trabajan con los genes comunes y ninguno de ellos implementa funciones que permita imputar la expresión de estos genes faltantes. Por este motivo, desarrollamos DExMA (*Differential Expression Meta Analysis*), un paquete de R que permite aplicar todos los pasos del meta-análisis de datos de expresión génica considerando la posible existencia de genes faltantes. DExMA

permite además la descarga de datos de GEO mediante los identificadores de los estudios, así como controles de calidad y visualización de los resultados. DExMA se encuentra disponible en el repositorio de Bioconductor (http://bioconductor.org/packages/release/bioc/html/DExMA.html). En esta sección haremos una descripción de las principales funcionalidades del paquete y de su aplicación a datos reales, aunque una descripción más detalla puede encontrarse en el artículo de *Villatoro-García et al.*[73] (proporcionado en el Anexo 9.2.) y en la guía del paquete que está disponible en Bioconductor[141].

3.3.2.1. Funciones y flujo de trabajo en DExMA

DExMA contiene implementadas diferentes funciones para realizar los pasos propios meta-análisis de expresión génica. En este flujo de trabajo podemos diferenciar 6 pasos principales: creación de objetos de meta-análisis, anotación de genes, control de calidad, imputación de genes faltantes, aplicación de métodos meta-análisis y visualización de resultados. En la Figura 8 podemos ver un esquema resumido del flujo de trabajo de DExMA.

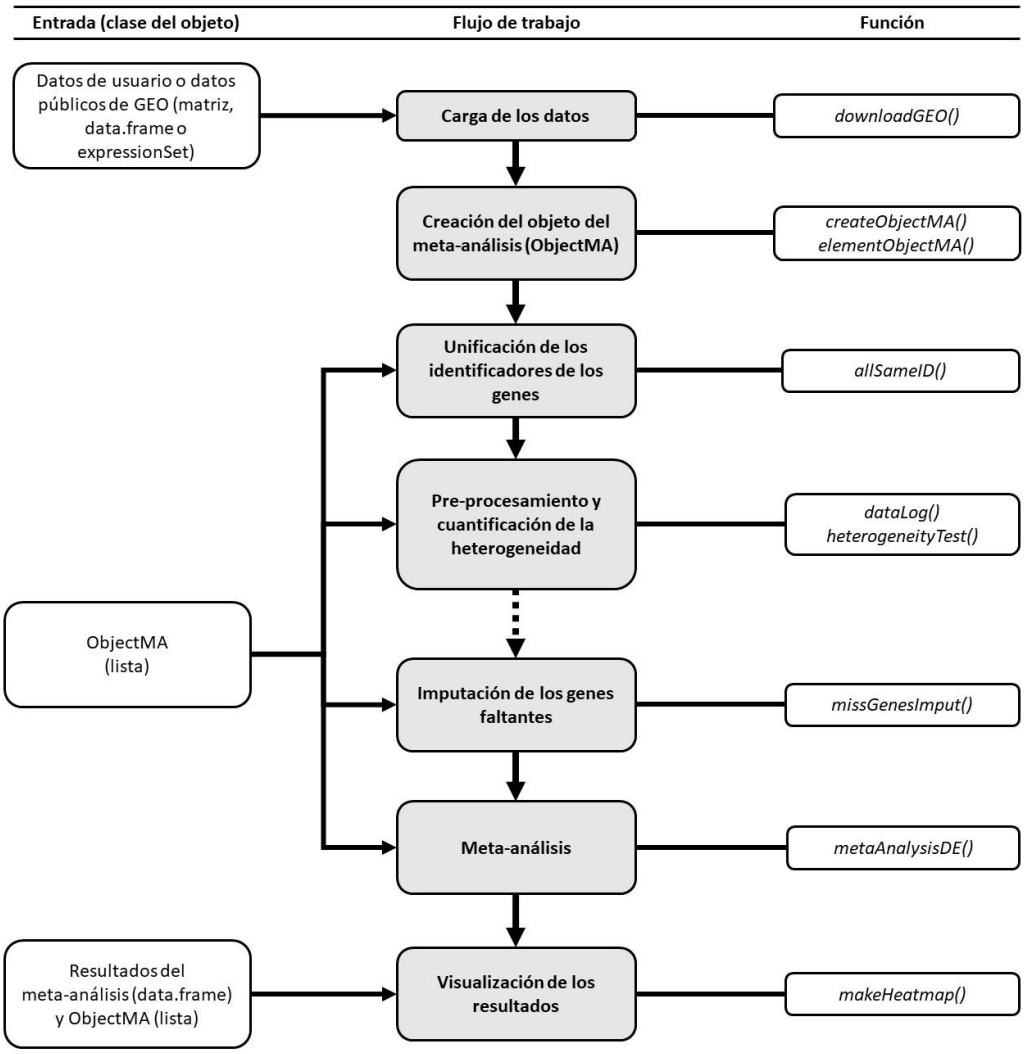


Figura 8. **Flujo de trabajo de DExMA**. La figura muestra los principales pasos del flujo de trabajo del paquete DExMA: (1) carga de datos y creación de objetos de meta análisis, (2) anotación de genes, (3) control de calidad, (4) imputación de genes ausentes (opcional), (5) meta análisis de expresión génica y (6) visualización. Figura adaptada del artículo *Villatoro-García et al* 2022[73].

Además, contiene unos datos de prueba simulados que ayudan al usuario a tener un primer contacto con el paquete y sus funciones. Este conjunto de datos llamado "*DExMAExampleData"* contiene diferentes objetos de R:

- Una lista formada por cuatro matrices de expresión *("listMatrixEX")*
- Una lista formada por cuatro data.frames o phenodatas con información de las muestras de las matrices de expresión *("listPhenodatas")*.
- Una lista formada por cuatro *ExpressionSet ("listExpressionSets")*. El objeto ExpressionSet es un objeto típico de los análisis bioinformáticos en R. En este caso contiene la misma información que el objeto *"listMatrixEX" y "listPhenodatas"*.
- Otro objeto llamado *"ExpressionSetStudy5"* similar al resto de *ExpressionSets*
- Un objeto llamado *"maObjectDif"* que es un ejemplo del tipo de con el que trabaja DExMA (llamada *objectMA* en las funciones) y que se ha formado a partir de los objetos *"listMatrixEx"* y *"listPhenodatas"*.
- Un objeto denominado *"maObject"* que es igual que el objeto *"maObjectDif"* pero obtenido a partir de establecer todos los estudios en la anotación Official Gene Symbol.

El primer paso en DExMA consiste en la creación de su propio objeto denominado *ObjectMA* y a partir del cual se pueden aplicar las diferentes funciones. Este objeto consiste en una lista de lista anidadas donde cada una contiene dos elementos: una matriz de expresión génica (con los genes en filas y las muestras en columnas) y un vector de ceros y unos que indica el grupo al que pertenece cada muestra (0 representa el grupo control y 1 el grupo experimental). Este objeto puede crearse con la función del paquete llamada *createObjectMA()*. Una vez creado el objeto, el resto de funciones se pueden aplicar de manera directa.

Como ya indicamos en *Toro-Domínguez et al.*[56], en uno de los pasos previos a aplicar el meta-análisis, es que todos los estudios tengan los genes anotados con los mismos identificadores para hacerlos comparables. En este caso DExMA proporciona la función *allSameID()* que permite traducir los identificadores de los diferentes estudios a un identificar común. Este caso los identificadores soportados por el paquete son los identificadores más comunes, es decir, Entrez, Ensembl y Official Gene Symbol.

A continuación, DExMA implementa funciones de pre-procesamiento que ayudan al usuario a transformar los datos (logaritmos) y cuantificar la heterogeneidad. En concreto, mediante la función *datalog()* se puede comprobar si los datos están en logaritmo (similar al procedimiento aplicado por GEO2R[142]) y, en el que caso de que no estén, aplica un logaritmo en base 2 (*log2*). En lo que se refiere al estudio de la heterogeneidad, DExMA proporciona la función *heterogeneityTest()* la cual presenta dos formas de cuantificarla: mediante un gráfico cuantil-cuantil o gráfico *Q-Q* (Figura 9) y el estadístico I^2 de inconsistencia. En lo que respecta al gráfico Q-Q este representa los diferentes estadísticos del test de Cochran de heterogeneidad. Este gráfico supone como distribución esperada (línea central) una distribución χ^2 de *K-1* grados de libertad (siendo K el número de estudios). Por lo tanto, en el caso de que la mayoría de los valores (cada valor está asociado a un gen) estén cerca de la línea central, podemos asumir que existe homogeneidad, y en el caso contrario heterogeneidad. Por lo otro lado, en cuanto al

estadístico I^2 para evitar que representar un estadístico por gen, la función devuelve los cuantiles del conjunto de valores I^2. En este caso, se podrá asumir homogeneidad si en un percentil elevado el valor de I^2 es menor de 0.25.

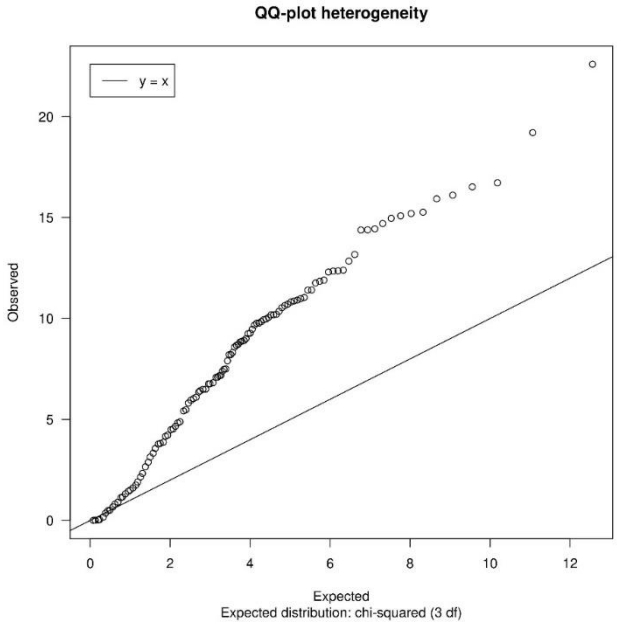


Figura 9. Ejemplo de gráfico Q-Q devuelvo por DExMA. Gráfico cuantil-cuantil donde la línea central representa la distribución esperada en caso de homogenidad.

Seguidamente, DExMA de manera opcional permite al usuario aplicar una imputación de los genes faltantes. Esto se realiza con la función *missGenesImput()*, la cual aplica el método *sampleKNN* (descrito en la sección de 3.2.6.) y devuelve el mismo objeto del paquete con los genes faltantes imputados. Además, proporciona una serie de indicadores para observar la calidad de la imputación: el número y porcentaje de genes faltantes imputados en cada muestra y el número y el porcentaje de valores faltantes imputados por gen y su porcentaje.

Posteriormente, DExMA tiene implementada la función *metaAnalysisDE()* la cual permite aplicar los diferentes métodos del meta-análisis basado la combinación de tamaños de efectos y en la combinación de p-valores descritos en la sección 3.2. de Metodología. En lo que respecta a los modelos basados en la combinación de tamaño de efecto, DExMA calcula la *g de Hedg*es para cada uno de los genes en cada uno de los estudios, mientras para el cálculo de los p-valores de los estudios individuales aplica el paquete de R limma[45].Adicionalmente, en el caso de que no se haya aplicado la imputación de los genes faltantes, en esta función podemos aplicar el segundo enfoque de tratamiento de genes faltantes indicando la proporción de estudios que deben contener a un gen para que este sea considerado en el meta-análisis.

Finalmente, DExMA proporciona al usuario la opción de representar los genes obtenidos como significativos en un mapa de calor o *heatmap* (ver Figura 10) mediante la función *makeHeatmap()*, lo que permite observar cómo están expresados cada uno de estos genes en cada uno de las muestras. Los mapas de calor están escalados para hacer comparables cada uno de los estudios.

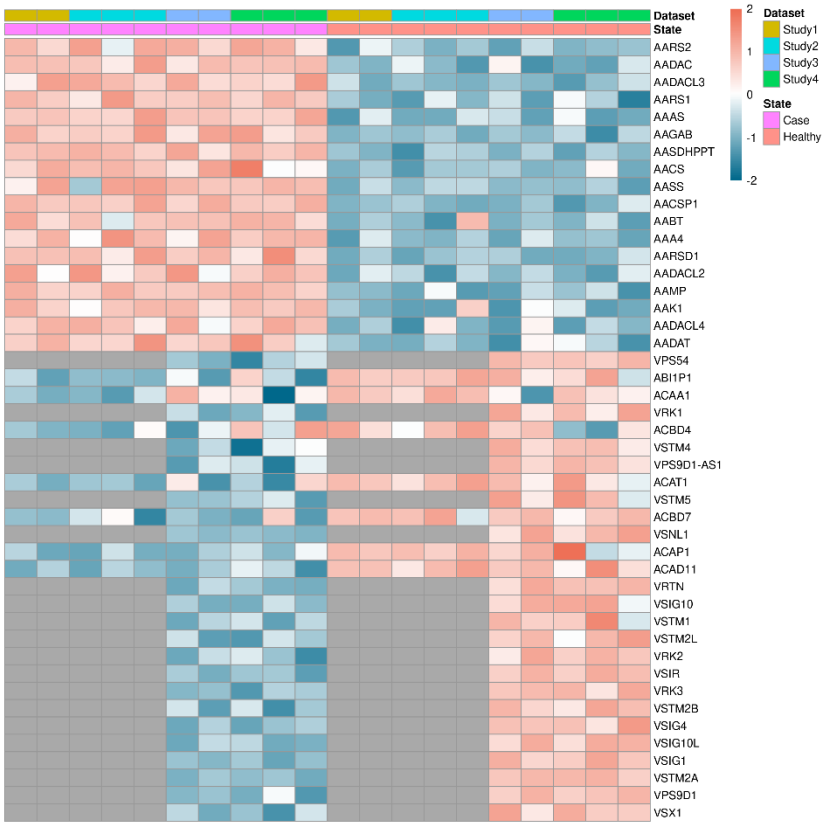


Figura 10. Ejemplo de mapa de calor devuelto por DExMA. Mapa de calor obtenido de aplicar el paquete DExMa los datos de ejemplos inlcuidos en él.

A parte de estas funciones hay que destacar que DExMA contiene implementadas otras funciones adicionales, como la función *downloadGEOData()* que permite la descarga directa de estudios de GEO devolviendo una lista donde cada elemento es el *ExpresionSet* del estudio. La función *batchRemove()* que elimina el efecto lote que puede tener la expresión corrigiendo por diferentes variables. En último lugar, para entender mejor los resultados obtenidos, las funciones *calculateES()* y *pvalueIndAnalysis()* devuelve los tamaños de efecto y los p-valores de cada uno de los estudios individuales.

3.3.2.2. Casos de uso en DExMA con datos reales

Para mostrar las utilidades DExMA, aplicamos el paquete a datos reales. Se usaron datos de la base de datos de enfermedades autoinmunes de ADEx[143]. Esta base contiene datos de enfermedades autoinmunes que también se encuentran disponibles en la base de datos de GEO, pero que ya han sido procesados y normalizados. Concretamente, se descargaron tres conjuntos de datos pertenecientes a la enfermedad autoinmune de Lupus Eritematoso Sistémico, la cual se caracteriza porque el sistema inmune ataca a los propios tejidos y órganos del cuerpo causando inflamación y daño en ellos. Estos estudios, cuyos identificadores en la base de datos son *GSE24706*[144], *GSE50772*[145], and *GSE82221_GPL10558*[146], fueron los seleccionados porque todas las muestras fueron extraídas del mismo tejido, lo que permite una mayor homogeneidad entre ellos. A pesar de ello, en la cuantificación de la heterogeneidad se obtuvo un gráfico Q-Q en los que la mayoría de valores distan mucho de la distribución de referencia (Ver figura 11) y con un 25% de los genes con un I^2 mayor de 0.71. Por lo tanto, se aplicó un meta-análisis basado en la combinación de tamaños de efectos con el MEA como método.

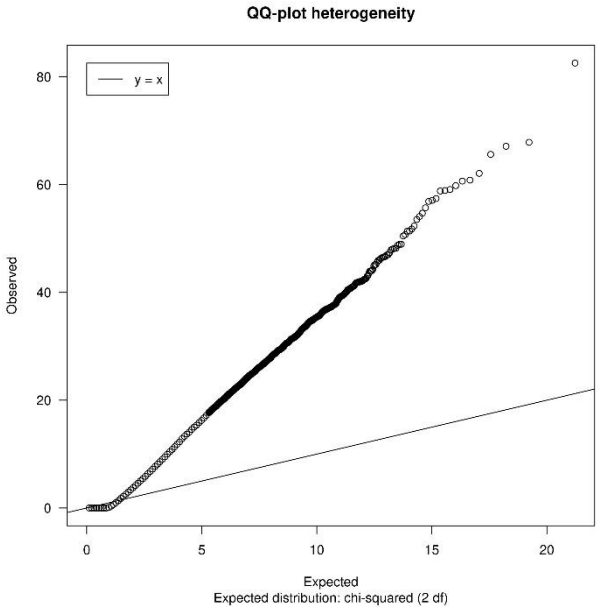


Figura 11. Gráfico Q-Q de la prueba de heterogeneidad para los datos de LES analizados. Gráfico QQ-plot obtenido tras aplicar la función de *heterogeneityTest()* a los datos de LES.

A la hora de abordar el problema de los genes faltantes y con el objetivo de demostrar los beneficios de tenerlos en consideración, se aplicaron los 3 enfoques distintos: considerando sólo los genes comunes (llamado *enfoque de genes comunes* a partir de ahora), considerando aquellos genes que estén en el al menos dos de los estudios (llamado *enfoque de proporción de genes* a partir de ahora) y aplicando la imputación de genes faltantes por el método de *sampleKNN* (llamado *enfoque de imputación* a partir de ahora). Esto permitió que mientras en el *enfoque de genes comunes* sólo se consideraran 11,298 genes, este número aumento a 14,548 en el *enfoque de proporción genes* y hasta 22,807 (todos los genes) en el *enfoque de imputación.* Esto implica que en el *enfoque de genes comunes* solo se hayan tenido en cuenta un 49.5% de los genes.

Considerando como significativos aquellos genes que tuvieran un p-valor ajustado menor de 0.05 por el método de Benjamini & Hochberg, tras aplicar el meta-análisis se obtuvieron como diferencialmente expresados y significativos 1896 genes en el *enfoque de genes comunes* (un 16.8% de los genes considerados), 2444 en el *enfoque de proporción de genes* (un 16.8% de los genes considerados) y 4830 en el enfoque de *imputación de genes* (un 21.1% de los genes considerados). Esto indica la pérdida de información que se produce si sólo se tienen en cuenta los genes comunes en el análisis. Esto también se puede observar si se genera un mapa de calor de los 50 genes más significativos (con p-valor más bajo) donde en algunos estudios presentan genes faltantes entre estos (Figura 12).

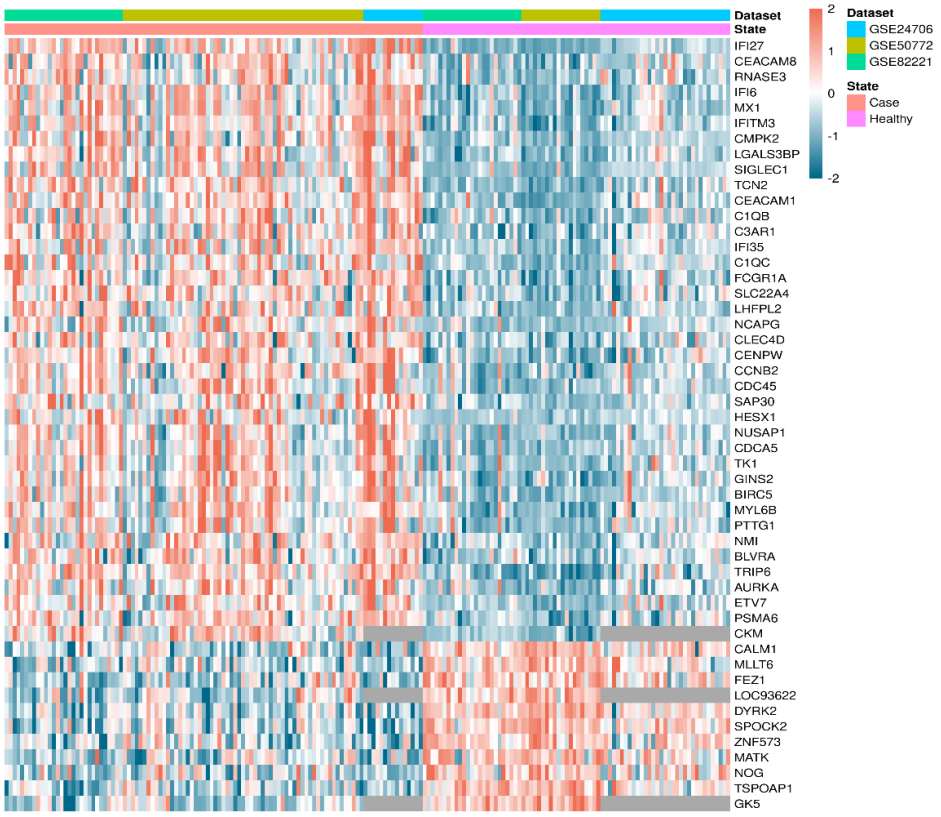


Figura 12. Mapa de calor de genes más significativos obtenidos por el *enfoque de proporción de genes.* Mapa de calor de los 50 genes con p-valor ajustado más bajo obtenidos por el *enfoque de proporción de genes.*

Finalmente, para comprobar si esta pérdida del número de genes considerados también implicaba una pérdida de información a nivel biológico, se llevó a cabo un análisis de enriquecimiento funcional, introduciendo los genes obtenidos como significativamente sobreexpresados en la aplicación GeneCodis4[147,148]. Los resultados de las rutas biológicas más significativamente desregulados en la base de datos de Bioplanet 2019[149] puede observase en la Figura 13. En este caso, el *enfoque de imputación de genes* obtiene la ruta propia del Lupus Sistémico Eritematoso (*Systemic Lupus Erythematosus*) como la ruta más significativa, algo que no ocurre en los otros dos enfoques. Esto demuestra la pérdida de información que pueden producir los genes faltantes y la inconsistencia de los resultados si estos no son tenidos en cuenta.

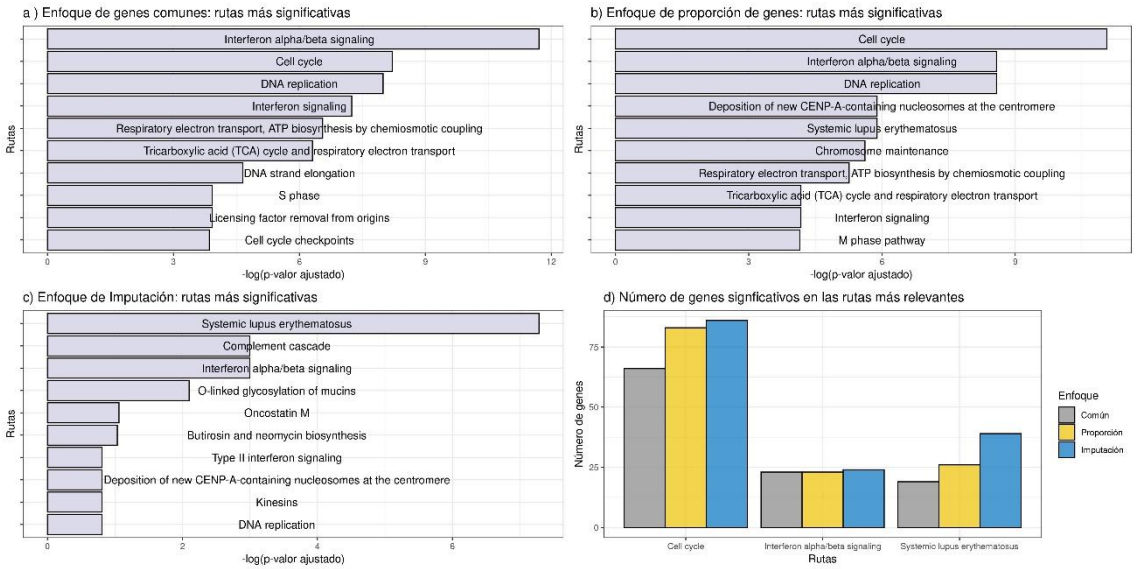


Figura 13. Representaciones gráficas de las rutas más significativas para cada uno de los enfoques de meta-análisis. (a) Diez rutas más significativas en el *enfoque de genes comunes*. (b) Diez rutas más significativas en el *enfoque de proporción de genes*. (c) Diez rutas más significativas en el *enfoque de imputación*. (d) Número de genes significativos en las rutas principales en cada enfoque. Figura adaptada de Villatoro-García et al. 2022[73].

Una explicación más detallada de los resultados obtenidos, así como el código empleado puede observarse en *Villatoro-García et al.*[73] y que se encuentra disponible en el Anexo 9.2.

3.4. Conclusiones

La acumulación y disponibilidad de datos de expresión génica en repositorios públicos ha impulsado el desarrollo de técnicas de meta-análisis como herramientas esenciales para integrar diferentes estudios de este tipo de datos. Estas técnicas se han utilizado con diferentes aplicaciones como el descubrimiento de biomarcadores de enfermedades o la búsqueda de patrones similares u opuestos entre diferentes condiciones. En este escenario, es crucial para la comunidad científica contar con paquetes de software que implementen métodos estadísticos adecuados y flujos de trabajo específicos, que permitan una correcta aplicación de estos métodos y la obtención de resultados fiables.

Además, uno de los problemas detectados a la hora de combinar los datos de expresión es la posible existencia de genes faltantes generados por diferentes plataformas de secuenciación. La práctica habitual a la hora de combinar conjuntos de datos que presentan genes faltantes es la considerar sólo los genes comunes. Sin embargo, este procedimiento puede producir que se produzca la pérdida de información relevante.

En todo este contexto, nuestros trabajos[56,73] son de especial relevancia. En el primer artículo proporcionamos una guía de los métodos y pasos necesarios para realizar un meta-análisis de expresión génica cubriendo todos los pasos necesarios. Estos procedimientos son implementados en la herramienta DExMA, la cual, a su vez, permite afrontar el problema de los genes faltantes desde dos enfoques distintos. Su aplicación a datos reales pone de manifiesto la obtención de resultados precisos y fiables cuando se integran estudios que presentan estos genes no medidos.

4. META-ANÁLISIS BASADO EN INFORMACIÓN DE RUTAS Y PROCESOS BIOLÓGICOS

4.1. Conceptos previos

4.1.1. El análisis de enriquecimiento funcional

El análisis de expresión diferencial, al igual que otros análisis de datos ómicos, tiene como resultado final grandes listas de genes significativos en el estudio. Estas listas por sí solas no proporcionan información para que los investigadores obtengan conclusiones claras e interpretaciones detalladas de los procesos biológicos que están actuando en relación con el contexto experimental o la enfermedad analizada. Por ello, es crucial realizar análisis posteriores que ayuden a contextualizar los resultados dentro de los mecanismos biológicos subyacentes, facilitando así una comprensión más integral y precisa de cómo las alteraciones en estas variables pueden influir en unas determinadas condiciones.

Este es el objetivo de las técnicas conocidas como análisis de enriquecimiento funcional (*functional enrichment analysis)*, las cuales comprenden un conjunto de métodos que permiten identificar qué procesos biológicos o funciones celulares están asociados con la lista de genes significativos. En la literatura estas técnicas también son conocidas como análisis de conjunto de genes (*Gene Set Analysis* o GSA).

En este proceso es esencial el uso de bases de datos que contengan las diferentes anotaciones funcionales, es decir, que proporcionen la información necesaria sobre la información funcional de los genes y su participación en los diferentes procesos biológicos. Algunos ejemplos de estas bases de datos son Gene Ontology (GO)[150], Kyoto Encyclopedia of Genes and Genomes (KEGG)[151] y Reactome[152],

En este tipo de análisis se han usado diferentes enfoques estadísticos, los cuáles evalúan si alguno de los procesos biológico de la base de datos empleada está sobrerrepresentado en la lista de variables en estudio mediante un valor o puntuación de enriquecimiento (*enrichment score* en inglés). Entre los métodos más comunes se encuentran el análisis de sobre-representación (Over-Representation Analysis, ORA) y el análisis de enriquecimiento de conjuntos de genes (Gene Set Enrichment Analysis, GSEA). En el caso de ORA a partir de los resultados del análisis de expresión diferencial, se selecciona una lista con los genes significativos. Posteriormente, a partir de esta lista se aplica un test estadístico, como el test exacto de Fisher o test basados en la distribución hipergeométrica, los cuales comparan la proporción de genes en una ruta o categoría específica con lo que esperaría obtenerse por azar. No obstante, este enfoque presenta ciertas limitaciones, dado que el umbral para determinar la significancia de un gen suele ser arbitrario, lo cual ignora el efecto coordinado de los elementos involucrados en una misma ruta biológica. Esto puede producir que se pierdan importantes biomarcadores o que el número sea demasiado elevado para obtener procesos realmente relevantes. El método GSEA se desarrolló para solventar estas limitaciones. En GSEA se ordenan todos los genes en base a sus valores de expresión diferencial (como el FC, por ejemplo) y a partir de este orden aplica un Test de Kolmogorov-Smirnov para evaluar si algún conjunto de genes muestra un perfil biológico distinto entre ambas condiciones.

Todos estos diferentes procedimientos permiten a los investigadores no solo identificar posibles biomarcadores, sino además conocer cuáles son los mecanismos biológicos que son relevantes en las diferentes condiciones en estudio.

4.1.2. Meta-análisis de enriquecimiento de rutas

En el caso del meta-análisis de expresión génica, el proceso habitual es aplicar un análisis de enriquecimiento funcional a las listas de genes obtenidas en los resultados del meta-análisis de expresión génica. Sin embargo, este procedimiento tiene algunas limitaciones ya que la variación de la expresión de los genes entre los diferentes estudios puede producir que la información de las rutas biológicas no se conserve de manera adecuada, obteniéndose resultados inconsistentes[153]. En este contexto, para integrar la información de las rutas biológicas de una manera más consistente, se han propuesto diferentes técnicas que tratan de combinar los métodos del meta-análisis con las técnicas de enriquecimiento funcional, conocidas como meta-análisis de enriquecimiento de rutas o en inglés *pathway enrichment meta-analysis* (MAPE). En la literatura, estas técnicas también se conocen como meta-análisis de enriquecimiento funcional o meta-análisis de enriquecimiento de conjunto de genes.

Originalmente, estas técnicas fueron descritas por *Kui Shen y George C. Tseng*[154], quienes inicialmente propusieron tres enfoques diferentes. El primer enfoque es aplicar el meta-análisis a nivel de genes (MAPE_G). Este método es similar al procedimiento estándar del análisis funcional en el meta-análisis de datos de expresión génica. En primer lugar, se realiza un meta-análisis de expresión génica basado en la combinación de p-valores (recomendando el método de *Wilkinson* para esta combinación) y, posteriormente, se aplica un análisis de enriquecimiento de rutas (GSEA) a los resultados obtenidos.

El segundo enfoque es aplicar el meta-análisis a nivel de rutas o *pathways* (MAPE_P). En este caso, se realiza un análisis de expresión diferencial en cada estudio y, tras un posterior GSEA, se combinan los resultados de los diferentes análisis de enriquecimiento funcional utilizando un método de combinación de p-valores (nuevamente recomendando el método de *Wilkinson*).

Finalmente, proponen un tercer enfoque que integra los resultados de ambos procedimientos (MAPE_I). Este método consiste en combinar los resultados de MAPE_G y MAPE_P aplicando el método de *Tippett* para combinar los p-valores de cada una de las rutas obtenidos por cada uno de los procedimientos.

Con posterioridad se desarrollaron otros procedimientos para aplicar un meta-análisis de enriquecimiento de rutas. *Chen et al.* desarrollaron una metodología que mediante modelización bayesiana combinaba toda la información de conjuntos de genes y datos de expresión génica[153]. *Lu et al.* implementaron iGSEA (Integrative Gene Set Enrichment Analysis), en el que desarrollan un método de meta-análisis denominado meta-análisis adaptativo, el cual integra los modelos MEF y MER para realizar un meta-análisis de expresión génica y posteriormente utiliza los tamaños de efectos obtenidos para obtener su propio valor de enriquecimiento para una ruta biológica determinada[155]. Por otro lado, *Meng H et al*. propusieron aplicar la metodología QuSAGE (Quantitative set analysis for gene expression)[156] para realizar meta-análisis de enriquecimiento de conjuntos de genes,

es decir, sugirieron combinar las funciones de densidad de probabilidad de conjuntos de genes obtenidos a partir de la expresión génica[157].

Sin embargo, estas técnicas también presentan algunas limitaciones. En primer lugar, exceptuando iGSEA, ninguna de las técnicas aplica los métodos del meta-análisis basados en la combinación de tamaños de efecto, lo que dificulta al investigador medir como está regulada exactamente cada una de las rutas biológicas en cada uno de los estudios o atribuir un peso a la hora de combinarlos. En su lugar, utilizan técnicas de combinación de p-valores (enfoques MAPE_G, MAPE_P o MAPE_I) o técnicas no propias del meta-análisis (QuSAGE y el modelo bayesiano de *Chen et al.*). Esto se debe a la dificultad de considerar medidas de puntuación de enriquecimiento o similares como tamaño de efecto, debido a sus propiedades intrínsecas. Por ejemplo, en el caso de GSEA, estimar la varianza del valor de enriquecimiento cuando éste se ha obtenido a partir de un test de Kolmogorov-Smirnov resulta extremadamente complejo. Del mismo modo, en el caso específico de iGSEA, aunque aplique modelos basado en la combinación de tamaños de efectos su procedimiento es similar al habitual o a MAPE_G, ya que el meta-análisis se realiza antes del enriquecimiento funcional.

Además, ninguna de las técnicas considera la posible presencia de genes faltantes a la hora de combinar los estudios y la influencia en los análisis de enriquecimiento funcional. Como hemos tratado en las secciones anteriores, esta pérdida de información puede producir que rutas biológicas de gran importancia en una condición o enfermedad no se obtengan como suficientemente enriquecidas o reguladas debido a la ausencia de algunos de los genes que las componen a lo largo de los estudios.

Con el objetivo de abordar los diversos problemas asociados con las diferentes técnicas de meta-análisis de enriquecimiento de rutas, durante esta tesis doctoral se ha desarrollado una nueva metodología para el análisis de enriquecimiento de rutas, denominada, Meta-análisis de conjuntos de genes o GSEMA (*Gene-set Enrichment Meta-analysis*). Esta nueva metodología combina distintas técnicas de meta-análisis de expresión con métodos de enriquecimiento de una sola muestra. Esta combinación permite trabajar primero en el espacio de las rutas en lugar del espacio de los genes, lo que reduce la influencia de los genes faltantes y además permite calcular un tamaño de efecto por estudio, similar al meta-análisis de expresión génica. Adicionalmente, GSEMA presenta algunas consideraciones metodológicas propias que difieren del meta-análisis de expresión génica tradicional. Estas diferencias se presentan en el cálculo del tamaño de efecto de cada una de las rutas, en la estimación de su varianza y en un filtrado previo de las rutas incluidas.

4.2. Metodología

4.2.1. Técnicas de enriquecimiento de una sola muestra

En el contexto del análisis de enriquecimiento funcional, en los últimos años se han desarrollado otras técnicas que, en lugar de obtener un valor de enriquecimiento para cada ruta biológica de manera general a partir de los genes diferencialmente expresados, permiten obtenerlo de manera individual para cada una de las muestras, es decir, permiten pasar de la matriz de expresión génica $M_{G\times N}$ a una matriz $R_{P\times N}$ en la que la que P es el

número de rutas de biológicas y cada uno de los elementos de la matriz, r_{ij}, son los valores de enriquecimiento de cada ruta biológica en cada muestra.

$$M_{G\times N} = \begin{pmatrix} m_{11} & \cdots & m_{1N} \\ \vdots & \ddots & \vdots \\ m_{G1} & \cdots & m_{GN} \end{pmatrix} \rightarrow R_{P\times N} = \begin{pmatrix} r_{11} & \cdots & r_{1N} \\ \vdots & \ddots & \vdots \\ r_{P1} & \cdots & r_{PN} \end{pmatrix}$$

Estos métodos son llamados técnicas de enriquecimiento de una sola muestra o en ingles *single sample enrichment* (SSE) y durante esta tesis doctoral se han considerado cuatro técnicas diferentes: Análisis de Enriquecimiento de Conjuntos de Genes en una sola muestra o *single sample Gene Set Enrichment* analysis (ssGSEA)[158], Análisis de Variación de Conjunto de Genes o Gene Set Variation Analysis (GSVA)[159], Análisis de enriquecimiento de conjuntos de genes con valor Z o Z-score Gene Set Enrichment Analysis (para abreviar a partir de ahora lo llamaremos *Z-score*)[160] y singscore[161].

El método **ssGSEA** descrito por *Barbie et al.*[158] considera los valores absolutos de expresión $\left(\left|m_{ij}\right|\right)$ de cada una de las muestras. Estos valores se normalizan en orden decreciente y se almacenan en una lista llamada *L*. Para esa muestra y una ruta biológica determinada, *L* se divide en dos grupos, los que están fuera y dentro del conjunto de genes. A continuación, se calcula la función de distribución acumulativa empírica (FDAE) para cada grupo. La FDAE para el primer grupo se calcula utilizando la forma estándar, mientras que la FDAE para el segundo grupo se pondera por sus valores en *L*. El valor de enriquecimiento para esa muestra y esa ruta biológica es la suma de las diferencias entre las FDAE de los dos grupos.

Por otro lado, el método ***Z-score*** definido por *Lee et al.*[160] obtiene valores de una normal estándar $\left(z_{ij}\right)$ tipificando por filas los valores de expresión de cada uno de los genes. Para cada una de las muestras, los diferentes valores z_{ij} se combinan utilizando una el método de Stouffer, es decir, se suman los diferentes valores z_{ij} y, posteriormente, se divide la suma por la raíz cuadrada del número de genes que componen la ruta biológica. De este modo, cada muestra obtiene un valor z_{ij} como valor de enriquecimiento para cada ruta.

En el caso de la metodología **GSVA** implementada por *Hänzelmann et al.*[159] considera cada fila de la matriz de expresión génica como el perfil de expresión del gen y calcula su función de distribución acumulada mediante una estimación de kernel gaussiano para microarrays o de poisson para RNA-Seq $\left(F_i\right)$. Posteriormente reasigna cada valor m_{ij} con el de la función estimada $\left(F_i\left(m_{ij}\right) = z_{ij}\right)$ y normaliza los valores obtenidos ordenando por filas, centrándolos y aplicando un valor absoluto. Finalmente, a estos valores normalizados y para cada muestra se aplica un procedimiento similar al ssGSEA para obtener los valores de enriquecimiento.

Por último, *Foroutan et al.* desarrollaron el método **singscore**[161], que al igual que ssGSEA, ordena los valores de expresión de un paciente. Sin embargo, singscore ofrece la ventaja adicional de considerar la dirección esperada del efecto en lugar del valor absoluto. Para el conjunto de genes de una determinada ruta, se calcula la media de los

rangos pertenecientes a ese conjunto y se normaliza utilizando la mediana, el mínimo y el máximo teóricos del rango medio. Este rango medio se calcula de manera diferente si la dirección del efecto es conocida de antemano, ya que se considera relevante la desviación absoluta de la mediana. Este proceso se repite para cada muestra y cada ruta, permitiendo teóricamente una evaluación más precisa y dirigida del impacto de los genes en las rutas biológicas.

4.2.2. Filtrado de rutas biológicas poco expresadas

Las técnicas de enriquecimiento de una sola muestran proporcionan valores de enriquecimiento para cada muestra que suelen estar comprendidos entre -1 y 1 denotando -1 como el valor máximo teórica de infra-expresión de la ruta en la muestra y el valor de 1 como el valor máximo teórico de sobre-expresión. Por lo tanto, valores cercanos a 0 indicarían rutas poco expresadas en una muestra. Sin embargo, al ser valores tan bajos, pequeñas diferencias con pequeñas varianzas tanto en el grupo experimental como en el grupo control pueden dar lugar a rutas diferencialmente enriquecidas cuando en realidad sea un falso positivo si la ruta tiene un valor de enriquecimiento cercano a 0 en ambos grupos. Para controlar este sesgo, es recomendable aplicar un filtrado de las rutas que tengan una baja expresión en ambos grupos como paso previo al cálculo del tamaño de efecto. Un filtrado previo es un paso común en el análisis de expresión diferencial, como por ejemplo en el caso de datos de RNA-Seq en el que recuentos de valores de expresión muy bajos pueden dar lugar a diferencias significativas[111].

No obstante, en el contexto de diferentes técnicas de enriquecimiento, establecer un umbral preciso para determinar cuándo un valor de enriquecimiento es demasiado bajo puede ser complejo. En el caso de ssGSEA y singscore, los valores de enriquecimiento experimentan transformaciones que limitan su rango, evitando que se acerquen a los límites de -1 y 1. Además, en el caso específico de Z-score sus valores se distribuyen bajo una distribución $N(0,1)$ para el caso de las rutas poco reguladas, pero se alejan de esta distribución cuando si están diferencialmente reguladas.

Por estos motivos y con el objetivo de establecer un filtro uniforme para todas las técnicas, en GSEMA se aplica una normalización cuando se emplea un método distinto al Z-score. Esta normalización implica estandarizar cada estudio, restando los valores de la matriz de rutas por su media y dividiendo por su desviación estándar. De esta manera, las matrices de rutas de los diferentes métodos presentan un rango similar y el umbral aplicado funciona de manera consistente entre todas las técnicas.

En el caso específico de GSEMA, como procedimiento estándar se recomienda filtrar las rutas cuya mediana absoluta esté por debajo de 0.65 tanto en el grupo de casos como en el grupo de control. Este filtro elimina aproximadamente el 50% de los valores centrales cuando la distribución es normal estándar, excluyendo así aquellas rutas cuyos valores están cercanos a 0 y que no muestran diferencias notables entre casos y controles. Por lo tanto, aunque el umbral óptimo puede variar según el estudio y la naturaleza de los datos, este filtro proporciona una primera aproximación para excluir rutas poco reguladas en ambos grupos.

4.2.3. Cálculo del tamaño de efecto en GSEMA

Al igual que el caso del meta-análisis de expresión génica, en GSEMA implementamos el estimador de la *g de Hedges* para calcular el tamaño de efecto de cada ruta en cada uno de los estudios. Además, como mencionamos anteriormente, estos datos enfrentan un problema de falsos positivos similar al que se observa en los datos de expresión génica, por lo que, a parte del filtrado previo, implementamos el mismo procedimiento para calcular la *g de Hedges* y corregir su varianza.

Por lo tanto, en GSEMA se calcula también la *t de Student moderada* del paquete de R de *limma* junto con sus grados de libertad para cada una de las rutas y haciendo uso de ambos se obtiene la *g de Hedges* a partir de la demostración de *Rosenthal and Rosnow*[81] (ecuación (3.7)), es decir el estimador de la *g de Hedges* para la ruta *i* en el estudio *j* se obtendría como:

$$g_{ij} = J_j \times d_{ij} = \left(1 - \frac{3}{4\times\left(n_{j_E} + n_{j_C} - 2\right) - 1}\right) \times \left(\frac{\left(n_{j_E} + n_{j_C}\right)\times t_{ij}}{\sqrt{n_{j_E} + n_{j_C}} \times \sqrt{df_{ij}}}\right) \tag{4.1}$$

Siendo t_{ij} el estadístico de la *t de Student* moderada obtenida en el estudio *j* para la ruta *i*, n_{j_E} y n_{j_C} los tamaños muestrales de los grupos experimental y de control respectivamente y df_{ij} los correspondientes grados de libertad de t_{ij}.

Finalmente, se obtiene la varianza del estimador haciendo uso también del cálculo alternativo propuesto por *Lin et al*[77] (ecuación (3.6):

$$V_{g_{ij}} = \frac{1}{n_{j_E}} + \frac{1}{n_{j_C}} + \frac{\overline{g_i}^2}{2\left(n_{j_E} \times n_{j_C}\right)} \tag{4.2}$$

Donde $\overline{g_i}$ es la media de los estimadores de la *g de Hedges* para la ruta *i*.

4.2.4. Generación de datos simulados y procesamiento de datos reales

Para evaluar la metodología de GSEMA, ésta fue aplicada tanto a datos reales como a datos simulados.

1. Datos simulados

En lo que se refiere a los datos simulados, se generaron datos de bulk RNA-Seq haciendo uso del paquete de R en Bioconductor *MOsim*[162]. En concreto, se simularon 5 estudios, cada uno con 50 muestras de controles y 50 muestras de casos con 39,359 genes en total. Del total de genes solo se asumió que el 1% de ellos estarían diferencialmente expresados en cada uno de los estudios. Además, con el objetivo de mejorar la evaluación metodológica, se asignaron nombres ficticios a 23 de los genes más diferencialmente sobreexpresados de cada uno de los estudios y estos a su vez fueron asociados una ruta biológica ficticia que denominamos *"Simulated_Pathway"*. Con este enfoque nos aseguramos que haya una ruta biológica que tenga que estar significativamente sobreexpresada en los resultados. Asimismo, nos evitamos que esta ruta haya sido

obtenida significativa por el solapamiento de genes compartidos con otras rutas al estar formada por genes diferentes al resto de rutas.

2. Datos reales

En lo que respecta a los datos reales se analizaron estudios que contuvieran un diferente número de genes faltantes. De manera específica, se trabajó de nuevo con datos de la enfermedad por LES, seleccionándose seis conjuntos de datos de expresión procedentes del mismo tejido: células mononucleares de sangre periférica (PBMCs por sus singlas en inglés). Cinco de estos estudios fueron descargados nuevamente de la base de datos de ADEx. Estos estudios tienen como identificadores GSE11909_GPL96[163], GSE11909_GPL97[163], GSE24706[144], GSE50772[145], GSE82221_GPL10558[146] y fueron secuenciado por diferentes plataformas de microarrays. El otro estudio de bulk RNA-Seq con identificador de GEO GSE122459[164] fue descargado de la base de datos de recount3[165]. La matriz de expresión del estudio GSE122459 fue procesada con el paquete de R de edgeR[47,50] para filtrar los genes con baja expresión y posteriormente se normalizó con el paquete NOISeq[52] mediante el método *tmm* o *trimmed Mean of M*.

Algunos de estudios fueron considerados en meta-análisis de datos de expresión génica previos[73,166], pero nunca han sido empleados los seis de forma simultánea. En la Tabla 1 se muestra un resumen de las características de cada uno de los estudios.

Conjunto de datos de GEO	Muestras sanas	Muestras LES	Tejido (tipo celular)	Plataforma de GEO	Número de genes
GSE11909_GPL96	10	118	Sangre periférica (PBMCs)	GPL96 (*microarray*)	13882
GSE11909_GPL97	7	53	Sangre periférica (PBMCs)	GPL97 (microarray)	11234
GSE24706	33	15	Sangre periférica (PBMCs)	GPL6884 (microarrray)	12467
GSE50772	20	61	Sangre periférica (PBMCs)	GPL570 (microarray)	21767
GSE82221_GPL10558	25	30	Sangre periférica (PBMCs)	GPL10558 (microarray)	14419
GSE122459	6	20	Sangre periférica (PBMCs)	GPL16791 (RNA-Seq)	24123

Tabla 1: Resumen de las características de cada uno de los estudios considerados. La tabla muestra un resumen de las características de estudios considerados: número de muestras sanas y de LES, tipo de tejido y tipo celular, el tipo de plataforma de GEO y el número total de genes considerados.

4.3. Resultados

4.3.1. Flujo de trabajo de GSEMA

El proceso metodológico para obtener las rutas diferencialmente reguladas mediante la metodología de GSEMA sigue una serie de pasos diferenciados (En la Figura 14 se puede ver un resumen esquemático de este flujo de trabajo).

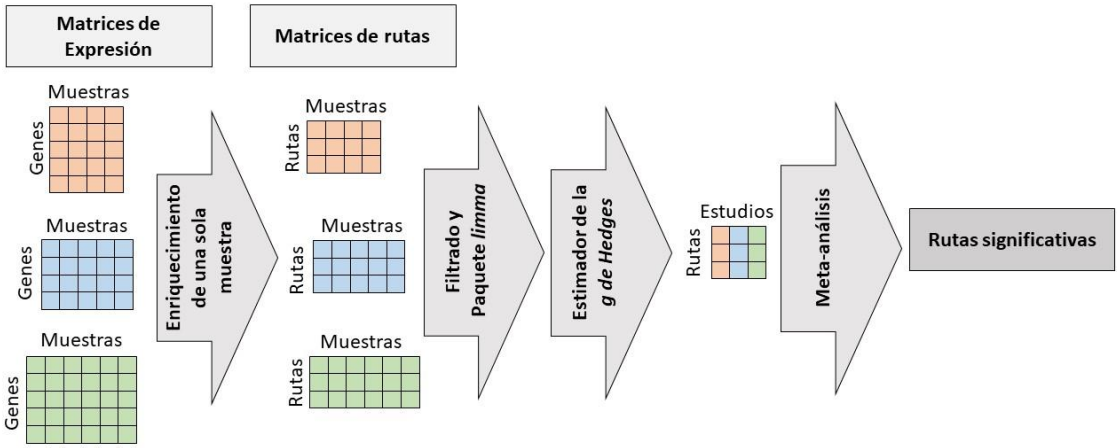


Figura 14. Resumen esquemático del flujo de trabajo de GSEMA. En la figura se representa de manera esquemática cada uno de los pasos que se llevan a cabo para aplicar la metodología de GSEMA.

En primer lugar, a partir de las matrices de expresión, $M_{G\times N}$, de cada uno de los estudios y haciendo uso de las técnicas de las técnicas de enriquecimiento de una sola muestra (*ssGSEA, Z-score, singscore, GSVA*) se obtienen las matrices las matrices de rutas, $R_{P\times N}$, de cada uno de los estudios. Posteriormente, las matrices de rutas se normalizan y se les aplica un filtro para descartar las rutas biológicas con un nivel bajo de expresión en el grupo experimental y control tal y como se describe en la sección 4.2.2.

En segundo lugar, se calculan los diferentes tamaños de efecto de cada una de las rutas en cada uno de los estudios. Para ello, se hace uso del paquete de R *limma*, en el que se obtienen los estadísticos de la *t de Student moderada* junto con sus correspondientes grados de libertad. A partir de estos valores se calculan los tamaños de efecto (*g de Hedges*) y sus correspondientes varianzas haciendo uso de la fórmula (4.1). Además, las varianzas de los tamaños de efecto son calculadas mediante la ecuación (4.2)

Para finalizar, se aplica un meta-análisis basado en la combinación de tamaños de efecto haciendo uso del modelo MEA y los p-valores obtenidos son corregidos por el método de *Benjamini & Hochberg*.

Esta metodología, así como funciones que permitan aplicar todos estos pasos fueron implementados en código de R y que está disponible de forma pública en https://github.com/Juananvg/GSEMA.

4.3.2. Evaluación comparativa de GSEMA

Para validar la su eficacia y consistencia, esta nueva metodología fue aplicada a datos simulados generados (sección 4.2.4.). Además, los resultados obtenidos fueron comparados con otros métodos previamente desarrollados en el campo. Específicamente, se consideraron los siguientes métodos:

- Procediendo habitual (denominado a partir de este momento ***MA_GSA)***: se aplica un meta-análisis de expresión génica basado en la combinación de tamaños de efectos haciendo uso del MEA y posteriormente se realiza un análisis de enriquecimiento funcional de los resultados.
- Métodos ***MAPE_G, MAPE_P y MAPE_I*** de *Kui Shen y George C. Tseng*[154].
- Meta-análisis haciendo uso de la metodología QuSAGE[157] (para abreviar a partir de ahora denominado solo ***QuSAGE***).
- Metodología GSEMA considerando las cuatro técnicas de enriquecimiento de una sola muestra descritas, GSVA, ssGSEA, Z-score y singscore. Por lo tanto, respectivamente nos referiremos a cada una de ellas como ***GSEMA_GSVA, GSEMA_ssGSEA, GSEMA_Z-score y GSEMA_singscore***.

Las rutas consideradas fueron obtenidas de la base de datos de GSEA, MSigDB[167–169]. En concreto se consideraron las rutas canónicas de los conjuntos de genes de humanos, siendo en total 3975 rutas. De estas rutas se eliminaron aquellas compuestas por menos de 7 genes, con el objetivo de mitigar el impacto de ellas en el número de falsos positivos. Para los análisis de enriquecimiento en MA_GSA y en las técnicas MAPE se aplicó el paquete de R de fgsea[170]. En todas las técnicas se consideraron como rutas significativas aquellas que obtuvieran un p-valor ajustado menor de 0.05 por el método de *Benjamini & Hochberg*.

Una vez aplicados los diferentes métodos a los datos simulados, se ordenaron las rutas obtenidas como significativas. En el caso de los métodos de GSEMA por su tamaño de efecto en valor absoluto (a mayor valor, más enriquecida está la ruta) y del mismo modo en el caso del resto de métodos se ordenan por en valor de enriquecimiento normalizado[170]. En la Tabla 2 se muestra un resumen de los resultados obtenidos.

Método	**Número de rutas significativas**	**Posición del "*Simulated_Pathway*"**
MA_GSA	1	1
MAPE-G	1	1
MAPE-P	1	1
MAPE-I	1	1
QuSAGE	1596	1
GSEMA_GSVA	1	1
GSEMA_Z-score	1	1
GSEMA_ssGSEA	8	1
GSEMA_singscore	10	1

Tabla 2: Resumen de resultados de los diferentes métodos de meta-análisis de enriquecimiento funcional. La columna *Posición del "Simulated_Pathway* indica la posición de la ruta *"Simulated_Pathway* ordenando por valor absoluto de tamaño de efecto o valor normalizado de enriquecimiento dentro de las rutas significativas obtenidas por el método

Los resultados obtenidos indican que todas las técnicas devuelven la ruta *"Simulated_Pathway"* como la ruta más importante. Sin embargo, GSEMA_ssGSEA y GSEMA_singscore muestran algunas rutas significativas más, aunque este número en proporción significante con respecto al número total de rutas consideradas. Esto mismo no ocurre con *QuSAGE*, la cual devuelve un número muy elevado de rutas significativas, lo que sugiere una tasa de falsos positivos relevante.

Para contrastar la consistencia de los resultados obtenidos por los diferentes métodos se llevó a cabo un análisis de la tasa de falsos positivos mediante un análisis de permutaciones. El problema de resultados con un alto número de falsos positivos es un problema recurrente en las técnicas de análisis de enriquecimiento[171,172]. En concreto, este análisis consistió en llevar a cabo 2000 simulaciones intercambiando aleatoriamente las condiciones (casos o controles) de las diferentes muestras. En el caso específico de GSEMA_ssGSEA, GSEMA_GSVA y *QuSAGE* sólo se realizaron 100 simulaciones, ya que estos métodos tienen un mayor coste computacional. Los resultados de un método serán más consistentes y fiables cuando obtengan menor número de rutas de manera aleatoria. En este contexto, se han considerado que una ruta es significativa si obtienen un p-valor sin ajustar menor de 0.05. En la Figura 15 se muestra un resumen del número de rutas obtenidas como significativas en cada una de las simulaciones por cada uno de los métodos.

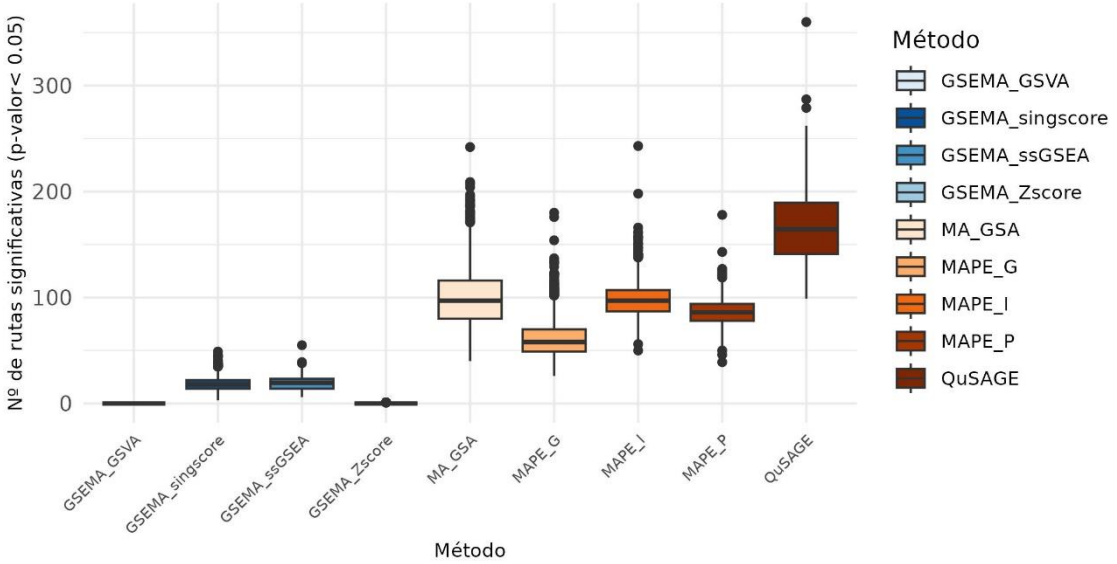


Figura 15: Diagrama de cajas del número de rutas significativas obtenidas por cada método en cada una de las simulaciones. Para todos los métodos se han aplicado 2000 simulaciones excepto para GSEMA_ssGSEA, GSEMA_GSVA y QuSAGE en los que sólo se han realizado 100 simulaciones. La figura muestra un diagrama de cajas de número de rutas significativas (p-valor <0.05) obtenidas por cada uno de los métodos

En este caso los resultados muestran un patrón diferente a los obtenidos en el análisis inicial. Mientras en la Tabla 2, se observaba que sólo *QuSAGE* mostraba una tasa elevada de falsos positivos, en el análisis de permutaciones los métodos *MAPE* y *MA_GSA* muestran un número de rutas significativas aleatorias superior a 50 en la mayoría de los casos, elevándose este número a más de 100 en el caso de *QuSAGE*. No obstante, en el caso de los diferentes métodos de GSEMA el número de rutas significativas aleatorias es mucho menor, lo que parece sugerir que controla mucho mejor el error tipo I, es decir, la

tasa de falsos positivos es mucho menor. Además, para el caso específico de la ruta *"Simulated_Pathway"* (Figura 16), se observa que el porcentaje de veces que se obtiene como significativa es mucho menor en los métodos GSEMA y *QuSAGE* (con un porcentaje que va del 0% al 5%) que en el caso de los métodos *MAPE* y *MA_GSA* en los cuales el porcentaje de veces que se obtiene como significativa es superior al 45%.

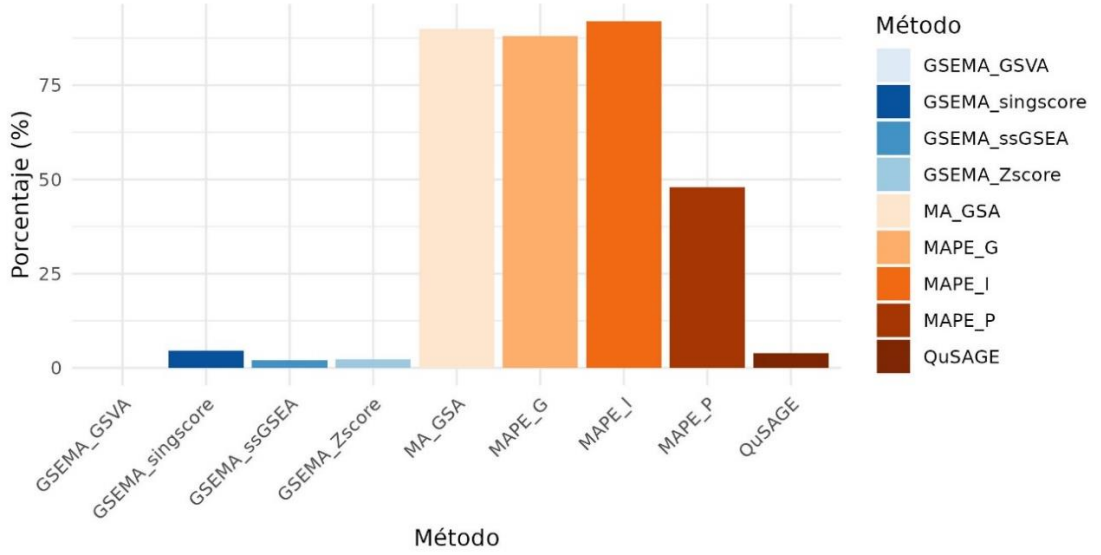


Figura 16: Diagrama de barras del porcentaje de veces que la ruta *"Simulated_Pathway"* es obtenida como significativa (p-valor < 0.05) en las permutaciones. El gráfico representa el porcentaje de veces que la ruta "*Simulated_Pathwat"* es obtenida como significativa (p-valor < 0.05) en cada uno de los métodos tras realizar las permutaciones.

Este resultado indica que la metodología GSEMA controla mucho mejor la tasa de falsos negativos que el resto de técnicas, devolviendo resultados mucho más consistentes al resto de metodologías.

4.3.3. Caso de uso con datos reales

GSEMA al pasar del espacio de genes al espacio de rutas puede ser de gran utilidad para combinar estudios que presenten una gran cantidad genes faltantes. Además, esto también puede resultar de gran relevancia a la hora de combinar estudios procedentes de diferentes plataformas en las que la heterogeneidad puede ser muy elevada. Para este caso se consideraron los estudios descritos en la sección 4.2.4.

Se aplicaron los mismos métodos que en la sección anterior, excepto el método *QuSAGE* que no pudo ser aplicado al no permitir la inclusión de estudios cuando estos presentan un alto número de genes no medidos. En el caso de los métodos *MAPE* y *MA_GSA* se trabajó con los genes comunes (3342), mientras con los métodos GSEMA si alguna ruta era filtrada en alguno de los estudios esta tampoco fue considerada en el meta-análisis.

4. META-ANÁLISIS BASADO EN INFORMACIÓN DE RUTAS Y PROCESOS BIOLÓGICOS

En este análisis, lo interesante es observar si las rutas más enriquecidas son estás relacionadas con la enfermedad del LES. Concretamente, al ser una enfermedad autoinmune, las rutas relevantes suelen estar relacionadas con el sistema inmune, como la respuesta inmune a enfermedades o la firma de interferón. En la Tabla 3 se muestran las rutas más enriquecidas obtenidas por cada uno de los métodos, tras ordenar en valor absoluto por tamaño de efecto o valor de enriquecimiento normalizado (VEN) las diferentes rutas significativas (p-valor ajustado < 0.05). Estos resultados muestran que en el caso de los métodos *MAPE* y *MA_GSA*, sólo el método *MAPE_P* obtiene, entre sus rutas más enriquecidas, rutas relacionadas con la enfermedad. El resto de métodos, aunque sus rutas pueden estar relacionadas con la enfermedad del LES (rutas sobre el ciclo y rutas ribosomales), estas no son tan relevantes para la enfermedad ni están relacionadas con el sistema inmune. Sin embargo, los métodos GSEMA si muestran rutas relacionados y relevantes con el sistema inmune, lo que demuestra una mejor conservación de la información y un mejor tratamiento de los valores faltantes. No obstante, es necesario mencionar que GSEMA_GSVA obtiene un número muy bajo de rutas significativas, lo que puede significar que es un método demasiado restrictivo en estos casos, lo que es un indicativo de una baja potencia o un error tipo II elevado.

4. META-ANÁLISIS BASADO EN INFORMACIÓN DE RUTAS Y PROCESOS BIOLÓGICOS

Método	Ruta	Expresión	Relación con el sistema inmune
MA_GSA	REACTOME_EUKARYOTIC_TRANSLATION_ELONGATION	Infraexpresada	No
	KEGG_MEDICUS_REFERENCE_TRANSLATION_INITIATION	Infraexpresada	No
	WP_CYTOPLASMIC_RIBOSOMAL_PROTEINS	Infraexpresada	No
	REACTOME_EUKARYOTIC_TRANSLATION_INITIATION	Infraexpresada	No
	REACTOME_RESPONSE_OF_EIF2AK4_GCN2_TO_AMINO_ACID_DEFICIENCY	Infraexpresada	No
MAPE_G	REACTOME_EUKARYOTIC_TRANSLATION_INITIATION	Infraexpresada	No
	WP_CYTOPLASMIC_RIBOSOMAL_PROTEINS	Infraexpresada	No
	KEGG_MEDICUS_REFERENCE_TRANSLATION_INITIATION	Infraexpresada	No
	REACTOME_EUKARYOTIC_TRANSLATION_ELONGATION	Infraexpresada	No
	KEGG_RIBOSOME	Infraexpresada	No
MAPE_P	REACTOME_INTERFERON_ALPHA_BETA_SIGNALING	Sobreexpresada	Sí
	REACTOME_NEUTROPHIL_DEGRANULATION	Sobreexpresada	Sí
	KEGG_MEDICUS_REFERENCE_TRANSLATION_INITIATION	Infraexpresada	Sí
	REACTOME_EUKARYOTIC_TRANSLATION_ELONGATION	Infraexpresada	Sí
	WP_CYTOPLASMIC_RIBOSOMAL_PROTEINS	Infraexpresada	Sí
MAPE_I	REACTOME_EUKARYOTIC_TRANSLATION_INITIATION	Infraexpresada	No
	WP_CYTOPLASMIC_RIBOSOMAL_PROTEINS	Infraexpresada	No
	KEGG_MEDICUS_REFERENCE_TRANSLATION_INITIATION	Infraexpresada	No
	REACTOME_EUKARYOTIC_TRANSLATION_ELONGATION	Infraexpresada	No
	KEGG_RIBOSOME	Infraexpresada	No
GSEMA_GSVA	WP_TYPE_II_INTERFERON_SIGNALING	Sobreexpresada	Sí
	WP_MYD88_DISTINCT_INPUT_OUTPUT_PATHWAY	Sobrexpresada	No
	KEGG_MEDICUS_REFERENCE_ORGANIZATION_OF_THE_OUTER_KINETOCHORE	Sobrexpresada	No
	WP_PHOTODYNAMIC_THERAPY_INDUCED_HIF_1_SURVIVAL_SIGNALING	Sobreexpresada	No
GSEMA_Zscore	WP_TYPE_I_INTERFERON_INDUCTION_AND_SIGNALING_DURING_SARS_COV_2_INFECTION	Sobrexpresada	Sí
	WP_TYPE_II_INTERFERON_SIGNALING	Sobrexpresada	Sí
	WP_NETWORK_MAP_OF_SARS_COV_2_SIGNALING_PATHWAY	Sobrexpresada	No
	REACTOME_INTERFERON_ALPHA_BETA_SIGNALING	Sobrexpresada	Sí
	WP_IMMUNE_RESPONSE_TO_TUBERCULOSIS	Sobrexpresada	Sí
GSEMA_ssGSEA	REACTOME_INTERFERON_ALPHA_BETA_SIGNALING	Sobrexpresada	Sí
	WP_TYPE_I_INTERFERON_INDUCTION_AND_SIGNALING_DURING_SARS_COV_2_INFECTION	Sobrexpresada	Sí
	WP_TYPE_II_INTERFERON_SIGNALING	Sobrexpresada	Sí
	WP_HOST_PATHOGEN_INTERACTION_OF_HUMAN_CORONAVIRUSES_INTERFERON_INDUCTION	Sobrexpresada	Sí
	WP_IMMUNE_RESPONSE_TO_TUBERCULOSIS	Sobrexpresada	Sí
GSEMA_singscore	WP_TYPE_I_INTERFERON_INDUCTION_AND_SIGNALING_DURING_SARS_COV_2_INFECTION	Sobrexpresada	Sí
	REACTOME_COMPLEMENT_CASCADE	Sobrexpresada	Sí
	WP_TYPE_II_INTERFERON_SIGNALING	Sobrexpresada	Sí
	WP_HOST_PATHOGEN_INTERACTION_OF_HUMAN_CORONAVIRUSES_INTERFERON_INDUCTION	Sobrexpresada	Sí
	WP_IMMUNE_RESPONSE_TO_TUBERCULOSIS	Sobrexpresada	Sí

Tabla 3: Resultados del meta-análisis de enriquecimiento de rutas a los datos de LES. La tabla recoge las 5 rutas con mayor tamaño de efecto o VEN obtenidas por cada método. Además, se indica si las rutas están relacionadas o no con el sistema inmune y si están sobrexpresadas o infraexpresadas.

4.4. Conclusiones

El meta-análisis de enriquecimiento de rutas mejora la conservación de las rutas biológicas implicadas al combinar distintos estudios, en comparación con el análisis de enriquecimiento basado en listas de genes obtenidas del meta-análisis de expresión génica. Sin embargo, muchas técnicas actuales para el meta-análisis de enriquecimiento de rutas emplean métodos de combinación de p-valores en lugar de tamaños de efectos, lo que provoca una pérdida de la direccionalidad en la regulación. Además, muchas de estas técnicas aplican el meta-análisis a la matriz de expresión en lugar de al espacio de rutas, lo que puede afectar los resultados debido a la presencia de genes faltantes.

Para resolver este problema, se desarrolló una metodología nueva llamada GSEMA. Esta metodología combina técnicas de enriquecimiento de una sola muestra con métodos de meta-análisis, permitiendo calcular un tamaño de efecto para cada ruta biológica y estudio. La aplicación de GSEMA a datos reales y simulados ha demostrado un mejor control de la tasa de falsos positivos en comparación con técnicas anteriores, mostrando mayor consistencia en los resultados. Además, GSEMA conserva mejor la información biológica al combinar estudios con una alta proporción de genes faltantes, haciéndola una herramienta valiosa para combinar estudios de bases de datos públicas provenientes de diversas plataformas de secuenciación.

5. INTEGRACIÓN DE DATOS EPIDEMIOLOGICOS EN COVID-19: EVALUACIÓN DEL EFECTO DE FACTORES AMBIENTALES EN PROPAGACIÓN DEL VIRUS

5.1. Conceptos previos

5.1.1 La enfermedad de COVID-19

La enfermedad por coronavirus 2019 (COVID-19) es una enfermedad infecciosa causada por el virus SARS-CoV-2, un nuevo tipo de coronavirus que fue identificado por primera vez en Wuhan, China, a finales de 2019[173]. Su rápida propagación a nivel mundial llevó a la Organización Mundial de la Salud (OMS) a declarar una pandemia en marzo del año 2020. En el caso de España, esto produjo que en marzo de 2020 el gobierno declarara un estado de alarma y un estricto confinamiento que duró hasta junio de ese año y con posteriores restricciones a lo largo de los siguientes meses.

Los síntomas de COVID-19 varían ampliamente y pueden incluir fiebre, tos, dificultad para respirar, fatiga, dolores musculares, pérdida del gusto o el olfato, y otros síntomas similares a los de otros virus respiratorios. En casos graves, la enfermedad puede llevar a neumonía, síndrome de dificultad respiratoria aguda (SDRA), insuficiencia multiorgánica y en casos extremos a la muerte. Al ser un virus respiratorio se transmite principalmente a través de las gotas respiratorias cuando una persona tose, estornuda o habla.

La pandemia de COVID-19 ha tenido un impacto profundo en la sociedad y la economía global. Las medidas de control, como el distanciamiento social y el uso de mascarillas, alteraron significativamente la vida cotidiana y han tenido repercusiones económicas y psicológicas de gran alcance. Por este motivo, desde el comienzo de la pandemia, diferentes científicos en todas partes del mundo han trabajado intensamente para entender mejor el virus, sus variantes y su capacidad de transmisión.

5.1.2. Importancia de los factores ambientales en la trasmisión del virus

Desde el comienzo de la pandemia, uno de los mayores esfuerzos llevados a cabo por los investigadores es el estudio de la relación entre los diferentes factores ambientales y la transmisión de la enfermedad. El conocimiento de esta asociación es imprescindible para entender la propagación de la enfermedad y las medidas gubernamentales y de control necesarias para controlar esta y futuras pandemias.

Una de las relaciones más estudiadas a lo largo de la pandemia, han sido los patrones estacionarios de la transmisión de la enfermedad, la cual, al igual que otros virus respiratorios como la gripe, se sospecha presente una mayor capacidad de transmisión ambientes secos y fríos [174–176].

Sin embargo, en este campo a pesar de haberse realizado numerosos estudios, en la mayoría de ellos se han encontrado muchos resultados contradictorios e inconsistentes. Mientras algunos muestran una correlación negativa[177–179] entre la temperatura y la

transmisión del virus, otros muestran una correlación positiva[180] o directamente que no existe ningún tipo de relación[181,182]. Esta incoherencia entre los resultados, pueden deberse a varios factores: 1º) la introducción de sesgos en los resultados a consecuencia de un uso inadecuado de las diferentes metodologías estadísticas[183–186]; 2º) el análisis de períodos iniciales de la pandemia en las que la mayoría de casos no eran reportados de manera adecuada[187,188]; 3º) la omisión de variables que pueden ser relevantes en la transmisión de la enfermedad como las medidas adoptadas por los diferentes gobiernos para controlar la propagación de la enfermedad o la aparición de nuevas variantes que influyen en la capacidad de transmisión del virus[189–191].

Uno de las variables que puede afectar considerablemente en los resultados es la falta de inmunidad poblacional al comienzo de la pandemia, la cual es un factor determinante en la transmisión del virus[192,193]. Esta hipótesis fue introducida por *Baker et al.* en un estudio publicado en *Science* [193,194] donde, usando datos de otros coronavirus humanos (HKU1 y HCoV-OC43), ya que ese momento aún no se disponía de datos suficientes en SARS-CoV-2, desarrolló un modelo que demostró que aunque los factores ambientales como la temperatura pueden afectar en la transmisión de enfermedad, es la baja inmunidad el principal factor que impulsa la transmisión del virus y lo que provoca una mitigación del efecto del resto de variables ambientales. Resultaba por tanto interesante disponer de datos actualizados en COVID-19 para poder analizar esta influencia más allá del plano teórico.

Estos son los objetivos de los artículos *Martorrel-Marugán, Villatoro-García et al.*[195] y *Villatoro-García et al.*[196]. El objetivo del primer trabajo fue el de recopilar, estandarizar e integrar datos de evolución del COVID-19 y factores ambientales con el fin de disponer de un recurso centralizado de información curada en España para llevar a cabo análisis sobre el impacto de factores como temperatura o humedad en la dispersión del virus. Posteriormente, aprovechando toda la información recopilada se analizaron la influencia de estos factores ambientales en la dispersión del virus teniendo en cuenta la inmunidad poblacional. En las siguientes secciones explicaremos los principales métodos empleados, así como los resultados más relevantes. Ambos artículos se proporcionan en los Anexos (9.3) y (9.4) de esta tesis doctoral una información mucho más amplia.

5.2. Metodología

5.2.1. Procesamiento de datos de COVID-19 y factores ambientales

Desde el comienzo de la pandemia, para recopilar y almacenar toda esta información referente a la evolución de la enfermedad, se crearon numerosas bases de datos. A nivel mundial, una de las más importantes fue el *John Hopkins University Coronavirus Resource Centre*[197] que desde principios de febrero de 2020 comenzó a recoger diariamente la información publicada por distintos países sobre el número de casos y muertes. Más adelante, adaptaron la base de datos para incluir también información sobre el número de dosis de vacunas administradas. Otra base de datos que ganó mucha popularidad fue *Our World in Data*, la cual recopila una gran cantidad de datos a nivel mundial y se adaptó para registrar información acerca de la pandemia[198]. En particular, esta base de datos hizo un gran esfuerzo en recopilar información sobre la vacunación a

nivel mundial, como el número de vacunas administradas, dosis recibidas y personas con pautas de vacunación completas[199]. [200].

Todas estas bases de datos son un gran fuerte de información para estudiar la transmisión del COVID-19 y adquirir nuevos conocimientos sobre la evolución del virus.

Durante esta tesis doctoral se recopiló información de diferentes bases de datos. En concreto distinguimos dos niveles de información. Por un lado, se recopiló información a nivel de España considerando sus diferentes comunidades y provincias. Por lo otro lado, se recopiló información a de Europa, considerando principalmente países mediterráneos y regiones de Italia.

5.2.1.1. Recopilación de datos a nivel de España

Para la integración de datos de COVID-19 y los diferentes factores ambientales a nivel de España se recopiló información procedente de diferentes fuentes y bases de datos.

En lo que respecta a los datos epidemiológicos, el número de casos totales acumulados, el número de casos diarias detectados y el número de casos detectado por la reacción en cadena de la polimerasa (PCR) de cada una de las comunidades autónomas y de las provincias fueron extraídos de la aplicación *PANEL COVID-19*[201] desarrollada y mantenida por el Ministerio de Sanidad de España y por el Instituto de Salud Carlos III. El número de defunciones diarias, el acumulado de pacientes hospitalizados, el acumulado de pacientes trasladados a unidades de cuidados intensivos (UCI) y el acumulado de pacientes recuperados de las diferentes comunidades autónomas fueron recopiladas del repositorio de datos Datadista[202]. De este repositorio de datos también se descargó la información referente al porcentaje diario de personas vacunadas con al menos dos dosis de vacunas.

Los datos climatológicos fueron descargados de la Agencia Estatal de Meteorología de España (AEMET)[203]. En concreto se descargaron datos diarios procedentes de las diferentes estaciones meteorológicas referentes a la temperatura media diaria (Cº), lluvia diaria (l/m^2), velocidad del viento (m/s) y horas diarias de insolación. No obstante, al no proporcional la AEMET los datos referentes a la humedad, del *European Centre for Medium-Range Weather Forecast ERA5 climate reanalysis*[204] se extrajeron datos referentes a la humedad específica (Kg/kg) de cada hora en cada día en cada comunidad.

Se recopiló también información acerca de datos de calidad del Aire. Para la comunidad autónoma de Andalucía y sus provincias se extrajeron los datos de la Consejería de Agricultura, Ganadería, Pesca y Desarrollo Sostenible de la Junta de Andalucía[205]. Para el resto de comunidades y provincias españolas fueron descargados del Portal Europeo de Calidad del Aire (*European Air Quality Portal)*[206]. Se recopilaron datos diarios referentes a las concentraciones (ug/m^3) de los gases contaminantes NO_2, CO, $PM_{2.5}$, PM_{10}, SO_2 y O_3.

Por último, del *Oxford COVID-19 Government Response Tracker* (OxCGRT)[200] se descargó información referente a las medidas adoptadas por el gobierno para controlar la pandemia. En concreto, se extrajo el índice de rigor o *Stringency Index* (SI) el cual mide del 0 al 100 el nivel de restricción de un gobierno, donde 0 indica ninguna restricción y 100 la máxima restricción posible. El índice se calcula a partir de nueve indicadores,

siendo estos: cierre de lugares de trabajo, cancelación de eventos públicos, restricciones a las reuniones públicas, cierre del transporte público, obligación de quedarse en casa, campañas de información pública, restricciones a los desplazamientos internos y controles de viajes internacionales.

5.2.1.2. Recopilación de los datos del resto de países

A parte de España, también se recopiló información del resto de países desde diferentes fuentes de información y bases de datos.

En esta ocasión, los datos referentes a la evolución de la pandemia por COVID-19 fueron descargados del *John Hopkins University Coronavirus Resource Centre*[197] y de *Our World in Data*[198,199]. Al igual que en el caso de España se recopilaron de ambas bases de datos los datos necesarios para obtener el número de casos diarios de COVID-19, el número de muertes por COVID-19 diarias, el porcentaje diario de personas vacunadas con al menos dos dosis.

Los datos climatológicos se descargaron del *European Centre for Medium-Range Weather Forecast ERA5 climate reanalysis*[204], obteniéndose la temperatura y la humedad específica media diarias al igual que en el caso de España.

Por último, el SI de los diferentes países también fue obtenido del *Oxford COVID-19 Government Response Tracker (OxCGRT)*[200].

5.2.1.3. Agregación de datos y cálculo de indicadores

La información obtenida de las diferentes bases de datos fue procesada y sometida a rigurosos controles de calidad para evitar la inclusión de datos erróneos. Se evitó la inclusión de valores negativos en variables cuantitativas discretas, como el número de casos o defunciones diarias, y se descartaron valores irreales en los niveles de algunos contaminantes ambientales.

Para agregar la información por comunidades y provincias de los factores ambientales, se calcularon las medias aritméticas de las diferentes estaciones meteorológicas pertenecientes a cada región. En el caso de los datos de COVID-19, cuando se disponía de datos diarios, los datos acumulativos se calcularon sumando los datos de los días anteriores. Del mismo modo, si se disponía de información acumulativa, se obtuvieron datos diarios restando el valor del día anterior del valor del día de referencia.

Una vez aplicados estos pasos, se determinaron nuevas métricas epidemiológicas a partir de los datos procesados. Se calculó la tasa de letalidad o el *case fatality rate* (*CFR*) dividiendo el número de fallecimientos acumulados (*F*) entre el número acumulado de enfermos (casos) acumulados:

$$CFR = \frac{F}{E} \times 100 \tag{5.1}$$

Del mismo modo se calculó la tasa de enfermos recuperados o *case recovery rate* (*CRR*):

$$CRR = \frac{R}{E} \times 100 \tag{5.2}$$

Donde *R* es el número de personas recuperadas de la enfermedad. Además, se calcularon las tasas de incidencia acumulada o *cumulative incidence rate* (*CI*). Esta tasa mide el número de personas que se han contagiado en un período de tiempo determinado respecto de la población total:

$$CI_x = \frac{E_x}{P_x} \times 100000 \tag{5.3}$$

Donde *x* representa el período de tiempo para la que se calcula la tasas, E_x el número de personas que han enfermado durante el período *x* y P_x la población media durante el período *x*. el número enfermos. Específicamente se calcularon las tasas para grupos de 7 y 14 días.

Por último, se estimó el número reproductivo efectivo (R_e) diario, el cual mide el número medio de contagios secundarios producidos por una persona infectada. El número reproductivo básico es un gran indicar de la evolución de la pandemia y del número de contagios, debido a que valores mayores de 1 indican una propagación rápida de la enfermedad y por el contrario valores menores de 1 indican una reducción de la propagación. Para su estimación se aplicó el paquete de R EpiEstim[207] el cual estima el R_e a partir de un enfoque bayesiano considerando que la el contagio de un infectado se modelo mediante un proceso de Poisson. En nuestro caso consideramos un intervalo de serie incierto con una media de 4.7 días y una desviación típica de 2.9 basándonos en el trabajo de *Baker et al.*[194].

5.2.2. Modelos para medir la relación entre la trasmisión del virus y los factores meteorológicos

Para medir la influencia de los factores meteorológicos en la transmisión de la enfermedad se consideraron dos modelos diferentes: los Modelos Aditivos Generalizados (GAM por sus siglas en inglés) y los Modelos no lineales de desfase distribuido (DLNM por sus siglas en inglés).

5.2.2.1. Datos y períodos considerados

En el análisis de la relación entre los factores ambientales y la evolución de la transmisión de la enfermedad se consideraron datos de las regiones de España y países europeos.

En el caso específico de España se consideraron 16 de las comunidades autónomas, excluyendo las Islas Canarias por su clima subtropical con temperaturas más suaves, lo que hace que su información sea menos comparable con el resto de comunidades. En el caso de las ciudades autónomas de Ceuta y Melilla, éstas fueron descartadas por tener un menor tamaño poblacional.

En lo que respecta a los países europeos, para que el clima fuera similar al de las comunidades españolas, se recopiló información acerca de países en latitudes similares a España y próximos al mar mediterráneo. Específicamente, los países considerados fueron: Francia, Portugal, Italia, Grecia, Eslovenia, Croacia, Serbia y Montenegro. Otros países con latitudes similares a los demás, como Albania y Macedonia del Norte, no fueron elegidos debido a la inferior calidad de los datos comunicados, como la presencia de casos diarios iguales a 0, que no se corresponde con la realidad de la pandemia del COVID-19. Además, como en el caso de Italia se disponía de datos regionales y sus condiciones

climáticas son muy similares a la de España, se recabó información sobre sus regiones. Se consideraron 19 de las 20 regiones italianas, excluyendo el Valle de Aosta debido a su clima significativamente más frío en comparación con el resto de las regiones, con una temperatura media inferior a 20°C debido a su proximidad con los Alpes.

Además, para considerar la influencia de la inmunidad adquirida se consideraron dos períodos diferentes en el análisis. Un primer período (P1) caracterizado por un bajo nivel de inmunidad poblacional que abarca desde el 1 de junio de 2020 al 31 de diciembre de 2020 y un segundo período (P2) comprendido entre el 1 de junio de 2021 al 31 de diciembre de 2021 en el cual gracias a los esfuerzos de vacunación ya se había alcanzado cierto grado de inmunidad en la población. Esto períodos fueron seleccionados de manera específica debido a que son las mismas fechas en dos años diferentes, lo que los hace mucho más comparables.

5.2.2.2. Modelos aditivos generalizados para el estudio de la relación entre los factores meteorológicos y la evolución del R_e

Los Modelos Aditivos Generalizados o por sus siglas en inglés modelos GAM (*Generalized additive model*) son una extensión de los modelos lineales generales en los que se permite la inclusión de relaciones no lineales entre las variables independientes y la variable respuesta[208]. En estos modelos las relaciones no lineales se incluyen como funciones suaves de las variables independientes, por lo que la forma general del modelo con *n* variables independientes:

$$g\left(E\left[Y\right]\right) = \beta_0 + f_1\left(X_1\right) + f_2\left(X_2\right) + \ldots + f_n\left(X_n\right) \tag{5.4}$$

En nuestro estudio para estimar la compleja relación entre un factor meteorológico y el R_e se consideró el siguiente modelo GAM para cada comunidad española, cada país europeo y cada región de Italia:

$$R_{e_{i,t}} = \beta_0 + s\left(MV_{i,t}\right) + \beta_1\left(VR_{i,t}\right) + \beta_2\left(SI_{i,t}\right) + V_t \tag{5.5}$$

donde:

$R_{e_{i,t}}$ es el número reproductivo efectivo en el día *t* en la región *i*.

$VM_{i,t}$ es la variable meteorológica en el día *t* en la región *i*.

$VR_{i,t}$ es la tasa de vacunación en el día *t* en la región *i*.

$SI_{i,t}$ es el *stringecy index* en el día *t* en la región *i.*

V_t es una variable categórica que indica la variante del COVID-19 dominante en el día *t*.

En este modelo, la variable meteorológica considerada (temperatura o humedad específica) es incluida como un spline cúbico (*s*). Además, se incluye la variante dominante en cada instante de tiempo (considerando que en todas las regiones en estudio la variante predominante es la misma) con el objetivo de considerar en el modelo el posible efecto de las distintas variantes en la transmisión de la enfermedad.

Además, las distintas variables independientes se incluyeron con ciertos desfases o *lags* para controlar el intervalo de tiempo entre la infección y la detección de la enfermedad.

En este caso, se consideraron diferentes *lags* para cada uno de los períodos en estudio debido a las características diferenciales en la detección de la enfermedad entre ambos. Concretamente, en P1 se empleó un *lag* de 14 días debido a que el tiempo entre el contagio y el reporte de los diferentes casos era sustancialmente largo llegando a extenderse entre 10 y 15 días tal y como indicaban estudios previos que abordaron este hecho[184,190,209]. No obstante, en P2, consideramos un *lag* de 7 días (una semana), motivado por las mejoras en las técnicas de detección y la mayor rapidez de los organismos a la hora de publicar los casos detectados.

Los diferentes modelos GAM de cada una de las regiones fueron calculados haciendo uso del paquete de R *mgcv*[210,211]. Además, los p-valores fueron corregidos haciendo uso del método de Benjamini & Hochberg[42].

5.2.2.3. Cuantificación del efecto de los diferentes factores meteorológicos en el riesgo de transmisión de la enfermedad

Una vez estudiadas las relaciones de los factores meteorológicos, el siguiente objetivo era medir la magnitud de la influencia y cómo esta podía afectar a la evolución de la pandemia. Por este motivo, se llevó a cabo un análisis de dos etapas similar al aplicado por *Nottmeyer et al*[184]. En primer lugar, se cuantificó el efecto de manera individual por cada región y posteriormente mediante un meta-análisis se combinaría los diferentes efectos para obtener un efecto global a mayor escala.

En la primera etapa, para cuantificar el efecto en cada una de las regiones de manera individual en cada uno de los períodos en estudio se implementó un modelo de modelo no lineal de desfase distribuido o por sus siglas en inglés **DLMN** (*distributed lag non-linear* model)[212]. Estos modelos fueron seleccionados ya que aparte de incluir relaciones no lineales, además consideran que las variables independientes pueden tener efectos retardados o *lag* sobre la variable respuesta y han sido muy empleados en el campo de la epidemiología medioambiental[213,214]. En nuestro caso hacemos uso del siguiente modelo:

$$R_{e_{i,t}} = CB\left(MV_{i,t}\right) + CB\left(SI_{i,t}\right) + V_t + Ind\left(Vac\right) + \text{int} + NS\left(fecha, df = 2\right) \qquad (5.6)$$

donde:

$CB\left(MV_{i,t}\right)$ es la matriz de base cruzada (cross-basic en inglés) para la variable meteorológica en el día *t* en la región *i*. Esta matriz, es una matriz de diseño que permite incorporar un *lag* con un rango de entre 7 a 14 días, permitiendo un mejor control del *lag* que en el caso de los modelos GAM.

$CB\left(SI_{i,t}\right)$ es la matriz de base cruzada para el *SI* en el día *t* en la región *i*. Su cálculo es similar al $CB\left(MV_{i,t}\right)$, permitiendo introducir un *lag* que va desde los 7 a los 14 días.

V_t es una variable categórica que indica la variante del COVID-19 dominante en el día *t*.

$Ind\left(Vac\right)$ es una variable binaria que presenta la falta o la presencia de vacunación.

int es un término iterativo que indica si el período es pre o post vacunación.

$NS\left(fecha, df = 2\right)$ es un término que modula la tendencia intraperiódica de la evolución de COVID-19. En este caso se considera una función spline natural de la fecha con 2

grados de libertad (*df*), lo que equivale aproximadamente a 1 grado de libertad por cada tres meses.

Para la construcción del modelo se hizo uso del paquete de R *dlmn*[215]. Se asumió que la varianza residual de $R_{e_{i,t}}$ se distribuye según una distribución de quasi-Poisson. Posteriormente para cada región y período se obtuvieron las curvas de la evolución (curvas de asociación) del R_e. Estas curvas de asociación fueron obtenidas calculando el riesgo relativo (RR) a partir de modelo tomando como referencia la media de la variable meteorológica.

Por último, en la segunda etapa, se aplicó un meta-análisis de los diferentes efectos (RR) con el objetivo de obtener la evolución del efecto a nivel global. Para ello se hizo uso del paquete de R *mvmeta*[216].

5.3. Resultados

5.3.1. Aplicación DatAC

Como se mencionó anteriormente, desde el inicio de la pandemia, muchos científicos y responsables públicos han realizado numerosos esfuerzos para analizar la relación entre la transmisión del COVID-19 y diversos factores ambientales. En este contexto, se crearon numerosas aplicaciones para recopilar datos sobre la evolución de la pandemia[197,217]. Sin embargo, en cuanto a la relación de toda esta información con las variables medioambientales, no se había desarrollado ninguna herramienta que proporcionara esta información de manera conjunta y ayudara a los investigadores en el análisis de estas relaciones. Con este objetivo, se desarrolló DatAC, una aplicación web que integra datos epidemiológicos de COVID-19 junto con datos medioambientales. DatAC proporciona una agregación espacio-temporal de todas estas fuentes de información a nivel de provincias y comunidades de España. En el momento de su desarrollo, DatAC fue la primera herramienta que integraba toda esta información. DatAC está disponible en https://covid19.genyo.es/. Para el desarrollo de DatAC se ha hecho uso del paquete de creación de aplicaciones web interactivas Shiny de R.

DatAC recopila datos curados procedentes de diferentes fuentes de información desde el 1 de enero de 2020 que se han ido actualizando de manera progresiva. Contiene desde datos epidemiológicos diarios y acumulados como por ejemplo el número de casos de COVID-19 detectados o el número de personas fallecidas a causa de enfermedad. Además, en lo que respecta a la información medioambiental, contiene variables meteorológicas diarias (temperatura, lluvia, velocidad del viento y radiación solar) y variables diarias relativas a la calidad del aire (NO_2, CO, $PM_{2.5}$, PM_{10}, SO_2 y O_3) disponibles tanto para estaciones meteorológicas urbanas, suburbanas y rurales.

La aplicación comprende tres módulos principales. El primero es el módulo *Map* (mapa), que agrega datos con información espacial y permite al usuario visualizarlos y generar gráficos a partir de las variables seleccionadas. En segundo lugar, está el módulo *Trend Analysis* (análisis de tendencias), que permite al usuario aplicar diferentes modelos de regresión y correlaciones para estudiar la relación entre dos variables. Por último, el módulo *Temporal Trends* (tendencias temporales) permite comparar dos variables en diferentes regiones a lo largo del tiempo. Además, tanto en *Trend Analysis* como en

Temporal Trends, la aplicación ofrece la posibilidad de añadir retrasos (lag) a alguna de las variables para estudiar la relación en diferentes instantes de tiempo. En la Figura 17 pueden observarse representaciones gráficas de los diferentes módulos.

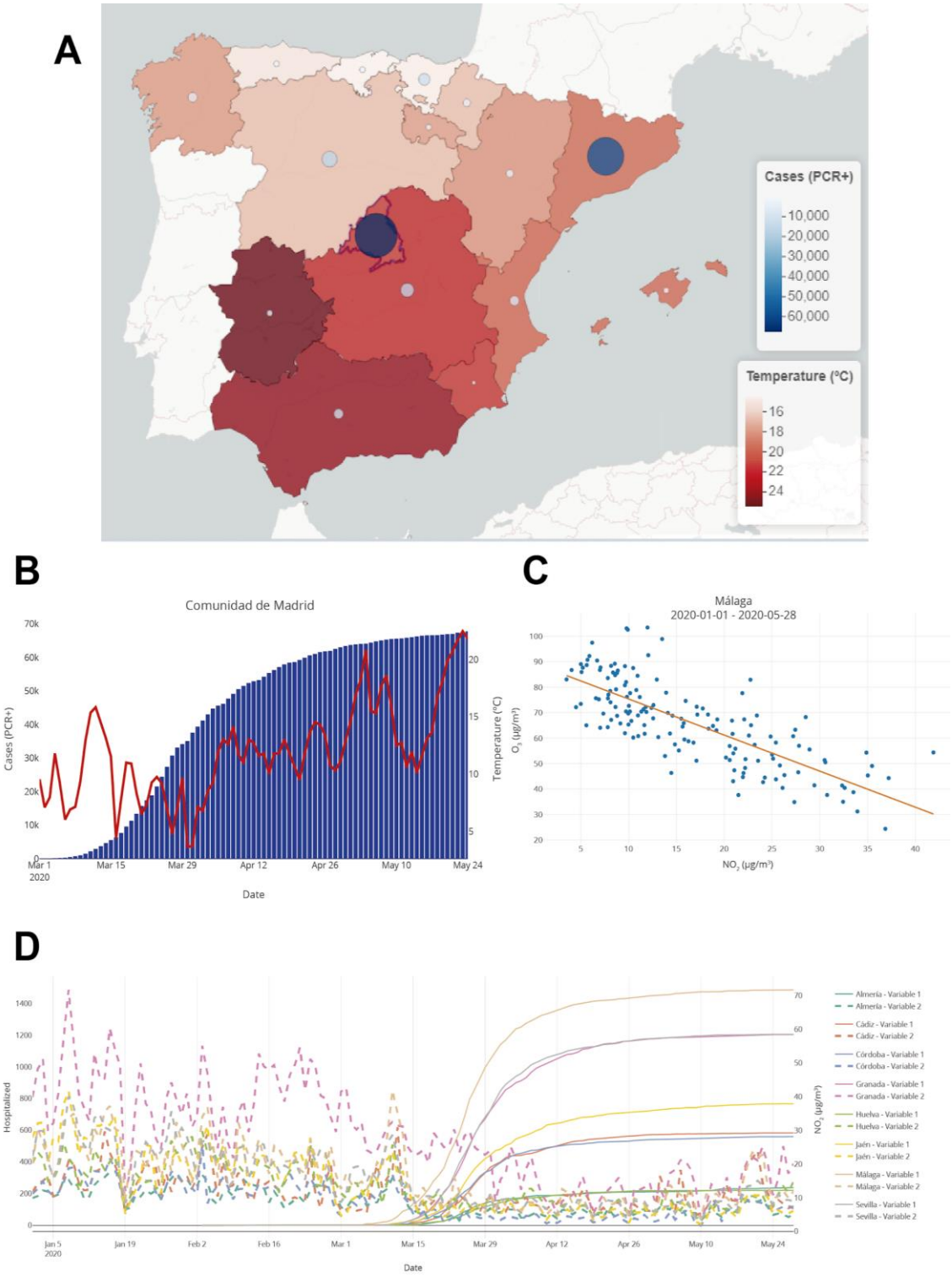


Figura 17. Ejemplo de resultados de DatAC. A) Mapa con los casos de COVID-19 confirmados por la prueba PCR (círculos azules) y la temperatura media (color de fondo) para las comunidades autónomas de España el 24 de mayo de 2020. B) Gráfico longitudinal con las mismas variables que A) para la Comunidad de Madrid desde el 1 de marzo hasta el 24 de mayo de 2020, representando los casos con barras y la temperatura con la línea roja. C) Gráfico de correlación entre las concentraciones de NO_2 y O_3 para la provincia de Málaga desde el 1 de enero hasta el 28 de mayo de 2020.

D) Gráfico longitudinal representando los pacientes hospitalizados (líneas continuas) y la concentración de NO_2 (líneas semicontinuas) para todas las provincias andaluzas desde el 1 de enero al 24 de mayo de 2020. Figura extraída del artículo *Martorell-Marugán, Villatoro García et al.*[195].

Los contenidos en la aplicación, así como el código fuente de la misma se encuentra disponibles bajo licencia libre a través del repositorio de GitHub (https://github.com/GENyO-BioInformatics/DatAC). Además, se puede encontrar información más detallada tanto de la aplicación en el artículo de *Martorell-Marugán et al*[195], proporcionado en el Anexo 9.3. y que es uno de los artículos que avala esta tesis.

5.3.2. Efecto de la temperatura y humedad en la transmisión del virus SARS-CoV-2

En el artículo *Villatoro-García et al.*[196] a partir de toda la información recopilada en DatAC para España y la extraída para otros países, se realizó un estudio de la influencia de los factores meteorológicos en la transmisión del virus SARS-CoV-2 en España y varios países mediterráneos. El objetivo fue evaluar las hipótesis del artículo de *Baket et al*. publicado en la prestigiosa revista *Science*, en el cual se teorizaba que el efecto de los factores meteorológicos sobre la transmisión del COVID-19 sólo tendría lugar cuando existiera una alta tasa de inmunidad poblacional. En las siguientes secciones se hace una descripción de los principales resultados obtenidos para la variable de temperatura media diaria, aunque una información más detallada y relativa a la humedad específica puede encontrarse en el Anexo 9.4 de esta tesis doctoral.

5.3.2.1. Estudio de relación entre la temperatura y la transmisión del virus en España

El análisis de la influencia de la temperatura en la evolución del R_e (transmisión del virus) se llevó a cabo mediante modelos aditivo no generalizados (GAMs), aplicados a cada una de las regiones españolas en los dos períodos anteriormente especificados. Los resultados de los modelos aplicados a cada una de las comunidades considerando la temperatura como factor meteorológico pueden observarse en la Tabla 4.

P1 (1 June 2020 to 31 December 2020)						
Autonomous community	**Deviance explained**	**R^2**	**Corrected P-values**			
			Temperature	**SI**	**Vaccination Rate**	**Variants**
Andalucía	38.54	0.36	<0.0001	<0.0001	NA	0.0731
Aragón	43.72	0.42	0.0050	<0.0001	NA	0.0413
Cantabria	13.32	0.12	0.0171	0.0002	NA	0.3109
Castilla y León	36.98	0.35	<0.0001	0.1917	NA	0.0731
Castilla-La Mancha	47.41	0.45	<0.0001	0.7907	NA	0.0614
Cataluña	62.55	0.61	<0.0001	<0.0001	NA	0.1952
Madrid	61.21	0.60	<0.0001	0.3729	NA	0.9196
Navarra	46.07	0.45	0.6030	<0.0001	NA	0.0413
C. Valenciana	29.67	0.28	0.0095	<0.0001	NA	0.2570
Extremadura	23.42	0.21	0.0009	<0.0001	NA	0.0824
Galicia	20.24	0.17	0.0030	<0.0001	NA	0.9196
Islas Baleares	11.44	0.09	0.0046	0.6126	NA	0.0731
La Rioja	38.83	0.37	0.0001	<0.0001	NA	0.0739
País Vasco	35.95	0.35	0.1301	<0.0001	NA	0.9196
Asturias	27.32	0.26	0.5348	<0.0001	NA	0.5804
Murcia	40.01	0.38	0.0001	<0.0001	NA	0.0001
Median	**37.76**	**0.36**	**0.0019**	**<0.0001**	**NA**	0.0782
P2 (1 June 2021 to 31 December 2021)						
Autonomous community	**Deviance explained**	**R^2**	**Corrected p-values**			
			Temperature	**SI**	**Vaccination Rate**	**Variants**
Andalucía	81.79	0.81	<0.0001	<0.0001	<0.0001	<0.0001
Aragón	56.17	0.55	<0.0001	<0.0001	<0.0001	<0.0001
Cantabria	59.54	0.58	<0.0001	<0.0001	<0.0001	<0.0001
Castilla y León	60.01	0.59	<0.0001	<0.0001	<0.0001	0.0007
Castilla-La Mancha	85.91	0.85	<0.0001	<0.0001	<0.0001	<0.0001
Cataluña	53.58	0.52	<0.0001	<0.0001	<0.0001	0.7814
Madrid	74.87	0.74	<0.0001	<0.0001	<0.0001	0.0001
Navarra	44.45	0.43	<0.0001	<0.0001	<0.0001	0.0011
C. Valenciana	68.00	0.67	<0.0001	<0.0001	<0.0001	<0.0001
Extremadura	76.66	0.76	<0.0001	<0.0001	<0.0001	0.0011
Galicia	72.83	0.72	<0.0001	<0.0001	<0.0001	0.7813
Islas Baleares	70.69	0.70	<0.0001	<0.0001	<0.0001	0.8998
La Rioja	66.19	0.65	<0.0001	<0.0001	<0.0001	0.0023
País Vasco	69.08	0.68	<0.0001	<0.0001	<0.0001	0.0873
Asturias	49.90	0.48	<0.0001	<0.0001	<0.0001	0.7848
Murcia	65.62	0.65	<0.0001	<0.0001	<0.0001	<0.0001
Median	**67.09**	**0.66**	**<0.0001**	**<0.0001**	**<0.0001**	**0.0009**

Tabla 4. Resultados de los modelos GAM con Temperatura como variable meteorológica para las diferentes comunidades autónomas españolas. La primera y segunda columnas representan la desviación explicada y el R^2 ajustado de los diferentes modelos. El resto de columnas representan los p-valores corregidos para las diferentes variables incluidas en los modelos. Los valores en negrita representan la mediana de las diferentes comunidades. Tabla adaptada del artículo *Villatoro García et al.*[196].

Los resultados de los modelos para P1 parecen indicar una asociación entre la temperatura y el R_e en la mayoría de las comunidades, siendo la gran parte de los p-valores corregidos obtenidos menores de 0,05 (con una mediana de 0.002). Sin embargo, si aislamos la

influencia del resto de variables y representamos la distribución de la temperatura y el R_e para los modelos GAM, no se observa una relación clara, llegando a haber un aumento de la transmisión a partir de los 17-20 °C (Figura 18).

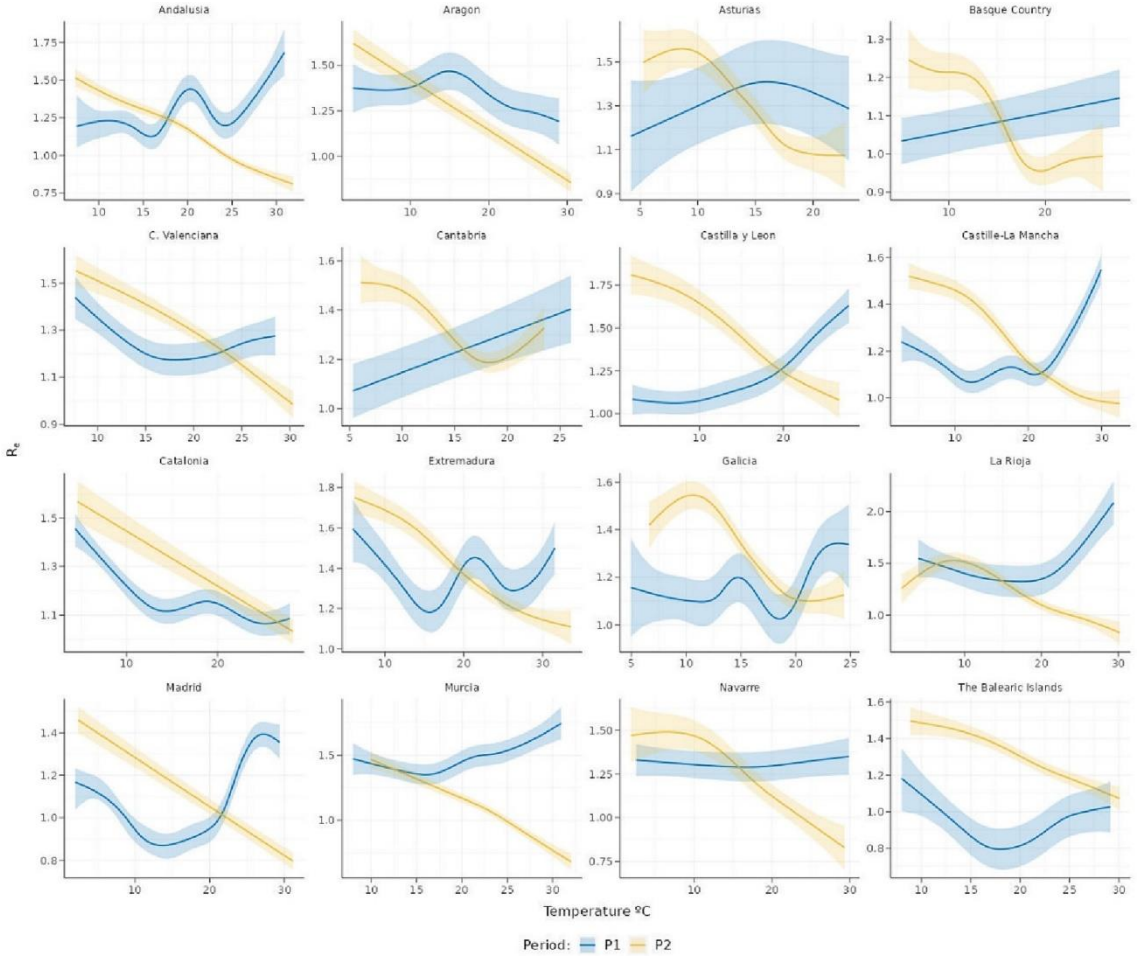


Figura 18. Estimación de la evolución R_e a partir de la influencia de la temperatura (°C) predicha por los modelos GAMs aislando el resto de variables en las diferentes comunidades autónomas de España. Representación gráfica de la R_e predicha por los GAMs con la temperatura como variable meteorológica y considerando el resto de variables constantes para las diferentes comunidades españolas. El color azul representa el periodo sin vacunación (P1) y el amarillo el periodo con vacunación (P2). Figura extraída del artículo *Villatoro-García et al. 2023*[196].

En lo que respecta a los resultados para P2, estos muestran también una asociación entre la temperatura y el R_e para la mayoría de comunidades (mediana de p-valores corregidos <0.001). Sin embargo, para este período cuando se representa la distribución de la temperatura y del R_e aislando la influencia del resto de variables si se observa una disminución de la transmisión con el aumento de la temperatura a excepción de algunas comunidades del norte de España (Cantabria, País Vasco y Galicia), en las cuales se observa un aumento por encima de los 20ºC, aunque esto puede deberse principalmente a sus veranos más suaves. Por lo tanto, en términos generales la transmisión del SARS-Cov-2 fue mayor en temperaturas más frías que en las más cálidas.

Además, si observamos el R^2 ajustado y el porcentaje de desviación explicada, en P2 sus valores son considerablemente mayores que en P1 para la mayoría de comunidades. Sin embargo, la mayoría de R^2 ajustado no son superiores a 0.75, por lo que el porcentaje de variabilidad explicada es moderado, lo que indica que pueden existir más factores de los incluidos que pueden afectar a la transmisión del virus. A pesar de ello, estos resultados

parecen indicar que un aumento de la temperatura puede conducir a una ligera reducción de la transmisión del virus cuando una parte significativa de la población está inmunizada.

Posteriormente, una vez confirmada que existe cierta influencia de los factores meteorológicos en función del período, el siguiente paso fue cuantificar el verdadero impacto de estas en el riesgo de contagio. Por lo tanto, para cuantificar la evolución de este efecto para cada una de las comunidades en cada uno de los períodos se aplicaron los modelos DLMN descritos previamente (Sección 5.2.2.3.). Posteriormente, haciendo uso de las técnicas de meta-análisis se combinaron los resultados de los modelos de las diferentes comunidades para obtener una curva de asociación general que representa la evolución del riesgo relativo de contagio en función de la evolución del factor meteorológico.

La curva de asociación general (Figura 19) muestra diferencias relevantes entre los dos períodos. Durante P1, se observa que el riesgo de contagio sube a medida que aumenta la temperatura, por lo que el riesgo es menor en temperaturas más bajas. Sin embargo, en el caso de P2, se observa un patrón totalmente opuesto, ya que el riesgo de contagio es menor conforme aumenta la temperatura, siendo el riesgo de contagio más elevado con las temperaturas más bajas. Por ejemplo, en P2, a temperaturas más elevadas, en torno a los 28ºC de temperatura media diaria, la probabilidad de contagio es 1,61 veces menor (RR = 0,62, IC = [0,59,0,66]) respecto a la temperatura media global usada como referencia, es decir, 17.5ºC. Por lo lado, en las temperaturas más bajas, entorno a los 5ºC, la probabilidad de contagio es 1,75 veces mayor (RR = 1,75, IC = [1,63,1,87]) con respecto a la temperatura de referencia. Además, en las curvas de asociación individuales de cada una de las comunidades españolas, la misma tendencia es observa en ambos períodos.

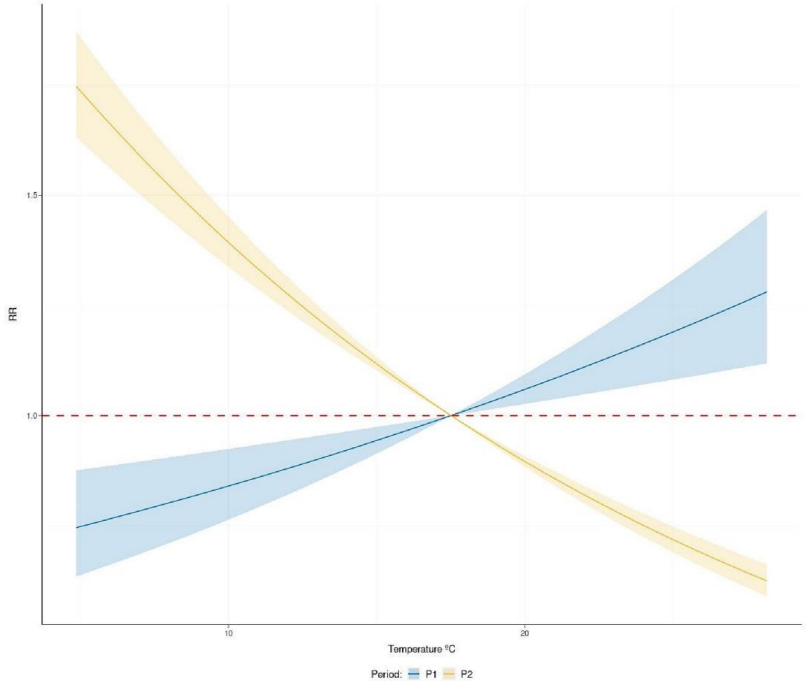


Figura 19: **Curva de asociación global para la temperatura en España**. Representación gráfica de la evolución del RR de infección por COVID-19 en función de la temperatura obtenida por el meta-análisis de los modelos DLNM de las diferentes regiones de España. Se utiliza como referencia la temperatura media global (17,5 °C). El color azul representa el periodo sin vacunación (P1) y el amarillo el periodo con vacunación (P2). Figura extraída del artículo *Villatoro-García et al. 2023*[196].

Estos mismos análisis fueron repetidos con datos de la humedad específica de cada una de las comunidades, obteniéndose resultados y conclusiones análogos a los de la temperatura, los cuales se pueden observar en el artículo[196] proporcionado en el Anexo 9.4.

Por lo tanto, se puede afirmar que, en el caso de las regiones de España, cuando un porcentaje considerable de población está inmunizada, temperaturas y humedades específicas más elevadas contribuyen a una ligera reducción del riesgo de contagio por COVID-19.

5.3.2.2. Validación de los resultados en otros países y regiones de Italia

Para validar los resultados obtenidos en las comunidades de España para la relación entre los factores meteorológicos y la transmisión del COVID-19, los mismos métodos fueron aplicados a países con latitudes similares a España (descritos en la sección5.2.1.2.) en los mismos períodos.

Cuando observamos los resultados de los modelos GAM, se observa una gran heterogeneidad en los resultados (Figura 20). En el caso de la influencia de la temperatura sobre el R_e controlando el resto de variables, se observa patrones similares a los de España en países como Francia, Portugal, Grecia e Italia. Sin embargo, en otros países como Serbia o Macedonia, se observan patrones opuestos para P2. Estas diferencias en las tendencias pueden deberse a porcentajes de vacunación distintos entre los países. En el caso de los países que presentan patrones similares a España, sus niveles de vacunación son similares también. No obstante, en el resto de países sus niveles de vacunación son inferiores, por lo que esto confirmaría el hecho de que es necesario una alta tasa de inmunidad poblacional para observar los efectos de la temperatura en la transmisión del COVID-19.

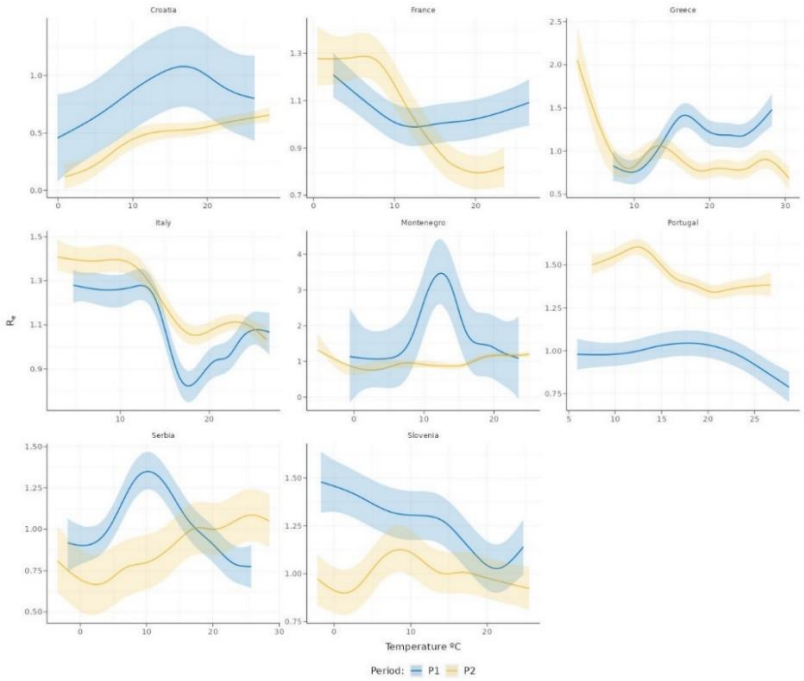


Figura 20: Estimación de la evolución R_e a partir de la influencia de la temperatura (°C) predicha por los modelos GAMs aislando el resto de variables en diferentes países mediterráneos. Representación gráfica de la R_e predicha por los GAMs con la temperatura como variable meteorológica y considerando el resto de variables constantes

para diferentes países mediterráneos. El color azul representa el periodo sin vacunación (P1) y el amarillo el periodo con vacunación (P2). Figura extraída del artículo *Villatoro-García et al. 2023*[196].

Para realizar más validaciones, como Italia muestra globalmente un patrón similar al de España, se aplicaron los mismos procedimientos a las diferentes regiones de Italia. Además, las regiones de Italia tienen la ventaja de que sus diferentes regiones presentan climas similares a las comunidades de España. En lo que se refiere a los modelos GAM, se obtienen resultados semejantes a los obtenidos en las comunidades de España (Figura 21).

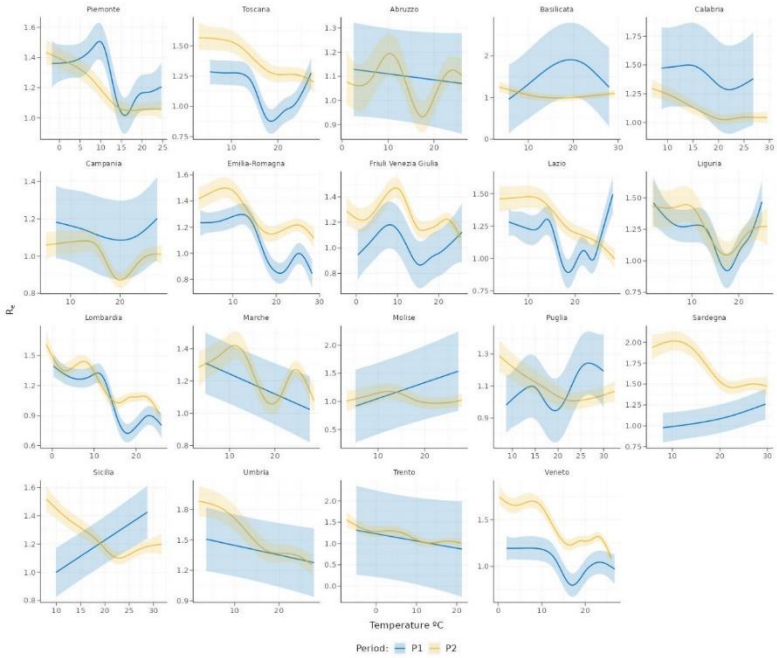


Figura 21: Estimación de la evolución R_e a partir de la influencia de la temperatura (°C) predicha por los modelos GAMs aislando el resto de variables en las diferentes regiones de Italia. Representación gráfica de la R_e predicha por los GAMs con la temperatura como variable meteorológica y considerando el resto de variables constantes para las diferentes regiones de Italia. El color azul representa el periodo sin vacunación (P1) y el amarillo el periodo con vacunación (P2). Figura extraída del artículo *Villatoro-García et al. 2023*[196].

Para la mayoría de regiones de Italia, tanto en P1 como en P2 los modelos son significativos, pero sólo en P2 se observa una reducción de la transmisión con el aumento de la temperatura. Además, en P2 el porcentaje de desviación explicada y el R^2 ajustado de los modelos son mayores que en el caso de P1. Del mismo modo, cuando aplicamos los modelos DLMN y el posterior meta-análisis, en la curva de asociación solo se observa una reducción del riesgo de contagio cuando las temperaturas aumentan durante P2 (Figura 22).

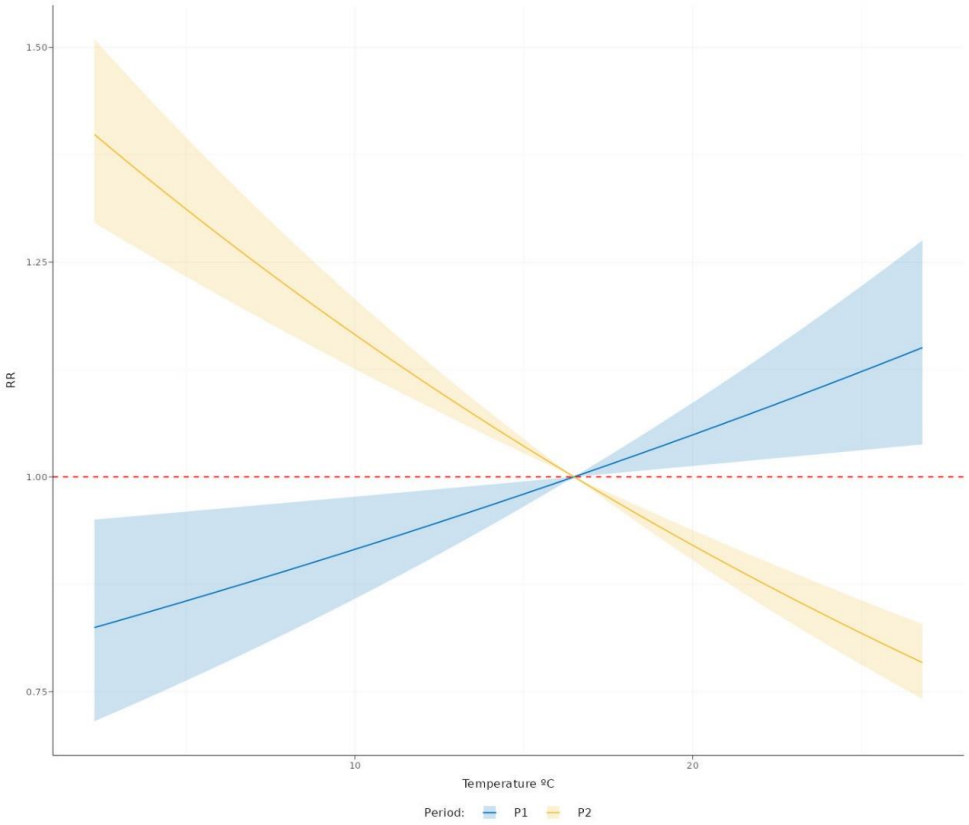


Figura 22: **Curva de asociación global para la temperatura en Italia**. Representación gráfica de la evolución del RR de infección por COVID-19 en función de la temperatura obtenida por el meta-análisis de los modelos DLNM de las diferentes regiones de Italia. Se utiliza como referencia la temperatura media global (17,5 °C). El color azul representa el periodo sin vacunación (P1) y el amarillo el periodo con vacunación (P2).

Por lo tanto, considerando todos los resultados, se puede deducir, que solo existe influencia de la temperatura en la transmisión del virus cuando un porcentaje relevante de población esté vacunada, produciéndose un aumento del contagio en épocas frías y secas.

5.4. Conclusiones

Los resultados de ambos artículos evidencian la importancia de integrar datos procedentes de diferentes fuentes de información para estudiar la evolución de la pandemia y su relación con las diferentes condiciones ambientales.

En lo que respecta a la aplicación de DatAC, hasta donde llega nuestro conocimiento fue la primera aplicación que integró datos epidemiológicos de COVID-19 con información meteorológica y de calidad del aire. Esta innovación la convirtió en una herramienta invaluable para investigaciones posteriores, al ofrecer acceso público a toda la información recopilada. Además, su potencial es significativo para el estudio de la evolución de futuras pandemias, permitiendo comparaciones con la pandemia del COVID-19.

Con esta información, junto con la recopilada de otros países, se analizó la influencia de la temperatura y la humedad en la transmisión del SARS-CoV-2, con el objetivo de esclarecer la controversia existente en la literatura sobre este tema. Basándonos en el

artículo publicado en *Science* por *Barker et al*[193] decidimos considerar la inmunidad poblacional como un factor clave en este análisis. Por ello, se evaluaron dos períodos distintos: uno con baja inmunidad poblacional, correspondiente a los meses posteriores a la etapa inicial pandemia, y el mismo período un año después, cuando la inmunidad poblacional ya se había fortalecido gracias a la vacunación. En ambos períodos se aplicaron modelos específicos que consideraban el tiempo entre contagio y la detección de la enfermedad, así como técnicas de meta-análisis para determinar el efecto global de los factores meteorológicos en la transmisión. Los resultados sugieren que los factores meteorológicos solo tuvieron una influencia leve en la transmisión del virus cuando un alto porcentaje de la población estaba vacunada, con una ligera disminución en la transmisión durante períodos cálidos y un leve aumento en períodos fríos, algo que concuerda con lo predicho en el artículo de *Barker et al*[193]. No obstante, hay que señalizar que nuestro estudio tiene algunas limitaciones. Estas incluyen que el análisis se restringe a personas que recibieron al menos dos dosis de la vacuna COVID-19, excluyendo a quienes adquirieron inmunidad por infección, y no se consideró la duración variable de la inmunidad inducida por la vacuna, lo cual puede afectar la definición de inmunización efectiva. Además, el estudio se centra en países mediterráneos europeos, lo que limita la generalización a otras regiones climáticas, y solo se consideraron la temperatura y la humedad específica, sin incluir otros factores ambientales que podrían influir en la estacionalidad del virus. Por último, la falta de datos completos y actualizados en algunos países limitó la inclusión de ciclos anuales completos en el análisis. A pesar de esto, nuestro estudio presenta un enfoque novedoso en la exploración de la relación entre la estacionalidad de COVID-19 y la inmunidad de la población sugiriendo que la temperatura y la humedad específica tienen un efecto diferencial en la transmisión del virus, y que los efectos observados son consecuencia de la inmunidad de la población. Los conocimientos obtenidos en este estudio proporcionan información valiosa para las estrategias de salud pública y gestión de esta y futuras enfermedades respiratorias.

6. DISCUSIÓN Y CONCLUSIONES FINALES

6.1. Discusión sobre los resultados

En el marco de esta tesis doctoral, se ha evidenciado la relevancia de la integración y análisis de grandes volúmenes de datos en la era del *Big Data*. En esta era, el papel de la Estadística es fundamental, ya que es la encargada de proporcionar los métodos y herramientas necesarios para examinar y analizar esta vasta cantidad de información.

En este escenario, la integración de datos es un proceso fundamental para unificar diferentes fuentes de información y obtener una estructura coherente que permita un análisis más exhaustivo y significativo. La evolución de las técnicas de integración de datos ha permitido manejar este tipo de información heterogéneas y no estructurada de manera más eficiente, facilitando la obtención de resultados más robustos y precisos. El objetivo final de la integración de datos es la obtención de los resultados y conclusiones robustas a partir de la información combinada, de cuyo procedimiento se encargan las técnicas de *Data Fusion* o Fusión de Datos, las cuales tienen como objetivo final la obtención de resultados y conclusiones a partir de la combinación de la información.

En este contexto, una de las técnicas que ha ganado gran popularidad en los últimos años es el meta-análisis, que permite sintetizar los resultados de diversos estudios en un único resultado común. Sin embargo, el uso generalizado de estas técnicas ha llevado en algunos casos a su aplicación incorrecta, lo que ha generado problemas de fiabilidad y reproducibilidad. Estos inconvenientes han sido señalados por expertos en distintos campos de estudio, subrayando la necesidad de contar con especialistas que desarrollen guías metodológicas e implementen las técnicas en software de código abierto, garantizando así su correcta aplicación y la fiabilidad de los resultados obtenidos.

En este contexto se ha desarrollado esta tesis doctoral, en la que se han implementado tanto métodos como herramientas de software con el objetivo de integrar y aplicar técnicas de meta-análisis a datos biomédicos. En particular, el trabajo se ha enfocado en los datos de expresión génica y en datos epidemiológicos relacionados con COVID-19, los cuales han experimentado una gran expansión en los últimos años. Los primeros, debido a los avances en las técnicas de secuenciación génica, y los segundos, como consecuencia de la pandemia de COVID-19 y los esfuerzos por comprender los factores que influían en su evolución.

En relación con los datos de expresión génica, esta tesis doctoral se ha centrado en la aplicación de técnicas de meta-análisis para combinar estudios de este tipo de datos. Se llevó a cabo, en primer lugar, una revisión exhaustiva de las técnicas y del software disponibles para aplicar correctamente el meta-análisis en expresión génica. Esto permitió también establecer un flujo de trabajo adecuado para aplicar los distintos métodos de manera rigurosa.

Posteriormente, con la información adquirida, se identificó un problema recurrente en la integración de este tipo de datos, ya mencionado previamente en la literatura: la posible presencia de genes faltantes. Este problema surge cuando el número de genes en los estudios combinados no coincide, y el enfoque habitual consiste en trabajar únicamente con los genes comunes, lo que puede resultar en una significativa pérdida de información.

Como respuesta, se desarrolló una nueva herramienta denominada DExMA. DExMA es un paquete de R que implementa las funciones necesarias para realizar un meta-análisis de datos de expresión génica, teniendo en cuenta la posible existencia de genes faltantes. Este software aborda el problema desde dos perspectivas: considerando los genes presentes en al menos un número mínimo de estudios y mediante la imputación de genes a partir de aquellos presentes en muestras de otros estudios con valores de expresión similares. La aplicación de DExMA a datos reales demostró que estos enfoques, especialmente el de imputación, minimizan la pérdida de información y generan mejores resultados en comparación con el enfoque tradicional.

Con posterioridad, se exploraron alternativas a los enfoques mencionados para abordar el problema de los genes faltantes. En particular, se investigó la posibilidad de combinar la meta-información de los genes, es decir, integrar la información de las rutas biológicas en lugar de centrarse en la expresión génica. Este enfoque corresponde al campo del meta-análisis de enriquecimiento de rutas. Sin embargo, se identificaron algunas debilidades en las técnicas actuales de este campo, ya que la mayoría utilizaban métodos de combinación de p-valores, lo que resultaba en una pérdida de direccionalidad. Además, muchas de estas técnicas aplicaban el meta-análisis en el contexto de la expresión génica, manteniendo así la influencia del problema de los genes faltantes. Como respuesta, se desarrolló una nueva metodología denominada GSEMA. Esta metodología fusiona técnicas de meta-análisis con enfoques de enriquecimiento de una sola muestra, lo que permite transformar la matriz de expresión génica en una matriz de rutas biológicas y trabajar directamente en el espacio de rutas. Además, los métodos de meta-análisis de GSEMA incorporan correcciones específicas para este tipo de datos, con el objetivo de reducir la tasa de falsos positivos, un problema común en los análisis de enriquecimiento de rutas. La aplicación de GSEMA a datos reales con un número considerable de genes faltantes demostró su capacidad para combinar de manera efectiva esta información, identificando rutas biológicas significativas que eran más relevantes para las condiciones estudiadas en comparación con otros métodos. Además, en su aplicación a datos simulados, GSEMA mostró un mejor control de la tasa de falsos positivos, produciendo resultados más consistentes.

En cuanto a los datos epidemiológicos, la investigación doctoral se centró en la integración de información relacionada con la enfermedad por COVID-19, con el objetivo de extraer conclusiones sobre su posible estacionalidad, es decir, la influencia de factores meteorológicos, como la temperatura, en la evolución de la transmisión del virus. Este aspecto había sido objeto de estudio a lo largo de la pandemia, pero con resultados contradictorios. Para abordar esta cuestión, se realizó una recopilación exhaustiva de datos provenientes de diversas fuentes y bases de datos, tanto de España como de otros países. Específicamente, se recopilaron datos sobre la evolución de la pandemia (número de casos, fallecimientos, tasas de vacunación, etc.), datos ambientales (temperatura, humedad, contaminantes, entre otros) y otros datos relevantes, como las variantes del virus y las medidas gubernamentales adoptadas. A nivel de comunidades autónomas y provincias en España, toda esta información se integró en una aplicación denominada DatAC. Esta herramienta ofrece a los investigadores el acceso público a esta gran cantidad de datos, permitiendo su descarga, visualización en mapas y gráficos, y la aplicación de modelos para obtener una visión preliminar de la evolución de las distintas variables.

Finalmente, con la información recopilada tanto de España como de otros países, se analizó la influencia de los factores meteorológicos en la transmisión del SARS-CoV-2, considerando la inmunidad poblacional como un factor clave. Se estudió la relación en dos períodos distintos: uno en el que la inmunidad poblacional era prácticamente inexistente y otro en el que había alcanzado un nivel alto gracias a la vacunación. Tras aplicar modelos específicos y realizar un meta-análisis para obtener el efecto global de la influencia, se observó que estos factores tenían una leve influencia solamente cuando un alto porcentaje de la población estaba vacunada. De este modo, la transmisión del virus se reducía ligeramente durante períodos cálidos y aumentaba durante los fríos, pero solo cuando existía un porcentaje significativo de inmunidad poblacional.

Estos resultados resaltan la importancia de integrar y reutilizar los datos almacenados en bases de datos públicas, para evitar que se conviertan en una simple acumulación de información y, de este modo, facilitar la obtención de nuevos conocimientos. Además, subrayan la necesidad de emplear y desarrollar técnicas estadísticas rigurosas, considerando los diversos supuestos y limitaciones. En consecuencia, demuestran que es crucial desarrollar metodologías de código abierto que permitan una aplicación adecuada por parte de todos los investigadores, promoviendo así la reproducibilidad y accesibilidad de los resultados a toda la comunidad científica.

6.2. Conclusiones finales

En esta tesis doctoral, se han desarrollado métodos y software para la integración y aplicación de técnicas de meta-análisis a datos biomédicos. Específicamente, las principales conclusiones de esta tesis son:

1. Las técnicas de meta-análisis aplicadas a datos de expresión génica permiten la integración de diferentes conjuntos de datos y la identificación de genes diferencialmente expresados de forma común. Hemos elaborado una guía detallada que describe los pasos necesarios para llevar a cabo de forma apropiada un meta-análisis en datos de expresión. Además, se ha realizado un análisis comparativo de las herramientas disponibles, proporcionando una descripción de sus características y métodos implementados.

2. El paquete de R DExMA implementa los diferentes pasos necesarios para realizar un meta-análisis de expresión génica a partir de datos públicos o datos generados por usuario. DExMA proporciona funcionaliades para tratar el problema de los genes faltantes, reduciendo significativamente la pérdida de información.

3. La metodología GSEMA ofrece un enfoque innovador para aplicar el meta-análisis de rutas biológicas, basado en técnicas de enriquecimiento de una sola muestra. Esto permite obtener un tamaño de efecto individual para cada estudio. Las simulaciones realizadas con GSEMA muestran resultados más consistentes y con una tasa menor de falsos positivos en comparación con otras técnicas. Además, su aplicación a datos reales de diferentes plataformas de secuenciación con genes faltantes conserva mejor la información biológica y produce resultados relevantes para las condiciones estudiadas.

4. La información recopilada en DatAC la convierte en un recurso único para el análisis de la influencia de factores ambientales en la incidencia de la COVID-19 a nivel nacional. Los análisis llevados a cabo con los datos recopilados proporcionan evidencias de que la asociación entre los factores ambientales y la transmisión del COVID-19 depende de la tasa de inmunidad poblacional, observando asociación negativa entre temperatura y humedad con la transmisión de la enfermedad cuando la población presenta una alta tasa de inmunidad.

7. TRABAJO FUTURO

En el contexto de los diferentes temas abordados a lo largo de esta tesis doctoral, se pueden deducir algunas nuevas líneas de trabajo para el futuro:

- En particular, la metodología GSEMA podría aplicarse en la integración de datos multiómicos. El meta-análisis de enriquecimiento ya ha demostrado su utilidad al combinar diferentes tipos de datos ómicos. Los resultados positivos de GSEMA en comparación con métodos previos sugieren que podría ser extremadamente útil para combinar estudios de diferentes ómicas bajo condiciones similares, preservando la direccionalidad de la regulación.
- Las técnicas de meta-análisis de datos de expresión génica en esta tesis doctoral se han aplicado a datos de *microarrays* y *bulk* RNA-Seq. Sin embargo, estas podrían extenderse al contexto de los datos de scRNA-Seq. Especialmente, estas podrían aplicarse a la hora de combinar datos de *pseduobulk* scRNA-Seq para combinar datos de expresión de los mismos tipos celulares.
- En lo que respecta a los análisis de la estacionalidad del COVID-19, desde el año 2022 los diferentes gobiernos ya no proporcionan datos diarios de la evolución de la pandemia, por lo que consideramos que todo el trabajo realizado ya es de suficiente relevancia en el campo y que resulta complicado obtener nuevos resultados que sean innovadores dentro del campo.

8. BIBLIOGRAFÍA

1. Data growth worldwide 2010-2025. Statista. Accessed June 3, 2024. https://www.statista.com/statistics/871513/worldwide-data-created/

2. Statista - El portal de estadísticas. Statista. Accessed June 3, 2024. https://es.statista.com/cuentas/

3. Lohr S. The Origins of "Big Data": An Etymological Detective Story. Bits Blog. 1359727843. Accessed July 16, 2024. https://archive.nytimes.com/bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story/

4. McDermott JE, Wang J, Mitchell H, et al. Challenges in biomarker discovery: combining expert insights with statistical analysis of complex omics data. *Expert Opinion on Medical Diagnostics*. 2013;7(1):37-51. doi:10.1517/17530059.2012.718329

5. Collins SH, Price D, Johnson LL. The Role of FDA CDER Statisticians in Response Efforts to the COVID-19 Pandemic. *Statistics in Biopharmaceutical Research*. 2022;14(1):3-4. doi:10.1080/19466315.2020.1841024

6. Srivastava K, Dehwal A. Data Integration Challenges and Solutions: A Study. In: ; 2012. Accessed July 17, 2024. https://www.semanticscholar.org/paper/Data-Integration-Challenges-and-Solutions%3A-A-Study-Srivastava-Dehwal/0b68afdc4d18c2d90aab32c2a7a5560e46c4a23c

7. Lenzerini M. Data integration: a theoretical perspective. In: *Proceedings of the Twenty-First ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. PODS '02. Association for Computing Machinery; 2002:233-246. doi:10.1145/543613.543644

8. Vyas S, Vaishnav P. A comparative study of various ETL process and their testing techniques in data warehouse. *Journal of Statistics and Management Systems*. 2017;20(4):753-763. doi:10.1080/09720510.2017.1395194

9. Maass W, Storey VC. Pairing conceptual modeling with machine learning. *Data & Knowledge Engineering*. 2021;134:101909. doi:10.1016/j.datak.2021.101909

10. Hui J, Li L, Zhang Z. Integration of Big Data: A Survey. In: Zhou Q, Gan Y, Jing W, Song X, Wang Y, Lu Z, eds. *Data Science*. Springer; 2018:101-121. doi:10.1007/978-981-13-2203-7_9

11. Bleiholder J, Naumann F. Data fusion. *ACM Comput Surv*. 2009;41(1):1-41. doi:10.1145/1456650.1456651

12. Schmitt M, Zhu XX. Data Fusion and Remote Sensing: An ever-growing relationship. *IEEE Geoscience and Remote Sensing Magazine*. 2016;4(4):6-23. doi:10.1109/MGRS.2016.2561021

13. Hassani S, Dackermann U, Mousavi M, Li J. A systematic review of data fusion techniques for optimized structural health monitoring. *Information Fusion*. 2024;103:102136. doi:10.1016/j.inffus.2023.102136

14. Dasarathy BV. Sensor fusion potential exploitation-innovative architectures and illustrative applications. *Proceedings of the IEEE*. 1997;85(1):24-38. doi:10.1109/5.554206

15. Alofi A, Alghamdi A, Alahmadi R, Aljuaid N, M. H. A Review of Data Fusion Techniques. *IJCA*. 2017;167(7):37-41. doi:10.5120/ijca2017914318

16. Castanedo F. A Review of Data Fusion Techniques. *The Scientific World Journal*. 2013;2013(1):704504. doi:10.1155/2013/704504

17. Fisher RA. Statistical Methods for Research Workers. In: Kotz S, Johnson NL, eds. *Breakthroughs in Statistics: Methodology and Distribution*. Springer; 1992:66-70. doi:10.1007/978-1-4612-4380-9_6

18. Pearson ES. The Probability Integral Transformation for Testing Goodness of Fit and Combining Independent Tests of Significance. *Biometrika*. 1938;30(1/2):134-148. doi:10.2307/2332229

19. Fisher RA. *Statistical Method For Research Workers*.; 1934. Accessed June 12, 2024. http://archive.org/details/in.ernet.dli.2015.205971

20. Glass GV. Primary, Secondary, and Meta-Analysis of Research. *Educational Researcher*. 1976;5(10):3-8. doi:10.2307/1174772

21. PubMed. PubMed. Accessed July 16, 2024. https://pubmed.ncbi.nlm.nih.gov/

22. Higgins J, Thomas J, Chandler J, et al. *Cochrane Handbook for Systematic Reviews of Interventions*. 2nd ed. John Wiley & Sons; 2019.

23. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med*. 2009;6(7):e1000097. doi:10.1371/journal.pmed.1000097

24. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. *Introduction to Meta-Analysis*. John Wiley & Sons; 2021.

25. Heard NA, Rubin-Delanchy P. Choosing between methods of combining p-values. *Biometrika*. 2018;105(1):239-246. doi:10.1093/biomet/asx076

26. Hong F, Breitling R, McEntee CW, Wittner BS, Nemhauser JL, Chory J. RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*. 2006;22(22):2825-2827. doi:10.1093/bioinformatics/btl476

27. Ioannidis JPA. The Mass Production of Redundant, Misleading, and Conflicted Systematic Reviews and Meta-analyses. *Milbank Q*. 2016;94(3):485-514. doi:10.1111/1468-0009.12210

28. Ioannidis JPA, Chang CQ, Lam TK, Schully SD, Khoury MJ. The Geometric Increase in Meta-Analyses from China in the Genomic Era. *PLOS ONE*. 2013;8(6):e65602. doi:10.1371/journal.pone.0065602

29. Park JH, Eisenhut M, van der Vliet HJ, Shin JI. Statistical controversies in clinical research: overlap and errors in the meta-analyses of microRNA genetic association studies in cancers. *Ann Oncol*. 2017;28(6):1169-1182. doi:10.1093/annonc/mdx024

30. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001;409(6822):860-921. doi:10.1038/35057062

31. Lewin HA, Robinson GE, Kress WJ, et al. Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences*. 2018;115(17):4325-4333. doi:10.1073/pnas.1720115115

32. Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). Accessed July 25, 2024. www.genome.gov/sequencingcostsdata

33. Manzoni C, Kia DA, Vandrovcova J, et al. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Briefings in Bioinformatics*. 2018;19(2):286-302. doi:10.1093/bib/bbw114

34. Parkinson H, Kapushesky M, Shojatalab M, et al. ArrayExpress—a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res*. 2007;35(Database issue):D747-D750. doi:10.1093/nar/gkl995

35. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 2018;46(D1):D8-D13. doi:10.1093/nar/gkx1095

36. Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res*. 2013;41(Database issue):D991-995. doi:10.1093/nar/gks1193

37. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002;30(1):207-210. doi:10.1093/nar/30.1.207

38. Dermitzakis ET. From gene expression to disease risk. *Nat Genet*. 2008;40(5):492-493. doi:10.1038/ng0508-492

39. McDermaid A, Monier B, Zhao J, Liu B, Ma Q. Interpretation of differential gene expression results of RNA-seq data: review and integration. *Brief Bioinform*. 2019;20(6):2044-2054. doi:10.1093/bib/bby067

40. Rosati D, Palmieri M, Brunelli G, et al. Differential gene expression analysis pipelines and bioinformatic tools for the identification of specific biomarkers: A review. *Comput Struct Biotechnol J*. 2024;23:1154-1168. doi:10.1016/j.csbj.2024.02.018

41. HOMMEL G. A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*. 1988;75(2):383-386. doi:10.1093/biomet/75.2.383

42. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1995;57(1):289-300. doi:10.1111/j.2517-6161.1995.tb02031.x

43. Liu S, Wang Z, Zhu R, Wang F, Cheng Y, Liu Y. Three Differential Expression Analysis Methods for RNA Sequencing: limma, EdgeR, DESeq2. *J Vis Exp*. 2021;(175). doi:10.3791/62528

44. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8(1):118-127. doi:10.1093/biostatistics/kxj037

45. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7):e47. doi:10.1093/nar/gkv007

46. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biology*. 2010;11(10):R106. doi:10.1186/gb-2010-11-10-r106

47. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26(1):139-140. doi:10.1093/bioinformatics/btp616

48. Sun S, Hood M, Scott L, et al. Differential expression analysis for RNAseq using Poisson mixed models. *Nucleic Acids Res*. 2017;45(11):e106. doi:10.1093/nar/gkx204

49. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*. 2014;15(12):550. doi:10.1186/s13059-014-0550-8

50. Chen Y, Chen L, Lun ATL, Baldoni PL, Smyth GK. edgeR 4.0: powerful differential analysis of sequencing data with expanded functionality and improved support for small counts and larger datasets. Published online January 24, 2024:2024.01.21.576131. doi:10.1101/2024.01.21.576131

51. Law CW, Chen Y, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*. 2014;15(2):R29. doi:10.1186/gb-2014-15-2-r29

52. Tarazona S, García F, Ferrer A, Dopazo J, Conesa A. NOIseq: a RNA-seq differential expression method robust for sequencing depth biases. *EMBnet.journal*. 2011;17(B):18-19. doi:10.14806/ej.17.B.265

53. Tarazona S, Furió-Tarí P, Turrà D, et al. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Res*. 2015;43(21):e140. doi:10.1093/nar/gkv711

54. Li J, Tibshirani R. Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. *Stat Methods Med Res*. 2013;22(5):519-536. doi:10.1177/0962280211428386

55. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*. 2013;14(1):91. doi:10.1186/1471-2105-14-91

56. Toro-Domínguez D, Villatoro-García JA, Martorell-Marugán J, Román-Montoya Y, Alarcón-Riquelme ME, Carmona-Sáez P. A survey of gene expression meta-analysis: methods and applications. *Brief Bioinform*. 2021;22(2):1694-1705. doi:10.1093/bib/bbaa019

57. Bell R, Barraclough R, Vasieva O. Gene Expression Meta-Analysis of Potential Metastatic Breast Cancer Markers. *Curr Mol Med*. 2017;17(3):200-210. doi:10.2174/1566524017666170807144946

58. O'Mara TA, Zhao M, Spurdle AB. Meta-analysis of gene expression studies in endometrial cancer identifies gene expression profiles associated with aggressive disease and patient outcome. *Scientific Reports*. 2016;6:36677. doi:10.1038/srep36677

59. Afroz S, Giddaluru J, Vishwakarma S, Naz S, Khan AA, Khan N. A Comprehensive Gene Expression Meta-analysis Identifies Novel Immune Signatures in Rheumatoid Arthritis Patients. *Front Immunol*. 2017;8:74. doi:10.3389/fimmu.2017.00074

60. Song GG, Kim JH, Seo YH, Choi SJ, Ji JD, Lee YH. Meta-analysis of differentially expressed genes in primary Sjogren's syndrome by using microarray. *Human Immunology*. 2014;75(1):98-104. doi:10.1016/j.humimm.2013.09.012

61. Patel H, Dobson RJB, Newhouse SJ. A Meta-Analysis of Alzheimer's Disease Brain Transcriptomic Data. *J Alzheimers Dis*. 2019;68(4):1635-1656. doi:10.3233/JAD-181085

62. Badr MT, Häcker G. Gene expression profiling meta-analysis reveals novel gene signatures and pathways shared between tuberculosis and rheumatoid arthritis. *PLoS ONE*. 2019;14(3):e0213470. doi:10.1371/journal.pone.0213470

63. Toro-Domínguez D, Carmona-Sáez P, Alarcón-Riquelme ME. Shared signatures between rheumatoid arthritis, systemic lupus erythematosus and Sjögren's syndrome uncovered through gene expression meta-analysis. *Arthritis Res Ther*. 2014;16(6). doi:10.1186/s13075-014-0489-x

64. Tuller T, Atar S, Ruppin E, Gurevich M, Achiron A. Common and specific signatures of gene expression and protein-protein interactions in autoimmune diseases. *Genes Immun*. 2013;14(2):67-82. doi:10.1038/gene.2012.55

65. Kelly J, Moyeed R, Carroll C, Albani D, Li X. Gene expression meta-analysis of Parkinson's disease and its relationship with Alzheimer's disease. *Molecular Brain*. 2019;12(1):16. doi:10.1186/s13041-019-0436-5

66. Ibáñez K, Boullosa C, Tabarés-Seisdedos R, Baudot A, Valencia A. Molecular Evidence for the Inverse Comorbidity between Central Nervous System Disorders and Cancers Detected by Transcriptomic Meta-analyses. *PLOS Genetics*. 2014;10(2):e1004173. doi:10.1371/journal.pgen.1004173

67. Toro-Domínguez D, Carmona-Sáez P, Alarcón-Riquelme ME. Support for phosphoinositol 3 kinase and mTOR inhibitors as treatment for lupus using in-silico drug-repurposing analysis. *Arthritis Res Ther*. 2017;19(1):54. doi:10.1186/s13075-017-1263-7

68. Lamb J, Crawford ED, Peck D, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*. 2006;313(5795):1929-1935. doi:10.1126/science.1132939

69. Bobak CA, McDonnell L, Nemesure MD, Lin J, Hill JE. Assessment of Imputation Methods for Missing Gene Expression Data in Meta-Analysis of Distinct Cohorts of Tuberculosis Patients. *Pac Symp Biocomput*. 2020;25:307-318.

70. Mancuso CA, Canfield JL, Singla D, Krishnan A. A flexible, interpretable, and accurate approach for imputing the expression of unmeasured genes. *Nucleic Acids Res*. doi:10.1093/nar/gkaa881

71. Chiu CC, Chan SY, Wang CC, Wu WS. Missing value imputation for microarray data: a comprehensive comparison study and a web tool. *BMC Syst Biol*. 2013;7(Suppl 6):S12. doi:10.1186/1752-0509-7-S6-S12

72. Stathias V, Turner J, Koleti A, et al. LINCS Data Portal 2.0: next generation access point for perturbation-response signatures. *Nucleic Acids Res*. 2020;48(D1):D431-D439. doi:10.1093/nar/gkz1023

73. Villatoro-García JA, Martorell-Marugán J, Toro-Domínguez D, Román-Montoya Y, Femia P, Carmona-Sáez P. DExMA: An R Package for Performing Gene Expression Meta-Analysis with Missing Genes. *Mathematics*. 2022;10(18):3376. doi:10.3390/math10183376

74. Hedges LV. Fitting Categorical Models to Effect Sizes from a Series of Experiments. *Journal of Educational Statistics*. 1982;7(2):119-137. doi:10.2307/1164961

75. Hedges LV. Distribution Theory for Glass's Estimator of Effect Size and Related Estimators. *Journal of Educational Statistics*. 1981;6(2):107-128. doi:10.2307/1164588

76. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Routledge; 1988. doi:10.4324/9780203771587

77. Lin L, Aloe AM. Evaluation of various estimators for standardized mean difference in meta-analysis. *Stat Med*. 2021;40(2):403-426. doi:10.1002/sim.8781

78. Doncaster CP, Spake R. Correction for bias in meta-analysis of little-replicated studies. *Methods in Ecology and Evolution*. 2018;9(3):634-644. doi:10.1111/2041-210X.12927

79. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*. 2004;3:Article3. doi:10.2202/1544-6115.1027

80. Smyth G, Hu Y, Ritchie M, et al. limma: Linear Models for Microarray Data. Published online 2021. doi:10.18129/B9.bioc.limma

81. Rosenthal R, Rosnow RL. *Essentials of Behavioral Research: Methods and Data Analysis*. Third Edition. New York: McGraw-Hill; 2008.

82. Marot G, Foulley JL, Mayer CD, Jaffrézic F. Moderated effect size and P-value combinations for microarray meta-analyses. *Bioinformatics*. 2009;25(20):2692-2699. doi:10.1093/bioinformatics/btp444

83. Law CW, Alhamdoosh M, Su S, et al. RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. *F1000Res*. 2018;5:ISCB Comm J-1408. doi:10.12688/f1000research.9005.3

84. Liu R, Holik AZ, Su S, et al. Why weight? Modelling sample and observational level variability improves power in RNA-seq analyses. *Nucleic Acids Res*. 2015;43(15):e97. doi:10.1093/nar/gkv412

85. DerSimonian R, Kacker R. Random-effects model for meta-analysis of clinical trials: an update. *Contemp Clin Trials*. 2007;28(2):105-114. doi:10.1016/j.cct.2006.04.004

86. Berkey CS, Hoaglin DC, Mosteller F, Colditz GA. A random-effects regression model for meta-analysis. *Statistics in Medicine*. 1995;14(4):395-411. doi:10.1002/sim.4780140406

87. Raudenbush SW. Analyzing effect sizes: Random-effects models. In: *The Handbook of Research Synthesis and Meta-Analysis, 2nd Ed*. Russell Sage Foundation; 2009:295-315.

88. Demerath NJ. The American Soldier: Volume I, Adjustment During Army Life. By S. A. Stouffer, E. A. Suchman, L. C. DeVinney, S. A. Star, R. M. Williams, Jr. Volume II, Combat and Its Aftermath. By S. A. Stouffer, A. A. Lumsdaine, M. H. Lumsdaine, R. M. Williams, Jr., M. B. Smith, I. L. Janis, S. A. Star, L. S. Cottrell, Jr. Princeton, New Jersey: Princeton University Press, 1949. Vol. I, 599 pp., Vol. II, 675 pp. $7.50 each; $13.50 together. *Social Forces*. 1949;28(1):87-90. doi:10.2307/2572105

89. Pearson K. On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. *Biometrika*. 1933;25:379-410. doi:10.1093/biomet/25.3-4.379

90. Zaykin DV. Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. *J Evol Biol*. 2011;24(8):1836-1841. doi:10.1111/j.1420-9101.2011.02297.x

91. Whitlock MC. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *J Evol Biol*. 2005;18(5):1368-1373. doi:10.1111/j.1420-9101.2005.00917.x

92. L.h.c Tippett. *The Methods Of Statistics*.; 1931. Accessed June 13, 2024. http://archive.org/details/in.ernet.dli.2015.189563

93. Song C, Tseng GC. HYPOTHESIS SETTING AND ORDER STATISTIC FOR ROBUST GENOMIC META-ANALYSIS. *Ann Appl Stat*. 2014;8(2):777-800.

94. Li J, Tseng GC. An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *Ann Appl Stat*. 2011;5(2A):994-1019. doi:10.1214/10-AOAS393

95. Wilkinson B. A statistical consideration in psychological research. *Psychological Bulletin*. 1951;48(2):156-158. doi:10.1037/h0059111

96. Liu Y, Chen S, Li Z, Morrison AC, Boerwinkle E, Lin X. ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. *Am J Hum Genet*. 2019;104(3):410-421. doi:10.1016/j.ajhg.2019.01.002

97. Liu Y, Xie J. Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *J Am Stat Assoc*. 2020;115(529):393-402. doi:10.1080/01621459.2018.1554485

98. Eisinga R, Breitling R, Heskes T. The exact probability distribution of the rank product statistics for replicated experiments. *FEBS Letters*. 2013;587(6):677-682. doi:10.1016/j.febslet.2013.01.037

99. Heskes T, Eisinga R, Breitling R. A fast algorithm for determining bounds and accurate approximate p-values of the rank product statistic for replicate experiments. *BMC Bioinformatics*. 2014;15(1):367. doi:10.1186/s12859-014-0367-1

100. Koziol JA. A cautionary note on the rank product statistic. *FEBS Letters*. 2016;590(11):1586-1591. doi:10.1002/1873-3468.12194

101. Chang LC, Lin HM, Sibille E, Tseng GC. Meta-analysis methods for combining multiple expression profiles: comparisons, statistical characterization and an application guideline. *BMC Bioinformatics*. 2013;14(1):368. doi:10.1186/1471-2105-14-368

102. Dreyfuss JM, Johnson MD, Park PJ. Meta-analysis of glioblastoma multiforme versus anaplastic astrocytoma identifies robust gene markers. *Molecular Cancer*. 2009;8(1):71. doi:10.1186/1476-4598-8-71

103. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med*. 2002;21(11):1539-1558. doi:10.1002/sim.1186

104. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ*. 2003;327(7414):557-560. doi:10.1136/bmj.327.7414.557

105. Waldron L, Riester M. Meta-Analysis in Gene Expression Studies. *Methods Mol Biol*. 2016;1418:161-176. doi:10.1007/978-1-4939-3578-9_8

106. Jaksik R, Iwanaszko M, Rzeszowska-Wolny J, Kimmel M. Microarray experiments and factors which affect their reliability. *Biol Direct*. 2015;10:46. doi:10.1186/s13062-015-0077-2

107. Ioannidis JPA. Why most published research findings are false. *PLoS Med*. 2005;2(8):e124. doi:10.1371/journal.pmed.0020124

108. Albert I. *The Biostar Handbook: 2nd Edition*. Accessed August 22, 2024. https://www.biostarhandbook.com/

109. Wu Z. A review of statistical methods for preprocessing oligonucleotide microarrays. *Stat Methods Med Res*. 2009;18(6):533-541. doi:10.1177/0962280209351924

110. Tarca AL, Romero R, Draghici S. Analysis of microarray experiments of gene expression profiling. *Am J Obstet Gynecol*. 2006;195(2):373-388. doi:10.1016/j.ajog.2006.07.001

111. Conesa A, Madrigal P, Tarazona S, et al. A survey of best practices for RNA-seq data analysis. *Genome Biology*. 2016;17(1):13. doi:10.1186/s13059-016-0881-8

112. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*. 2011;39(Database issue):D52-57. doi:10.1093/nar/gkq1237

113. Harrison PW, Amode MR, Austine-Orimoloye O, et al. Ensembl 2024. *Nucleic Acids Res*. 2023;52(D1):D891-D899. doi:10.1093/nar/gkad1049

114. Seal RL, Braschi B, Gray K, et al. Genenames.org: the HGNC resources in 2023. *Nucleic Acids Research*. 2023;51(D1):D1003-D1009. doi:10.1093/nar/gkac888

115. Miller JA, Cai C, Langfelder P, et al. Strategies for aggregating gene expression data: The collapseRows R function. *BMC Bioinformatics*. 2011;12(1):322. doi:10.1186/1471-2105-12-322

116. Kwak SK, Kim JH. Statistical data preparation: management of missing values and outliers. *Korean J Anesthesiol*. 2017;70(4):407-411. doi:10.4097/kjae.2017.70.4.407

117. de Souto MC, Jaskowiak PA, Costa IG. Impact of missing data imputation methods on gene expression clustering and classification. *BMC Bioinformatics*. 2015;16(1):64. doi:10.1186/s12859-015-0494-3

118. Liew AWC, Law NF, Yan H. Missing value imputation for gene expression data: computational techniques to recover missing data from available information. *Brief Bioinformatics*. 2011;12(5):498-513. doi:10.1093/bib/bbq080

119. Aittokallio T. Dealing with missing values in large-scale studies: microarray data imputation and beyond. *Brief Bioinformatics*. 2010;11(2):253-264. doi:10.1093/bib/bbp059

120. Filzmoser P, Maronna R, Werner M. Outlier identification in high dimensions. *Computational Statistics & Data Analysis*. 2008;52(3):1694-1711. doi:10.1016/j.csda.2007.05.018

121. Hadi AS. Identifying Multiple Outliers in Multivariate Data. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1992;54(3):761-771. doi:10.1111/j.2517-6161.1992.tb01449.x

122. Shieh AD, Hung YS. Detecting outlier samples in microarray data. *Stat Appl Genet Mol Biol*. 2009;8:Article 13. doi:10.2202/1544-6115.1426

123. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*. 2012;28(6):882-883. doi:10.1093/bioinformatics/bts034

124. Zhou G, Soufan O, Ewald J, Hancock REW, Basu N, Xia J. NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. *Nucleic Acids Research*. 2019;47(W1):W234-W241. doi:10.1093/nar/gkz240

125. Toro-Domínguez D, Martorell-Marugán J, López-Domínguez R, et al. ImaGEO: integrative gene expression meta-analysis from GEO database. *Bioinformatics*. 2019;35(5):880-882. doi:10.1093/bioinformatics/bty721

126. Zoubarev A, Hamer KM, Keshav KD, et al. Gemma: a resource for the reuse, sharing and meta-analysis of expression profiling data. *Bioinformatics*. 2012;28(17):2272-2273. doi:10.1093/bioinformatics/bts430

127. Sharov AA, Schlessinger D, Ko MSH. ExAtlas: An interactive online tool for meta-analysis of gene expression data. *J Bioinform Comput Biol*. 2015;13(6):1550019. doi:10.1142/S0219720015500195

128. Shashirekha HL, Wani AH. ShinyMDE: Shiny tool for microarray meta-analysis for differentially expressed gene detection. In: *2016 International Conference on Bioinformatics and Systems Biology (BSB)*; 2016:1-5. doi:10.1109/BSB.2016.7552152

129. Ma T, Huo Z, Kuo A, et al. MetaOmics: analysis pipeline and browser-based software suite for transcriptomic meta-analysis. *Bioinformatics*. 2019;35(9):1597-1599. doi:10.1093/bioinformatics/bty825

130. Haynes WA, Vallania F, Liu C, et al. Empowering Multi-Cohort Gene Expression Analysis to Increase Reproducibility. *Pac Symp Biocomput*. 2016;22:144-153.

131. Dewey M. metap: meta-analysis of significance values. Published online 2024. R package version 1.11 https://cran.r-project.org/web/packages/metap/metap.pdf

132. Lara Lusa, Gentleman R, Ruschhaupt M. GeneMeta: MetaAnalysis for High Throughput Experiments. Published online 2024. R package version 1.76.0. doi:10.18129/B9.bioc.GeneMeta

133. Rau A, Marot G, Jaffrézic F. Differential meta-analysis of RNA-seq data from multiple studies. *BMC Bioinformatics*. 2014;15(1):91. doi:10.1186/1471-2105-15-91

134. Del Carratore F, Jankevics A, Eisinga R, Heskes T, Hong F, Breitling R. RankProd 2.0: a refactored bioconductor package for detecting differentially expressed features in molecular profiling datasets. *Bioinformatics*. 2017;33(17):2774-2775. doi:10.1093/bioinformatics/btx292

135. Pihur V, Datta S, Datta S. RankAggreg, an R package for weighted rank aggregation. *BMC Bioinformatics*. 2009;10:62. doi:10.1186/1471-2105-10-62

136. Lottaz C, Yang X, Scheid S, Spang R. OrderedList--a bioconductor package for detecting similarity in ordered gene lists. *Bioinformatics*. 2006;22(18):2315-2316. doi:10.1093/bioinformatics/btl385

137. Stevens JR, Nicholas G. metahdep: Hierarchical Dependence in Meta-Analysis. Published online 2024. R package version 1.62.0. doi:10.18129/B9.bioc.metahdep

138. Tsuyuzaki K, Nikaido I. metaSeq: Meta-analysis of RNA-Seq count data in multiple studies. Published online 2024. R package version 1.44.0. doi:10.18129/B9.bioc.metaSeq

139. Prada C, Lima D, Nakaya H. MetaVolcanoR: Gene Expression Meta-analysis Visualization Tool. Published online 2022. doi:10.18129/B9.bioc.MetaVolcanoR

140. Pickering A. crossmeta: Cross Platform Meta-Analysis of Microarray Data. Published online 2022. doi:10.18129/B9.bioc.crossmeta

141. Villatoro-García JA, Carmona-Saez P. Differential Expression Meta Analysis with DExMA package. Published online 2024. https://www.bioconductor.org/packages/release/bioc/html/DExMA.html

142. Clough E, Barrett T, Wilhite SE, et al. NCBI GEO: archive for gene expression and epigenomics data sets: 23-year update. *Nucleic Acids Research*. 2024;52(D1):D138-D144. doi:10.1093/nar/gkad965

143. Martorell-Marugán J, López-Domínguez R, García-Moreno A, et al. A comprehensive database for integrated analysis of omics data in autoimmune diseases. *BMC Bioinformatics*. 2021;22(1):343. doi:10.1186/s12859-021-04268-4

144. Li QZ, Karp DR, Quan J, et al. Risk factors for ANA positivity in healthy persons. *Arthritis Res Ther*. 2011;13(2):R38. doi:10.1186/ar3271

145. Kennedy WP, Maciuca R, Wolslegel K, et al. Association of the interferon signature metric with serological disease manifestations but not global activity scores in multiple cohorts of patients with SLE. *Lupus Sci Med*. 2015;2(1):e000080. doi:10.1136/lupus-2014-000080

146. Zhu H, Mi W, Luo H, et al. Whole-genome transcription and DNA methylation analysis of peripheral blood mononuclear cells identified aberrant gene regulation pathways in systemic lupus erythematosus. *Arthritis Res Ther*. 2016;18:162. doi:10.1186/s13075-016-1050-x

147. Carmona-Saez P, Chagoyen M, Tirado F, Carazo JM, Pascual-Montano A. GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. *Genome Biology*. 2007;8(1):R3. doi:10.1186/gb-2007-8-1-r3

148. Garcia-Moreno A, López-Domínguez R, Villatoro-García JA, et al. Functional Enrichment Analysis of Regulatory Elements. *Biomedicines*. 2022;10(3):590. doi:10.3390/biomedicines10030590

149. Huang R, Grishagin I, Wang Y, et al. The NCATS BioPlanet – An Integrated Platform for Exploring the Universe of Cellular Signaling Pathways for Toxicology, Systems Biology, and Chemical Genomics. *Frontiers in Pharmacology*. 2019;10:445. doi:10.3389/fphar.2019.00445

150. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res*. 2019;47(D1):D330-D338. doi:10.1093/nar/gky1055

151. Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M. New approach for understanding genome variations in KEGG. *Nucleic Acids Res*. 2019;47(D1):D590-D595. doi:10.1093/nar/gky962

152. Croft D, O'Kelly G, Wu G, et al. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res*. 2011;39(Database issue):D691-D697. doi:10.1093/nar/gkq1018

153. Chen M, Zang M, Wang X, Xiao G. A powerful Bayesian meta-analysis method to integrate multiple gene set enrichment studies. *Bioinformatics*. 2013;29(7):862-869. doi:10.1093/bioinformatics/btt068

154. Shen K, Tseng GC. Meta-analysis for pathway enrichment analysis when combining multiple genomic studies. *Bioinformatics*. 2010;26(10):1316-1323. doi:10.1093/bioinformatics/btq148

155. Lu W, Wang X, Zhan X, Gazdar A. Meta-analysis approaches to combine multiple gene set enrichment studies. *Stat Med*. 2018;37(4):659-672. doi:10.1002/sim.7540

156. Yaari G, Bolen CR, Thakar J, Kleinstein SH. Quantitative set analysis for gene expression: a method to quantify gene set differential expression including gene-gene correlations. *Nucleic Acids Research*. 2013;41(18):e170. doi:10.1093/nar/gkt660

157. Meng H, Yaari G, Bolen CR, Avey S, Kleinstein SH. Gene set meta-analysis with Quantitative Set Analysis for Gene Expression (QuSAGE). *PLOS Computational Biology*. 2019;15(4):e1006899. doi:10.1371/journal.pcbi.1006899

158. Barbie DA, Tamayo P, Boehm JS, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*. 2009;462(7269):108-112. doi:10.1038/nature08460

159. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics*. 2013;14(1):7. doi:10.1186/1471-2105-14-7

160. Lee E, Chuang HY, Kim JW, Ideker T, Lee D. Inferring Pathway Activity toward Precise Disease Classification. *PLOS Computational Biology*. 2008;4(11):e1000217. doi:10.1371/journal.pcbi.1000217

161. Foroutan M, Bhuva DD, Lyu R, Horan K, Cursons J, Davis MJ. Single sample scoring of molecular phenotypes. *BMC Bioinformatics*. 2018;19(1):404. doi:10.1186/s12859-018-2435-4

162. Martínez-Mira C, Conesa A, Tarazona S. MOSim: Multi-Omics Simulation in R. Published online September 20, 2018:421834. doi:10.1101/421834

163. Chaussabel D, Quinn C, Shen J, et al. A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus. *Immunity*. 2008;29(1):150-164. doi:10.1016/j.immuni.2008.05.012

164. Tokuyama M, Kong Y, Song E, Jayewickreme T, Kang I, Iwasaki A. ERVmap analysis reveals genome-wide transcription of human endogenous retroviruses. *Proc Natl Acad Sci U S A*. 2018;115(50):12565-12572. doi:10.1073/pnas.1814589115

165. Wilks C, Zheng SC, Chen FY, et al. recount3: summaries and queries for large-scale RNA-seq expression and splicing. *Genome Biol*. 2021;22(1):323. doi:10.1186/s13059-021-02533-6

166. Kröger W, Mapiye D, Entfellner JBD, Tiffin N. A meta-analysis of public microarray data identifies gene regulatory pathways deregulated in peripheral blood mononuclear cells from individuals with Systemic Lupus Erythematosus compared to those without. *BMC Med Genomics*. 2016;9:66. doi:10.1186/s12920-016-0227-0

167. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst*. 2015;1(6):417-425. doi:10.1016/j.cels.2015.12.004

168. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*. 2011;27(12):1739-1740. doi:10.1093/bioinformatics/btr260

169. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545-15550. doi:10.1073/pnas.0506580102

170. Korotkevich G, Sukhov V, Budin N, Shpak B, Artyomov MN, Sergushichev A. Fast gene set enrichment analysis. Published online February 1, 2021:060012. doi:10.1101/060012

171. Gatti DM, Barry WT, Nobel AB, Rusyn I, Wright FA. Heading Down the Wrong Pathway: on the Influence of Correlation within Gene Sets. *BMC Genomics*. 2010;11(1):574. doi:10.1186/1471-2164-11-574

172. Zheng Y, Lunetta KL, Liu C, et al. An evaluation of the genome-wide false positive rates of common methods for identifying differentially methylated regions using illumina methylation arrays. *Epigenetics*. 2022;17(13):2241-2258. doi:10.1080/15592294.2022.2115600

173. Guan W jie, Ni Z yi, Hu Y, et al. Clinical Characteristics of Coronavirus Disease 2019 in China. *New England Journal of Medicine*. 2020;382(18):1708-1720. doi:10.1056/NEJMoa2002032

174. Martinez ME. The calendar of epidemics: Seasonal cycles of infectious diseases. *PLOS Pathogens*. 2018;14(11):e1007327. doi:10.1371/journal.ppat.1007327

175. Baker RE, Mahmud AS, Metcalf CJE. Dynamic response of airborne infections to climate change: predictions for varicella. *Climatic Change*. 2018;148(4):547-560.

176. Moriyama M, Hugentobler WJ, Iwasaki A. Seasonality of Respiratory Viral Infections. *Annu Rev Virol*. 2020;7(1):83-101. doi:10.1146/annurev-virology-012420-022445

177. D'Amico F, Marmiere M, Righetti B, et al. COVID-19 seasonality in temperate countries. *Environ Res*. 2022;206:112614. doi:10.1016/j.envres.2021.112614

178. Hoogeveen MJ, Kroes ACM, Hoogeveen EK. Environmental factors and mobility predict COVID-19 seasonality in the Netherlands. *Environ Res*. 2022;211:113030. doi:10.1016/j.envres.2022.113030

179. Yamasaki L, Murayama H, Hashizume M. The impact of temperature on the transmissibility and virulence of COVID-19 in Tokyo, Japan. *Sci Rep*. 2021;11(1):24477. doi:10.1038/s41598-021-04242-3

180. Xie J, Zhu Y. Association between ambient temperature and COVID-19 infection in 122 cities from China. *Sci Total Environ*. 2020;724:138201. doi:10.1016/j.scitotenv.2020.138201

181. Kassem AZE. Does Temperature Affect COVID-19 Transmission? *Frontiers in Public Health*. 2020;8. Accessed March 18, 2022. https://www.frontiersin.org/article/10.3389/fpubh.2020.554964

182. Liu M, Li Z, Liu M, et al. Association between temperature and COVID-19 transmission in 153 countries. *Environ Sci Pollut Res*. 2022;29(11):16017-16027. doi:10.1007/s11356-021-16666-5

183. Dong Z, Fan X, Wang J, Mao Y, Luo Y, Tang S. Data-related and methodological obstacles to determining associations between temperature and COVID-19 transmission. *Environ Res Lett*. 2021;16(3):034016. doi:10.1088/1748-9326/abda71

184. Nottmeyer L, Armstrong B, Lowe R, et al. The association of COVID-19 incidence with temperature, humidity, and UV radiation – A global multi-city analysis.

Science of The Total Environment. 2023;854:158636. doi:10.1016/j.scitotenv.2022.158636

185. Villeneuve PJ, Goldberg MS. Methodological Considerations for Epidemiological Studies of Air Pollution and the SARS and COVID-19 Coronavirus Outbreaks. *Environmental Health Perspectives*. 2020;128(9):095001. doi:10.1289/EHP7411

186. Weaver AK, Head JR, Gould CF, Carlton EJ, Remais JV. Environmental Factors Influencing COVID-19 Incidence and Severity. *Annual Review of Public Health*. 2022;43(1):271-291. doi:10.1146/annurev-publhealth-052120-101420

187. Chatterjee P. Is India missing COVID-19 deaths? *The Lancet*. 2020;396(10252):657. doi:10.1016/S0140-6736(20)31857-2

188. Pifarré i Arolas H, Vidal-Alaball J, Gil J, López F, Nicodemo C, Saez M. Missing Diagnoses during the COVID-19 Pandemic: A Year in Review. *International Journal of Environmental Research and Public Health*. 2021;18(10):5335. doi:10.3390/ijerph18105335

189. Mecenas P, Bastos RT da RM, Vallinoto ACR, Normando D. Effects of temperature and humidity on the spread of COVID-19: A systematic review. *PLoS One*. 2020;15(9):e0238339. doi:10.1371/journal.pone.0238339

190. Sera F, Armstrong B, Abbott S, et al. A cross-sectional analysis of meteorological factors and SARS-CoV-2 transmission in 409 cities across 26 countries. *Nat Commun*. 2021;12(1):5968. doi:10.1038/s41467-021-25914-8

191. Smit AJ, Fitchett JM, Engelbrecht FA, Scholes RJ, Dzhivhuho G, Sweijd NA. Winter Is Coming: A Southern Hemisphere Perspective of the Environmental Drivers of SARS-CoV-2 and the Potential Seasonality of COVID-19. *Int J Environ Res Public Health*. 2020;17(16):5634. doi:10.3390/ijerph17165634

192. Carlson CJ, Gomez ACR, Bansal S, Ryan SJ. Misconceptions about weather and seasonality must not misguide COVID-19 response. *Nat Commun*. 2020;11(1):4312. doi:10.1038/s41467-020-18150-z

193. Baker RE, Yang W, Vecchi GA, Metcalf CJE, Grenfell BT. Susceptible supply limits the role of climate in the early SARS-CoV-2 pandemic. *Science*. 2020;369(6501):315-319. doi:10.1126/science.abc2535

194. Baker RE, Yang W, Vecchi GA, Metcalf CJE, Grenfell BT. Assessing the influence of climate on wintertime SARS-CoV-2 outbreaks. *Nat Commun*. 2021;12(1):846. doi:10.1038/s41467-021-20991-1

195. Martorell-Marugán J, Villatoro-García JA, García-Moreno A, et al. DatAC: A visual analytics platform to explore climate and air quality indicators associated with the COVID-19 pandemic in Spain. *Sci Total Environ*. 2021;750:141424. doi:10.1016/j.scitotenv.2020.141424

196. Villatoro-García JA, López-Domínguez R, Martorell-Marugán J, Luna J de D, Lorente JA, Carmona-Sáez P. Exploring the interplay between climate, population immunity and SARS-CoV-2 transmission dynamics in Mediterranean countries. *Sci Total Environ*. 2023;897:165487. doi:10.1016/j.scitotenv.2023.165487

197. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*. 2020;20(5):533-534. doi:10.1016/S1473-3099(20)30120-1

198. Hannah Ritchie DB Edouard Mathieu, Lucas Rodés Guirao, Cameron Appel, Charlie Giattino, Esteban Ortiz Ospina, Joe Hasell, Bobbie Macdonald, Roser M. Coronavirus Pandemic (COVID-19). *Our World in Data*. Published online 2020.

199. Mathieu E, Ritchie H, Ortiz-Ospina E, et al. A global database of COVID-19 vaccinations. *Nat Hum Behav*. 2021;5(7):947-953. doi:10.1038/s41562-021-01122-8

200. Hale T, Angrist N, Goldszmidt R, et al. A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). *Nat Hum Behav*. 2021;5(4):529-538. doi:10.1038/s41562-021-01079-8

201. PANEL COVID-19. Accessed August 19, 2024. https://cnecovid.isciii.es/covid19/

202. Datadista. Datadista. GitHub. 2021. Accessed June 9, 2022. https://github.com/datadista

203. AEMET. AEMET OpenData. 2023. Accessed February 8, 2023. https://opendata.aemet.es/centrodedescargas/inicio

204. Hersbach H, Bell B, Berrisford P, et al. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*. 2020;146(730):1999-2049. doi:10.1002/qj.3803

205. Junta de Andalucía. Informes diarios de calidad del aire :: Red de Información Ambiental de Andalucía :: Consejería de Medio Ambiente y Ordenación del Territorio :: Junta de Andalucía :: 2020. Accessed June 17, 2020. http://www.juntadeandalucia.es/medioambiente/site/rediam/

206. European Environment Agency. European Air Quality Portal – e-Reporting. 2020. Accessed June 9, 2021. https://aqportal.discomap.eea.europa.eu/

207. Cori A, Ferguson NM, Fraser C, Cauchemez S. A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics. *American Journal of Epidemiology*. 2013;178(9):1505-1512. doi:10.1093/aje/kwt133

208. Hastie T, Tibshirani R. Generalized Additive Models. *Statistical Science*. 1986;1(3):297-310. doi:10.1214/ss/1177013604

209. Ma Y, Pei S, Shaman J, Dubrow R, Chen K. Role of meteorological factors in the transmission of SARS-CoV-2 in the United States. *Nat Commun*. 2021;12(1):3602. doi:10.1038/s41467-021-23866-7

210. Wood SN. *Generalized Additive Models: An Introduction with R, Second Edition*. 2nd ed. Chapman and Hall/CRC; 2017. doi:10.1201/9781315370279

211. Wood S. mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation. Published online October 21, 2022. Accessed January 24, 2023. https://CRAN.R-project.org/package=mgcv

212. Gasparrini A, Armstrong B, Kenward MG. Distributed lag non-linear models. *Stat Med*. 2010;29(21):2224-2234. doi:10.1002/sim.3940

213. Chien LC, Guo Y, Li X, Yu HL. Considering spatial heterogeneity in the distributed lag non-linear model when analyzing spatiotemporal data. *J Expo Sci Environ Epidemiol*. 2018;28(1):13-20. doi:10.1038/jes.2016.62

214. Gasparrini A, Armstrong B. Reducing and meta-analysing estimates from distributed lag non-linear models. *BMC Medical Research Methodology*. 2013;13(1):1. doi:10.1186/1471-2288-13-1

215. Gasparrini A. Distributed Lag Linear and Non-Linear Models in R: The Package dlnm. *J Stat Softw*. 2011;43(8):1-20.

216. Gasparrini A, Armstrong B, Kenward MG. Multivariate meta-analysis for non-linear and other multi-parameter associations. *Stat Med*. 2012;31(29):3821-3839. doi:10.1002/sim.5471

217. Xu B, Kraemer MUG, Open COVID-19 Data Curation Group. Open access epidemiological data from the COVID-19 outbreak. *Lancet Infect Dis*. 2020;20(5):534. doi:10.1016/S1473-3099(20)30119-5

9. ANEXO: ARTÍCULOS

9.1.A survey of gene expression meta-analysis: methods and applications

Este artículo fue publicado en la revista *Briefings in Bioinformatics*, Volumen 22, Número 2, Marzo 2021, Páginas 1694–1705, DOI: https://doi.org/10.1093/bib/bbaa019. Esta es la versión aceptada del artículo. De acuerdo con la editorial (Oxford University Press) esta versión del artículo tiene un permiso de reutilización no comercial después de 12 meses de embargo que terminaron el 1 de marzo de 2022.

A survey of gene expression meta-analysis: methods and applications

Daniel Toro-Dominguez [1, *], Juan Antonio Villatoro-García [1, *], Jordi Martorell-Marugan [1], Yolanda Román-Montoya [2], Marta E. Alarcón-Riquelme [1,3], Pedro Carmona-Sáez [1]

[1] GENYO. Centre for Genomics and Oncological Research: Pfizer / University of Granada / Andalusian Regional Government. PTS Granada - Avenida de la Ilustración, 114 - 18016, Granada, Spain
[2] Department of Statistics and Operations Research, Universidad de Granada, Spain.
[3] Unit of Inflammatory Diseases, Department of Environmental Medicine, Karolinska Institute, 171 67, Solna, Sweden.
Corresponding author: Pedro Carmona-Sáez, pedro.carmona@genyo.es, Centre for Genomics and Oncological Research: Pfizer / University of Granada / Andalusian Regional Government. PTS Granada - Avenida de la Ilustración, 114 - 18016, Granada, Spain. Tel: +34 958715500.
[*] These authors contributed equally to this work

Abstract

The increasing use of high-throughput gene expression quantification technologies over the last two decades and the fact that most of the published studies are stored in public databases has triggered an explosion of studies available through public repositories. All this information offers an invaluable resource for reuse to generate new knowledge and scientific findings. In this context, great interest has been focused on meta-analysis methods to integrate and jointly analyze different gene expression datasets. In this work we describe the main steps in the gene expression meta-analysis, from data preparation to the state-of-the art statistical methods. We also analyze the main types of applications and problems that can be approached in gene expression meta-analysis studies and provide a comparative overview of the available software and bioinformatics tools. Moreover, a practical guide for choosing the most appropriate method in each case is also provided.

1. Introduction

The development of high-throughput gene expression quantification technologies (gene expression arrays and RNA-Seq) has been key in the advancement of biomedical research, allowing researchers to measure genome-wide gene expression patterns in an experimental condition. Their increasing use over two decades and the fact that most of the published studies are stored in public databases has triggered an explosion in the amount of gene expression datasets that are available through public repositories. There are general databases such as the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) [1], which stores expression data from over 100000 different studies comprising a total of more than 3 million samples, or ArrayExpress [2],

which contains expression data from more than 2.3 million curated samples. These resources store data from a broad range of platforms and organisms, but there are also more specific repositories focused on particular diseases, organisms or tissues, such as The Cancer Genome Atlas (TCGA) [3], which stores multi-omics data from 33 cancer subtypes, or GTEx [4], that collects gene expression data from 54 non-diseased human tissue sites across nearly 1000 individuals. All this information offers an invaluable resource to reuse and carry out integrated analysis to generate new knowledge and scientific findings. Meta-analysis techniques are a set of statistical methods that combine multiple and independent studies to obtain a single common and significant result. In the context of gene expression, datasets from different cohorts or studies can be analyzed in a meta-analysis in order to define common molecular signatures, improving reproducibility or obtaining more reliable biomarkers [5,6].

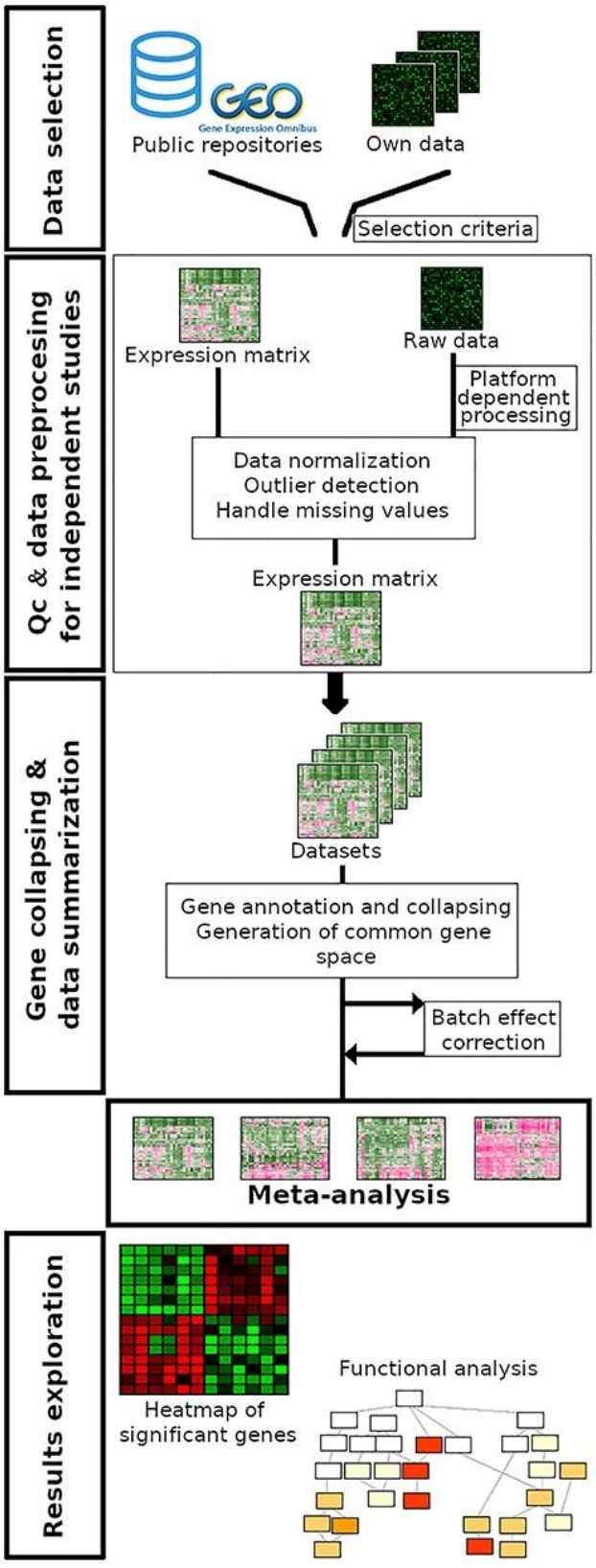


Figure 1 Overview of gene expression meta-analysis workflow. First, input data is selected from public repositories or own data. Secondly, the data must be preprocessed to get gene expression matrices and quality controls are performed. Then, a common gene space is created for all the datasets where batch effect correction can be applied and, finally, the meta-analysis is performed. Qc: Quality control.

In recent years, the number of publications based on gene expression meta-analysis has increased enormously. For instance, Huan et al. [7] combined a total of 7017 blood pressure and hypertension samples from different studies and identified sets of differentially expressed genes related to specific clinical manifestations within the disease. In the study by de Magalhães et al. [8], they used 27 datasets from Gene Aging Nexus [9] and GEO to discover age-related gene expression profiles obtaining 56 genes consistently overexpressed with age. Meta-analyses have been widely used in multiple types of cancer for the identification of biomarkers of activity and progression of the disease [10,11]. Meta-analyses are also especially useful in the case of rare or less frequently studied diseases, with a limited number of studies and small cohorts of patients. Meta-analysis allows increasing the sample size by incorporating samples from different cohorts, increasing the statistical power and the robustness of the results. As examples of successful meta-analyses, two genes altered in schizophrenia as compared to healthy controls was the result obtained by Piras et al. [6] following meta-analysis of three GEO datasets, while Lining Su et al. [12] identified a series of altered molecular mechanisms combining several independent Alzheimer's disease expression studies. Some meta-analyses have been also performed in the context of autoimmune disorders [13] and other rare diseases [14–16].

In this review, we provide an overview of the main statistical methods used in gene expression meta-analysis and the main applications of these types of techniques. We also analyze and review the most popular tools and software available for gene expression meta-analysis.

2. Meta-Analysis Workflow in Gene Expression data

In this section, we explain important considerations and key steps in gene expression meta-analysis [17–19]. Figure 1 summarizes the different points collected in this section.

- *Hypothesis and data selection*. Based on the aim of the study, eligibility criteria are defined to search and select datasets to be included. These criteria could be biological (disease, tissue, age, etc.) and/or technical (platform, sample size, quality controls, etc.). Public repositories allow automatic quick searches using keywords or ontologies. Different parameters must be considered for data selection since these might influence the results. Inter-study heterogeneity refers to how diverse, technically or biologically, the studies are. Different patient cohorts generated on different profiling platforms at different laboratories often reveal conflicting conclusions for the same question. Selecting high or low inter-study heterogeneity can serve different purposes: high inter-study heterogeneity reduces statistical power, but in fact increases the generalizability of the results [19,20]. On the other hand, it is important to work with balanced data regarding the number of samples from each class and the sample size of each study in order to homogenize the weight that each study exerts on the results. As a general rule, the isolated effect of a study is minimized with a larger number of studies, prioritizing the general meaning of the results [21].
- *Data preprocessing and normalization*. The first step, if starting from raw data, is to generate an expression matrix for each dataset. This process is different depending on the technology used to generate the data. For instance, RNA-Seq reads must be aligned against a reference genome to obtain gene counts, while microarrays data must be processed with different methods depending on the

platform. A common and important preprocessing step is normalization to minimize non biological variations [22]. There are excellent reviews about this topic for RNA-Seq [23] and for microarrays [24]. Public data repositories often have available preprocessed data. Data preprocessing should be as standardized as possible between studies in order to minimize technical heterogeneity.

- *Individual quality control*. Quality controls must be carried out to detect outliers, inconsistent measurements, technical problems in samples, etc. Outliers in data introduce bias into posterior analysis leading to under- or over-estimated resulting values. Handling missing data on samples if an example of common problem that can produce unreliable results when inferences about a population are drawn based on such samples. Therefore, the results are considerably dependent on methods used to process missing values and outliers [25]. Commonly, in expression studies, the identification of outlier samples is based on abnormal values distribution respect to the normality or on measures of similarity between samples, such as correlations, clustering, Mahalanobis distances or principal component analysis [26–28]. The problem of missing values in genes within a dataset can be assessed by imputation. Some of the most common imputation methods are the replacement of the missing value by expression average for that gene or the use of models derived from similarity between genes, such as K-nearest neighbor algorithm [29,30].
- *Gene collapsing and data summarization.* If datasets are from different platforms, it is necessary to annotate the probe sets to common and standard identifiers such as Entrez Gene identifiers, Ensembl ID or gene symbol and, then, collapsing expression values for all probes belonging to the same gene. The main approaches are based on summarizing each gene as the probe that contains the highest or lowest mean, absolute mean or variance, or summarize a gene as the average values of all their probes [31]. One of the problems at the next point is that there may be genes not shared across studies, mainly if studies from different platforms are used. The classic solution is based on the selection of only common genes, although this can cause important genes to be lost for the context studied [32]. New alternatives are emerging to address this problem, such as the imputation of gene expression through regression models using the genes shared in all datasets [32,33].
- *Batch effect*. Optionally, if the reduction of technical inter-heterogeneity is desired, batch-effect correction techniques can be applied in order to reduce or eliminate the effect of external factors and technical biases that may overshadow the gene expression caused directly by the study condition on gene expression. That is, the batch effect correction is applied to eliminate unwanted variation sources. Some of the most used approaches are the empirical Bayes methods [34], the adjustment through the use of covariates or the elimination of variation caused by surrogate variables, being the limma and sva R packages two of the most important tools to apply them [35,36].
- *Meta-analysis method selection*. The convenience of using one method over another depends both on the data features, goals and analytical design. In the Methods section, we discuss in detail the selection of the meta-analysis method. In this context, the technical heterogeneity (different platforms) and the biological heterogeneity (different study conditions or phenotypes) must be considered, in addition to the number of studies and how stringent the analysis should be. There are different tests to evaluate heterogeneity, like Cochran's Q test, advisable when the number of studies is large, and the I^2 statistic (derived from Cochran's Q)

which describes the percentage of variation across studies due to heterogeneity rather than by chance that is not dependent on the number of studies [37,38].

- *Result interpretatio*n. Generally, the output of a gene expression meta-analysis consists of a list of differentially expressed genes through different studies. Visualization tools such as heatmaps or interaction network plots of these genes in the different studies provide help in the explorative analysis, and further functional analysis is usually applied to characterize the main biological functions associated with the gene signature.

3. Meta-analysis methods

There are three main types of meta-analysis methods: meta-analysis based on effect sizes, meta-analysis based on P-values combination and meta-analysis based on ranks combination. Supplementary Table 1 shows a summary of the characteristics, advantages and disadvantages of each method and Figure 2 serves as a guide to decide which method is most appropriate based on the characteristics of the analysis.

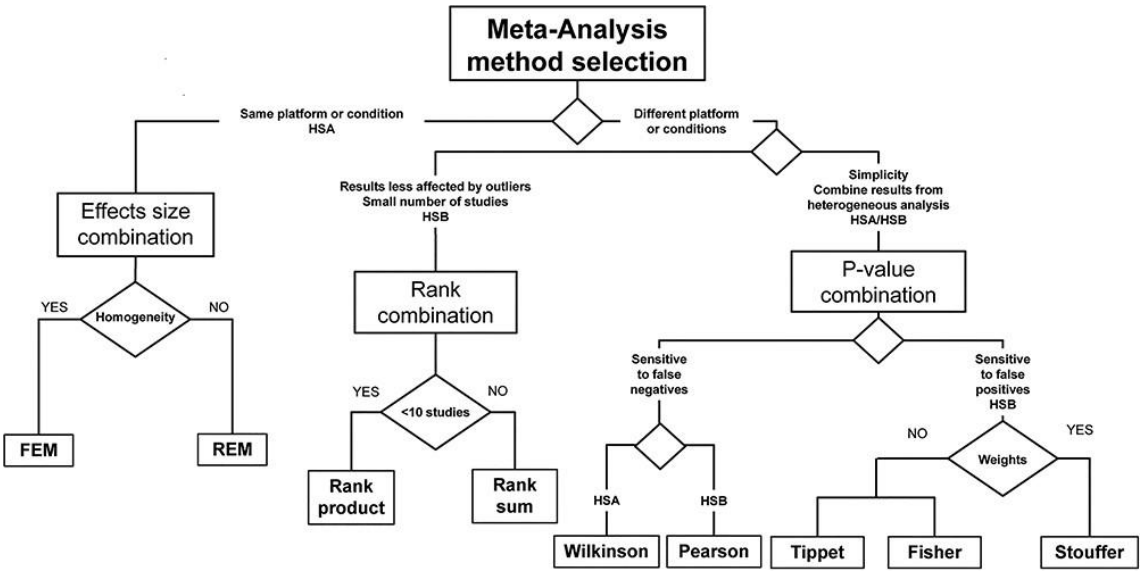


Figure 2 Meta-analysis method decision scheme. The figure summarizes the main recommendations for the selection of the most appropriate meta-analysis method based on the characteristics of the data. HSA: when the gene is significant in the meta-analysis it is significant in all studies. HSB: when the gene is significant in the meta-analysis, it is significant in at least one study. These are recommendations especially related to the type of the data. In the final selection of the method, it is also important to consider the objective of the meta-analysis.

3.1. Methods based on effect size combination

This type of methods aims to explain the difference between the strength of a phenomenon (called effect) in different studies. The calculated effect size depends on the type of data and the type of studies that are being carried out [39–41]. For example, in Genome-Wide Association Studies (GWAS), in which this type of methods has been widely used [42], the calculated effect is related to the odds ratio [43–45]. In the case of gene expression data, the effect size is the differential expression between two groups (e.g. cases and controls), which must follow a normal distribution [46]. The Hedges' g estimator [46–48] is one of the most recommended measurements to calculate effect size:

$$T_i = c(m)\frac{\bar{y}_E - \bar{y}_C}{s} \quad (1)$$

where:

- T_i is the effect size of one gene for the i-th dataset
- $m = n_E + n_C - 2$ represents the appropiate degrees of freedom,
- $c(m) = 1 - \frac{3}{4m-1} = 1 - \frac{3}{4(n_E+n_C)-9}$, is a factor that corrects the positive bias.
- $\bar{y}_E$ and $\bar{y}_C$ are the gene expression means of the case (experimental) and control groups.
- $S = \sqrt{\frac{(n_E-1)S_E^2+(n_C-1)S_C^2}{n_E+n_C-2}}$, standard deviation between studies, where:
    - n_E and n_C are the sample sizes of the experimental and control groups.
    - S_E^2 and S_C^2 are the variances of the experimental and control groups.

Moreover, the variance of an effect size is calculated:

$$V(T_i) = \frac{n_E+n_C}{n_E \times n_C} + \frac{(T_i)^2}{2(n_E+n_C)} \quad (2)$$

Once effects sizes and their variances are calculated, the aim is to obtain a *combined effect* of all of them for each gene. This *combined effect* will allow to discern if one gene is differentially expressed across all the studies. There are two possible models to obtain the *combined effect*, the Fixed Effects Model (FEM) and the Random Effects Model (REM).

Fixed Effects Model

FEM is a linear model that considers that the different studies share a common effect size called true effect. The *combined effect,* $\overline{T}_.$, is calculated [49] as:

$$\overline{T}_. = \frac{\Sigma \omega_i T_i}{\Sigma \omega_i} \quad (3)$$

where ω_i are the different weights assigned to each study, that is, the inverse within-study variance, $V(T_i)$ [49]:

$$\omega_i = \frac{1}{V(T_i)} \quad (4)$$

The variance of the combined effect is defined as:

$$V(\overline{T}_.) = \frac{1}{\Sigma \omega_i} \quad (5)$$

The *combined effect* value for a standard normal [49]:

$$Z = \frac{\overline{T}_.}{\sqrt{V(\overline{T}_.)}} \quad (6)$$

Therefore, we can calculate the two-tailed P-value [49]:

$$P = 2[1 - (\Phi(|Z|))] \quad (7)$$

where $\Phi$is the standard normal cumulative distribution function.

This model is recommended when there is homogeneity between the different studies or if all samples are assumed to come from the same population [39,50]. Therefore, this method is very restrictive and it should be applied with caution.

Random Effects Model

REM assumes that there is a distribution of the true effect sizes (i.e. the true effect differs between studies). Consequently, in this model the *combined effect* represents the mean of the population of true effect sizes instead of representing an identical effect size shared by all studies like in FEM [47,49].

In this model, the *combined effect*, $\overline{T}_{\cdot}^{*}$, is calculated similarly to the FEM [47,49]:

$$\overline{T}_{\cdot}^{*} = \frac{\sum_{i=1}^{k} \omega_i^* T_i}{\sum_{i=1}^{k} \omega_i^*} \quad (8)$$

The main difference is in the way the weights, ω_i^*, are obtained.

For assigning weights, in this model there are two sources of error: within-study variance and between-study variance. The within-study variance (v_i) is obtained when one observed effect is calculated, but it is necessary to obtain the between-study variance (τ^2). For obtaining between-study variance, the total variance must be calculated and then within-study variance must be isolated.

The statistic that represents the total variance, *Q*, is defined as:

$$Q = \sum_{i=1}^{k} \omega_i (T_i - \overline{T}_{\cdot}) \quad (9)$$

where:

- T_i is the observed effect
- ω_i is the calculated weights for the FEM
- $\overline{T}_{\cdot}$ is the *combined effect* calculated for the FEM

Once Q statistic has been obtained, it is decomposed in order to obtain the between-study variance (τ^2) [49]:

$$\tau^2 = \begin{cases} \frac{Q-df}{C} \; if \; Q > df \\ 0 \; if \; Q \leq df \end{cases} \quad (10)$$

where:

- $C = \sum_{i=1}^{k} \omega_i - \frac{\sum_{i=1}^{k} {\omega_i}^2}{\sum_{i=1}^{k} \omega_i}$
- *df = (number of studies - 1)*, degrees of freedom in case the only source of variance were the within-study error.

Once between-study variance is obtained, the weights assigned to each study can be calculated:

$$\omega_i{}^* = \frac{1}{v_i^*} \tag{11}$$

where v_i^* is:

$$v_i{}^* = V(T_i) + \tau^2 \tag{12}$$

The variance of the *combined effect* is defined:

$$V(\overline{T}_{\cdot}^*) = \frac{1}{\sum_{i=1}^{k} \omega_i^*} \tag{13}$$

The *combined effect* value for a standard normal [25]:

$$Z = \frac{\overline{T}_{\cdot}^*}{\sqrt{V(\overline{T}_{\cdot}^*)}} \tag{14}$$

Therefore, we can calculate the two-tailed p-value [49]:

$$P = 2[1 - (\Phi(|Z|))] \tag{15}$$

where Φ is the standard normal cumulative distribution function.

This model is the most frequently used among meta-analysis techniques [51]. It is recommended when there is some heterogeneity between studies, but without large differences between them. Its main advantage is that it is much less restrictive than FEM and usually fits the biological truth better [19,39,50].

3.2. Methods based on P-values combination

These techniques are aimed to integrate the P-values of individual analyses into one single combined P-value per gene.

One of the characteristics of these methods is that all studies influence the results equally regardless of their size, since the P-values obtained individually are combined directly, unless weights can be included. Moreover, these methods are more suitable to combine studies from different platforms or conditions than effects size combination approaches [52,53].

However, P-value based methods have the important disadvantage of losing the directionality of the expression pattern [51]. For example, when common patterns are searched for, if a gene is significantly overexpressed in some studies and significantly underexpressed in others, a significant combined P-value can be obtained regardless of the expression patterns in the different studies, which may not reflect the biological reality. One possible solution is to select only the genes that have the same fold-change (FC) direction and remove the conflicting genes [54].

As in the case of effects size combination, the differential expression between the case and control groups must be tested. Once the P-values for each of the studies are obtained, there are several methods that allow their combination.

Fisher's method

This method uses as statistic the sum of the logarithms of the P-values [51,53,55], that is to say:

$$-2 \times \sum_{i=1}^{k} ln\,(p_i) \tag{16}$$

where p_i is each of the P-values of the different studies. In this case, the null hypothesis is that there is no difference in gene expression between the different studies and it is distributed as a χ^2 with $2 \times k$ degrees of freedom (being *k* the number of studies).

When one gene is obtained as significant, this means that this gene is significant in one or more studies. Furthermore, Fisher's method is sensitive to very small P-values, therefore a single very significant study can lead to a significant combined P-value [56,57].

Pearson's method

The statistic of this method is:

$$-2 \times \sum_{i=1}^{k} ln\,(1 - p_i) \tag{17}$$

This method has similar characteristics to the Fisher's method and it can be used in the same situations. The only difference is that this method is most sensitive to large P-values, therefore, more false negatives are obtained [56].

Stouffer's method

This technique assumes that:

$$Z_i = \Phi^{-1}(1 - p_i) \tag{18}$$

Where Φ is the standard normal cumulative distribution function.

The statistic used in this method is [55]:

$$\frac{\sum_{i=1}^{k} Z_i}{\sqrt{k}} \tag{19}$$

Under the null hypothesis, it is distributed as a standad normal distribution, *N(0,1)*

One advantage of this method is that it allows including weights for the studies. In this case, the statistic is:

$$\frac{\sum_{i=1}^{k} \omega_i Z_i}{\sqrt{\sum_{i=1}^{k} \omega_i^2}} \tag{20}$$

The assigned weights (ω_i) should be the inverse variance of the statistics used to obtain the different P-values. However, if these variances could not be calculated, the square roots of sample sizes ($\sqrt{n_i}$) can also be used because they provide good results too [58,59].

Taking all of the above into consideration, Stouffer's method is recommended where weights can be calculated, because weighted Stouffer's method usually provides more reliable results than Fisher's or the unweighted Stouffer's method [59].

Tippet's method (minimum of P-values)

The statistic of this method is the minimum of P-values of all studies:

$$min\ (p_1, p_2, \ldots, p_i, \ldots, p_k) \tag{21}$$

Under the null hypothesis it is distributed as *Beta (1,k)* [55].

When one gene shows significance, this means that this gene is significant in one or more studies, so many false positives may be obtained. Therefore, this method is recommended when the aim is to discard genes [56].

Wilkinson's method (maximum of P-values)

The statistic of this method is the maximum of P-values of all studies:

$$max\ (p_1, p_2, \ldots, p_i, \ldots, p_k) \tag{22}$$

Under the null hypothesis it is distributed as *Beta (k,1)* [55].

Unlike the rest of P-value combination methods, this technique has the advantage that if one gene shows significance, this gene is significant in all the studies. However, this may

also generate many false negatives. Therefore, this method is recommended when the aim is to identify the most robust genes [56].

3.3. Non-parametric methods: Rank combination.

Rank combination methods allow combining different studies based on any statistic that can be ordered, although FC is usually used. After the estimation of the FCs, their values are converted to ranks, that is to say, the gene with the smallest value of FC would be ranked as 1 and the gene with the highest value would be in the last position in order to calculate the underexpressed genes and the opposite order to estimate the overexpressed genes. The advantages of these methods are that they allow combining any kind of data (including heterogeneous data or even data from different platforms) and they prevent very significant P-values (outliers) coming from individual studies to influence the results [51,60]. Moreover, these methods usually obtain more accurate and robust results than P-values combination methods [61].

However, these methods have the disadvantage that they are very sensitive to diversity of variance. For example, diversity of variance can severely reduce the accuracy of these methods for sorting genes by differential expression or it could lead to obtain false significant P-values as Breitling and Herzyk described previously [62]. Furthermore, if a two-sided P-value is used in these methods, they may not detect the direction in opposite FC [17,51,60], therefore, it is recommendable to use a one-sided P-value to apply these methods.

Rank product:

In this method, the rank combination for each gene is calculated as the product of their ranks between the different studies [51,63]:

$$RP_g = \prod_i^k \ r_{ig} \tag{23}$$

Then, an empirical P-value is calculated randomizing the values within the matrix of ranks and recalculating the product of the ranks for the new values. If the rank obtained randomly for a gene is greater than the original, 1 is added to the error of that gene. This process is repeated n times and finally the empirical P-value for a gene is the total error obtained divided by n [51,62,64]. This method is recommended when the number of studies is small (less than 10 studies), because reliable results are obtained, but given that a product is applied the computational cost is very high [62].

Rank Sum

This method is similar to the previous one with the only difference that the sum of the rank combination is calculated instead of the product [51,63]:

$$RS_g = \sum_i^K \ r_{ig} \tag{24}$$

Although this method may obtain less robust results than the rank product method, it has the advantage that it is more efficient when there is a large number of studies, that is to say, it has a lower computational cost [62].

4. Applications

Gene expression meta-analysis techniques have been used in a wide range of contexts, although they can be summarized in three main types of applications (see Figure 3).

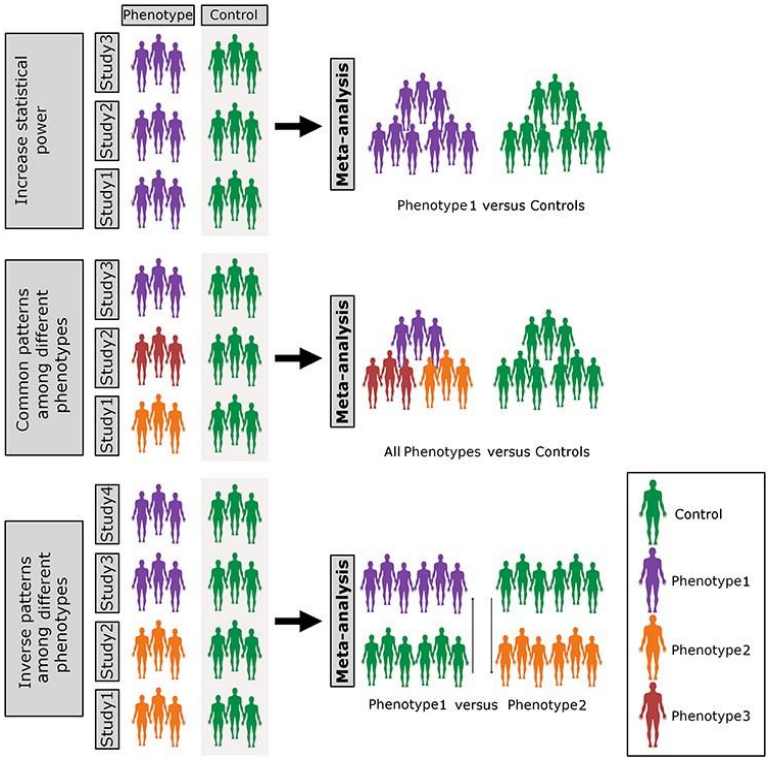


Figure 3 Applications of gene expression meta-analysis. The figure shows the three main applications or objectives of the gene expression meta-analysis. The increase in the statistical power, the search for homogeneous patterns between different phenotypes and inverse patterns.

The most common application is the analysis of different cohorts with the same phenotype in order to increase the statistical power to detect genes showing consistent differences between groups of cases and control. This meta-analysis application is therefore useful for extracting consistent biomarkers and has been widely applied in cancer [10,65], autoimmune diseases [66,67] or mental disorders [68].

The second application consists in looking for common gene expression patterns between different conditions, for example, genes differentially expressed between a set of different diseases respect to healthy samples, different disease states or a set of drug-derived gene profiles. This application is useful to identify shared molecular mechanisms and biological pathways across different phenotypes. This has been for example applied to analyze common gene expression patterns across autoimmune diseases [69–71] or neural disorders [72].

The third application is based on integrating gene expression data to discover inverse gene expression patterns between different conditions or diseases. The main idea is to identify sets of genes that are over-expressed in one phenotype but at the same time under-expressed in another. As an example of this applicability, Ibañez et al. [73] jointly analyzed several gene expression datasets to identify inverse gene patterns between Alzheimer's disease and cancer, pathologies described to have an inverse comorbidity. This workflow can be also applied for drug repurposing analysis. Briefly, a drug repurposing analysis searches for new drug indications for existing drugs, looking for

inverse gene patterns between a disease and a drug-caused gene profile. This is based on the hypothesis that if a drug induces gene patterns inverse to that of the disease, it could potentially reverse the pathogenic phenotype [74,75].

These three applications involve conceptual changes in the way the meta-analysis is conducted to respond to different questions, although they can share statistical methods. If the aim is to define common or inverse patterns across different phenotypes, the researcher could expect to find genes that are significantly expressed in all datasets to make sure that the findings significant across all phenotypes and, therefore, it would be advisable to use high restrictive approaches, such as the Wilkinson's method. Nevertheless, when the objective is to increase the sample size by integrating studies of the same phenotype, the use of methods that find significant patterns in most of the studies is appropiated.

5. Tools

In this section, we provide a description of publicly available tools and software suites for gene expression meta-analysis, focusing on web tools and R packages. Supplementary Table 2 provides a summary of the software presented and Figure 4 classify the different tools according to their implemented methods. Here we briefly review some characteristics, strengths and weaknesses of each tool, but it should be noticed that, for all tools, we consider the implemented methods as strengths and the unimplemented methods as weaknesses.

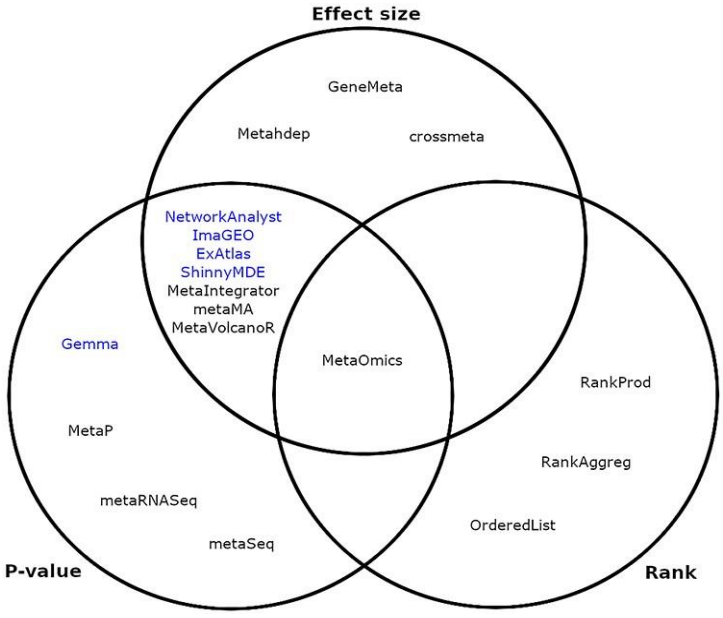


Figure 4 Summary of available gene expression meta-analysis tools. The tools are classified according to whether they implement the methods of combining effect size, P value, ranks or several of them. The font color represents the type of tool implementation (blue: web tool, black: R package).

5.1. Web based applications

- NetworkAnalyst [76] is one of the most popular web-based meta-analysis software, developed in Java Server Faces 2.0 technology. It contains different quality control steps (including batch effect correction and several normalization approaches for microarray and RNA-Seq data) and meta-analysis methods. It provides also different visualization and exploration functions including

functional analysis using KEGG [77] and Gene Ontology [78] databases. Possibly the main disadvantage of this software is that it lacks functionality for the use of data from public repositories. Users have to download and format the gene expression datasets before uploading them into the application. This specific format requirement may be a handicap for non-bioinformatician users. The application is available at: https://www.networkanalyst.ca/.

- ImaGEO [79] is a web tool developed in Shiny and R mainly focused on the meta-analysis of public gene expression data stored in the GEO database, directly selecting the GEO identifiers of the studies to be analyzed. It also allows users to upload their own datasets. ImaGEO identifies outlier samples, filtering of data based on missing values, and it offers the most used meta-analysis methods, including effect-based meta-analysis and P-value integration. The tool also performs quality control steps and functional analysis using the Gene Ontology database. It is available at: http://bioinfo.genyo.es/imageo/.
- Gemma [80] is a web tool designed to perform several types of analyses, including meta-analysis. It contains more than 3000 curated and preloaded datasets from GEO, including human, rat, and mouse studies, but users can also upload their own data (registration required). Although Gemma is a very useful resource for analyzing and exploring public expression data, available meta-analysis methods are limited to P-value combination. It is available at: https://gemma.msl.ubc.ca.
- ExAtlas [81] is another web tool designed to perform meta-analysis on both public (preloaded from GEO) and user defined data. It offers several analysis methods and results, including co-expression analysis and functional analysis using several databases like ENCODE [82], Gene Ontology and KEGG, among others. It is available at: https://lgsun.grc.nia.nih.gov/exatlas/.
- ShinyMDE [83] is a web tool based on R and Shiny that allows integrating different gene expression datasets from Affymetrix and Illumina microarrays allowing to start from raw data, and providing P-values combination approaches. A major limitation of ShinyMDE is that it cannot integrate datasets generated with different platforms. In addition, the input is a table of precomputed P-values, so differential gene expression analysis must be performed prior to using the tool. Available at: https://hussain.shinyapps.io/App-1/.

5.2. R packages

- MetaOmics [84,85] is an interactive Shiny based application, but we include it in this section because it requires some R knowledge to install it's dependencies and to launch it locally. MetaOmics includes different modules to perform all the meta-analysis steps interactively. It includes most of the meta-analysis methods based on effect size, P-value and rank combination. In addition, it offers modules to analyze the meta-analysis results (network analysis, pathway analysis, principal components analysis, etc.).
- MetaIntegrator [86] is an R package that implements some meta-analysis and visualization methods. It allows applying REM and Fisher's method.
- MetaP [87]. This R package implements 10 different methods of meta-analysis based on P-value combination, including the approaches described in the Methods section, such as the Wilkinson's method or Fisher's method, as well as some variations of these.

- GeneMeta [88] is an R package that uses the expression matrices from the individual studies and implements meta-analyses based on effect size combination.
- metaMA and metaRNASeq [52,54]. These are R packages that contain functions to perform meta-analysis based on combination of P-values and effect sizes (only metaMA). MetaMA and metaRNASeq are dedicated to microarray and RNA-Seq data, respectively. MetaMA allows starting from the expression matrices, as well as from lists of genes with their t-statistics. metaRNASeq combines lists of P-values with Fisher's or Stouffer'smethods. In addition, both metaMA and metaRNASeq are implemented as an interactive Galaxy tool [89].
- RankProd [90]. Is an R package specialized in the use of the product and the sum of ranks, both for differential expression between two conditions and for meta-analysis.
- RankAggreg [91]. This R package allows meta-analysis of ranked lists (for example, genes ordered based on FC) coming from different studies, using the rank sum algorithm.
- OrderedList [92]. This R package works with ranked lists of genes. The algorithm computes a similarity score based on the number of genes shared on the top across all lists.
- metahdep [93]. This R package starts from raw data from the Affymetrix microarray platform and allows performing meta-analysis based on effect size.
- metaSeq [94]. R package designed to apply Fisher's and Stouffer's methods to RNA-Seq differential expression analysis.
- MetaVolcanoR [95]. This R package applies REM and Fisher's method to expression matrices. Its main advantage is the graphical output: volcano plots and forest plots can be generated from the meta-analysis results.
- crossmeta [96]. This R package is focused on downloading and preprocessing microarray raw data from GEO. It applies effect size combination and performs pathways meta-analysis.

6. Conclusions

In the last few years, an important concern in biomedical research has been the reproducibility of results [97,98], specially in the context of high-throughput experiments such as gene expression microarray or genome-wide association studies where thousands of hypotheses are tested simultaneously. Technical and biological variation are major sources of irreproducibility combined with the analytical focus on significant P-values rather than effect sizes or independent verification [18].

The popularization of high-throughput technologies, together with the requirement of depositing experimental data in public repositories before publication, has made an unprecedented amount of data available. This scenario has opened alternatives to improve reproducibility by integrating multiple datasets via gene expression meta-analysis but also new ways to test and generate new hypotheses.

Consequently, there has been a tremendous growth in the number of systematic reviews and meta-analyses over the past decade. Similarly, different methods and variations of these have been developed, offering a wide range of possibilities to the researcher. While in GWAS there are consolidated software packages for meta-analysis such as METAL [99] or PLINK [100] as well as a good number of reviews, the field of gene expression

meta-analysis is less exploited. This is in part because there is much more heterogeneity in the number of gene expression platforms, experimental designs and the data itself.

In this review, we provided an overview of the most widely used meta-analysis methods applied on gene expression data, remarking pros and cons of each one depending on the application. We also discussed the main applications of gene expression meta-analysis and the workflow used in a common meta-analysis experiment, highlighting the key points that the user has to consider when conducting the analysis. Finally, we provided a comparative review of the different public tools available to perform gene expression meta-analysis, focusing on aspects such as meta-analysis methods, data input, quality control and functionalities for exploring the results.

This review has the aim of helping users to understand methodologies to perform meta-analyses based on gene expression data.

References

1. Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res 2013; 41:D991–D995

2. Athar A, Füllgrabe A, George N, et al. ArrayExpress update - from bulk to single-cell expression data. Nucleic Acids Res. 2019; 47:D711–D715

3. Weinstein JN, Collisson EA, Mills GB, et al. The Cancer Genome Atlas Pan-Cancer Analysis Project. Nat Genet 2013; 45:1113–1120

4. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat. Genet. 2013; 45:580–585

5. Cho H, Kim H, Na D, et al. Meta-analysis method for discovering reliable biomarkers by integrating statistical and biological approaches: An application to liver toxicity. Biochemical and Biophysical Research Communications 2016; 471:274–281

6. Piras IS, Manchia M, Huentelman MJ, et al. Peripheral Biomarkers in Schizophrenia: A Meta-Analysis of Microarray Gene Expression Datasets. Int J Neuropsychopharmacol 2018; 22:186–193

7. Huan T, Esko T, Peters MJ, et al. A Meta-analysis of Gene Expression Signatures of Blood Pressure and Hypertension. PLoS Genet 2015; 11:

8. de Magalhães JP, Curado J, Church GM. Meta-analysis of age-related gene expression profiles identifies common signatures of aging. Bioinformatics 2009; 25:875–881

9. Pan F, Chiu C-H, Pulapura S, et al. Gene Aging Nexus: a web database and data mining platform for microarray data on aging. Nucleic Acids Res 2007; 35:D756–D759

10. Bell R, Barraclough R, Vasieva O. Gene Expression Meta-Analysis of Potential Metastatic Breast Cancer Markers. Curr. Mol. Med. 2017; 17:200–210

11. Chen R, Khatri P, Mazur PK, et al. A meta-analysis of lung cancer gene expression identifies PTK7 as a survival gene in lung adenocarcinoma. Cancer Res. 2014; 74:2892–2902

12. Su L, Chen S, Zheng C, et al. Meta-Analysis of Gene Expression and Identification of Biological Regulatory Mechanisms in Alzheimer's Disease. Front Neurosci 2019; 13:

13. Kröger W, Mapiye D, Entfellner J-BD, et al. A meta-analysis of public microarray data identifies gene regulatory pathways deregulated in peripheral blood mononuclear cells from individuals with Systemic Lupus Erythematosus compared to those without. BMC Medical Genomics 2016; 9:66

14. Hamda CB, Sangeda R, Mwita L, et al. A common molecular signature of patients with sickle cell disease revealed by microarray meta-analysis and a genome-wide association study. PLOS ONE 2018; 13:e0199461

15. Zhang Z, Hailat Z, Falk MJ, et al. Integrative analysis of independent transcriptome data for rare diseases. Methods 2014; 69:315–325

16. Ch'ng C, Kwok W, Rogic S, et al. Meta-analysis of gene expression in autism spectrum disorder. Autism Res 2015; 8:593–608

17. Ramasamy A, Mondry A, Holmes CC, et al. Key Issues in Conducting a Meta-Analysis of Gene Expression Microarray Datasets. PLoS Med 2008; 5:

18. Sweeney TE, Haynes WA, Vallania F, et al. Methods to increase reproducibility in differential gene expression via meta-analysis. Nucleic Acids Res 2017; 45:e1

19. Waldron L, Riester M. Meta-Analysis in Gene Expression Studies. Statistical Genomics 2016; 1418:161–176

20. Jaksik R, Iwanaszko M, Rzeszowska-Wolny J, et al. Microarray experiments and factors which affect their reliability. Biol Direct 2015; 10:

21. Ioannidis JPA. Why most published research findings are false. PLoS Med. 2005; 2:e124

22. Wu Z. A review of statistical methods for preprocessing oligonucleotide microarrays. Stat Methods Med Res 2009; 18:533–541

23. Conesa A, Madrigal P, Tarazona S, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016; 17:13

24. Tarca AL, Romero R, Draghici S. Analysis of microarray experiments of gene expression profiling. Am. J. Obstet. Gynecol. 2006; 195:373–388

25. Kwak SK, Kim JH. Statistical data preparation: management of missing values and outliers. Korean J Anesthesiol 2017; 70:407–411

26. Filzmoser P, Maronna R, Werner M. Outlier identification in high dimensions. Computational Statistics & Data Analysis 2008; 52:1694–1711

27. Hadi AS. Identifying Multiple Outliers in Multivariate Data. Journal of the Royal Statistical Society: Series B (Methodological) 1992; 54:761–771

28. Shieh AD, Hung YS. Detecting outlier samples in microarray data. Stat Appl Genet Mol Biol 2009; 8:Article 13

29. Aittokallio T. Dealing with missing values in large-scale studies: microarray data imputation and beyond. Brief. Bioinformatics 2010; 11:253–264

30. Liew AW-C, Law N-F, Yan H. Missing value imputation for gene expression data: computational techniques to recover missing data from available information. Brief. Bioinformatics 2011; 12:498–513

31. Miller JA, Cai C, Langfelder P, et al. Strategies for aggregating gene expression data: The collapseRows R function. BMC Bioinformatics 2011; 12:322

32. Bobak CA, McDonnell L, Nemesure MD, et al. Assessment of Imputation Methods for Missing Gene Expression Data in Meta-Analysis of Distinct Cohorts of Tuberculosis Patients. Pac Symp Biocomput 2020; 25:307–318

33. Wang KY, Vankov ER, Lin DDM. Predictors of clinical outcome in pediatric oligodendroglioma: meta-analysis of individual patient data and multiple imputation. J Neurosurg Pediatr 2018; 21:153–163

34. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 2007; 8:118–127

35. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015; 43:e47

36. Leek JT, Johnson WE, Parker HS, et al. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics 2012; 28:882–883

37. Higgins JPT, Thompson SG, Deeks JJ, et al. Measuring inconsistency in meta-analyses. BMJ 2003; 327:557–560

38. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Statistics in Medicine 2002; 21:1539–1558

39. Nakagawa S, Noble DWA, Senior AM, et al. Meta-evaluation of meta-analysis: ten appraisal questions for biologists. BMC Biol. 2017; 15:18

40. Nakagawa S, Cuthill IC. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biol Rev Camb Philos Soc 2007; 82:591–605

41. Tang LL, Caudy M, Taxman F. A statistical method for synthesizing meta-analyses. Comput Math Methods Med 2013; 2013:732989

42. Kavvoura FK, Ioannidis JPA. Methods for meta-analysis in genetic association studies: a review of their potential and pitfalls. Hum. Genet. 2008; 123:1–14

43. Jakobsdottir J, Gorin MB, Conley YP, et al. Interpretation of genetic association studies: markers with replicated highly significant odds ratios may be poor classifiers. PLoS Genet. 2009; 5:e1000337

44. Waltoft BL, Pedersen CB, Nyegaard M, et al. The importance of distinguishing between the odds ratio and the incidence rate ratio in GWAS. BMC Med. Genet. 2015; 16:71

45. Stringer S, Wray NR, Kahn RS, et al. Underestimated effect sizes in GWAS: fundamental limitations of single SNP analysis for dichotomous phenotypes. PLoS ONE 2011; 6:e27964

46. Hedges LV. Fitting Categorical Models to Effect Sizes from a Series of Experiments. Journal of Educational Statistics 1982; 7:119–137

47. Cohn LD, Becker BJ. How meta-analysis increases statistical power. Psychol Methods 2003; 8:243–253

48. Ellis PD. The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results. 2010;

49. Borenstein M, Hedges L, Rothstein H. Meta-Analysis Fixed effect vs. random effects. 162. Introduction to Meta-Analysis. 2009;

50. Nakagawa S, Santos ESA. Methodological issues and advances in biological meta-analysis. Evol Ecol 2012; 26:1253–1274

51. Siangphoe U, Archer KJ. Estimation of random effects and identifying heterogeneous genes in meta-analysis of gene expression studies. Brief. Bioinformatics 2017; 18:602–618

52. Marot G, Foulley J-L, Mayer C-D, et al. Moderated effect size and P-value combinations for microarray meta-analyses. Bioinformatics 2009; 25:2692–2699

53. Sutton AJ, Jones DR, Sheldon T, et al. Methods for Meta-analysis in Medical Research. 2003; 22:

54. Rau A, Marot G, Jaffrézic F. Differential meta-analysis of RNA-seq data from multiple studies. BMC Bioinformatics 2014; 15:91

55. Li J, Tseng GC. An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. Ann. Appl. Stat. 2011; 5:994–1019

56. Heard N, Rubin-Delanchy P. Choosing Between Methods of Combining p-values. Biometrika 2018; 105:239–246

57. Song C, Tseng GC. HYPOTHESIS SETTING AND ORDER STATISTIC FOR ROBUST GENOMIC META-ANALYSIS. Ann Appl Stat 2014; 8:777–800

58. Zaykin DV. Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. J Evol Biol 2011; 24:1836–1841

59. Whitlock MC. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. J. Evol. Biol. 2005; 18:1368–1373

60. Tseng GC, Ghosh D, Feingold E. Comprehensive literature review and statistical considerations for microarray meta-analysis. Nucleic Acids Res. 2012; 40:3785–3799

61. Hong F, Breitling R. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. Bioinformatics 2008; 24:374–382

62. Breitling R, Herzyk P. Rank-based methods as a non-parametric alternative of the T-statistic for the analysis of biological microarray data. J Bioinform Comput Biol 2005; 3:1171–1189

63. Chang L-C, Lin H-M, Sibille E, et al. Meta-analysis methods for combining multiple expression profiles: comparisons, statistical characterization and an application guideline. BMC Bioinformatics 2013; 14:368

64. Breitling R, Armengaud P, Amtmann A, et al. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS Lett. 2004; 573:83–92

65. O'Mara TA, Zhao M, Spurdle AB. Meta-analysis of gene expression studies in endometrial cancer identifies gene expression profiles associated with aggressive disease and patient outcome. Scientific Reports 2016; 6:36677

66. Afroz S, Giddaluru J, Vishwakarma S, et al. A Comprehensive Gene Expression Meta-analysis Identifies Novel Immune Signatures in Rheumatoid Arthritis Patients. Front Immunol 2017; 8:74

67. Song GG, Kim J-H, Seo YH, et al. Meta-analysis of differentially expressed genes in primary Sjogren's syndrome by using microarray. Human Immunology 2014; 75:98–104

68. Patel H, Dobson RJB, Newhouse SJ. A Meta-Analysis of Alzheimer's Disease Brain Transcriptomic Data. J Alzheimers Dis 68:1635–1656

69. Badr MT, Häcker G. Gene expression profiling meta-analysis reveals novel gene signatures and pathways shared between tuberculosis and rheumatoid arthritis. PLOS ONE 2019; 14:e0213470

70. Toro-Domínguez D, Carmona-Sáez P, Alarcón-Riquelme ME. Shared signatures between rheumatoid arthritis, systemic lupus erythematosus and Sjögren's syndrome uncovered through gene expression meta-analysis. Arthritis Res. Ther. 2014; 16:489

71. Tuller T, Atar S, Ruppin E, et al. Common and specific signatures of gene expression and protein-protein interactions in autoimmune diseases. Genes Immun. 2013; 14:67–82

72. Kelly J, Moyeed R, Carroll C, et al. Gene expression meta-analysis of Parkinson's disease and its relationship with Alzheimer's disease. Molecular Brain 2019; 12:16

73. Ibáñez K, Boullosa C, Tabarés-Seisdedos R, et al. Molecular Evidence for the Inverse Comorbidity between Central Nervous System Disorders and Cancers Detected by Transcriptomic Meta-analyses. PLOS Genetics 2014; 10:e1004173

74. Toro-Domínguez D, Carmona-Sáez P, Alarcón-Riquelme ME. Support for phosphoinositol 3 kinase and mTOR inhibitors as treatment for lupus using in-silico drug-repurposing analysis. Arthritis Res. Ther. 2017; 19:54

75. Lamb J, Crawford ED, Peck D, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science 2006; 313:1929–1935

76. Zhou G, Soufan O, Ewald J, et al. NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. Nucleic Acids Res. 2019; 47:W234–W241

77. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000; 28:27–30

78. . The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 2004; 32:D258–D261

79. Toro-Domínguez D, Martorell-Marugán J, López-Domínguez R, et al. ImaGEO: integrative gene expression meta-analysis from GEO database. Bioinformatics 2019; 35:880–882

80. Zoubarev A, Hamer KM, Keshav KD, et al. Gemma: a resource for the reuse, sharing and meta-analysis of expression profiling data. Bioinformatics 2012; 28:2272–2273

81. Sharov AA, Schlessinger D, Ko MSH. ExAtlas: An interactive online tool for meta-analysis of gene expression data. J Bioinform Comput Biol 2015; 13:1550019

82. . An Integrated Encyclopedia of DNA Elements in the Human Genome. Nature 2012; 489:57–74

83. Shashirekha HL, Wani AH. ShinyMDE: Shiny tool for microarray meta-analysis for differentially expressed gene detection. 2016 International Conference on Bioinformatics and Systems Biology (BSB) 2016; 1–5

84. Ma T, Huo Z, Kuo A, et al. MetaOmics: analysis pipeline and browser-based software suite for transcriptomic meta-analysis. Bioinformatics 2019; 35:1597–1599

85. Forero DA. Available software for meta-analyses of genome-wide expression studies. 2019;

86. Haynes WA, Vallania F, Liu C, et al. Empowering Multi-Cohort Gene Expression Analysis to Increase Reproducibility. Pac Symp Biocomput 2016; 22:144–153

87. Dewey M. metap: meta-analysis of significance values. 2019;

88. Lusa L, Gentleman R, Ruschhaupt M. GeneMeta: MetaAnalysis for High Throughput Experiments. 2019;

89. Blanck S, Marot G. SMAGEXP: a galaxy tool suite for transcriptomics data meta-analysis. arXiv:1802.08251 [q-bio, stat] 2018;

90. Hong F, Breitling R, McEntee CW, et al. RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. Bioinformatics 2006; 22:2825–2827

91. Pihur V, Datta S, Datta S. RankAggreg, an R package for weighted rank aggregation. BMC Bioinformatics 2009; 10:62

92. Lottaz C, Yang X, Scheid S, et al. OrderedList--a bioconductor package for detecting similarity in ordered gene lists. Bioinformatics 2006; 22:2315–2316

93. Stevens JR, Nicholas G. metahdep: meta-analysis of hierarchically dependent gene expression studies. Bioinformatics 2009; 25:2619–2620

94. Tsuyuzaki K, Nikaido I. metaSeq: Meta-analysis of RNA-Seq count data in multiple studies. 2019;

95. Prada C, Lima D, Nakaya H. MetaVolcanoR: Gene Expression Meta-analysis Visualization Tool. 2020;

96. Pickering A. crossmeta: Cross Platform Meta-Analysis of Microarray Data. 2020;

97. Goodman SN, Fanelli D, Ioannidis JPA. What does research reproducibility mean? Sci Transl Med 2016; 8:341ps12

98. Shi L, Jones WD, Jensen RV, et al. The balance of reproducibility, sensitivity, and specificity of lists of differentially expressed genes in microarray studies. BMC Bioinformatics 2008; 9:S10

99. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 2010; 26:2190–2191

100. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 2007; 81:559–575

Key points

- Public gene expression databases have growth exponentially providing and unprecedented number of datasets and studies that can be integrated to establish robust and reproducible results. It is essential to use proper gene expression meta-analysis techniques and methods to attain the best integration and combination of different studies.
- A standardized workflow is key for obtaining reliable results, which range from the selection and processing of data to the application of the most appropriate method for the desired analysis. Different statistical methods for gene expression meta-analysis have been developed to cover different data scenarios.

- Gene expression meta-analysis can be applied to answer different biological questions. These can be summarized into three main applications: integrating datasets from the same phenotype to increase statistical power, combine data from different phenotypes to define common biomarkers and combine data from different phenotypes to establish opposed gene expression signatures.
- There are several web tools and R packages available for gene expression meta-analyses with different characteristics and scope.

Keywords

Meta-analysis, gene expression, biomarker discovery, data integration, omics data, public tools

Funding

This work was partially supported by the Junta de Andalucía (PI-0173-2017) and the IMI-JU GA# 115565 from the European Union.

Daniel Toro-Dominguez is a doctor specialized in bioinformatics, biomedicine and integrated analysis of omic data. He is currently working in the Bioinformatics Unit and the Medical Genomics group of GENYO (Granada).

Juan Antonio Villatoro-García is a bioestatistician at GENYO Bioinformatics Unit (Granada). Msc in Applied Statistics and mainly interested in developing statistical methods for gene expression meta-analysis.

Jordi Martorell-Marugan is a PhD student in Bioinformatics applied to Biomedicine. He holds a MSc degree in Omics Data Analysis and he is interested in the development of new omics data analysis algorithms.

Yolanda Román-Montoya is a professor in the Department of Statistics and Operational Research at the University of Granada (Spain). She belongs to the research group Computational and Applied Statistics in which she developes her research.

Marta E. Alarcón-Riquelme, MD, PhD is Head of Medical Genomics at GENYO in Granada (Spain). She has had a longstanding career in genetics at Uppsala University, Sweden and the Oklahoma Medical Research Foundation in the US.

Pedro Carmona-Sáez is the head of the Bioinformatics Unit at the Centre for Genomics and Oncological Research (Granada, Spain). He mainly engages in bioinformatics research, focusing on developing of methods for integrating omics data to decipher molecular mechanisms of complex diseases.

Figures

Figure 1: Guided scheme of the key steps to perform a meta-analysis.

Figure 2: Guide for the selection of the meta-analysis method.

Figure 3: Applications of gene expression meta-analysis.

Figure 4: Summary of available gene exprexion meta-analysis tools. The tools are classified according to whether they implement the methods of combining effect size, p-value, ranks or several of them. The font color represents the type of tool implementation (blue: web tool, black: R package).

Supplementary files

Supplementary material can be downloaded at:
https://academic.oup.com/bib/article/22/2/1694/5753843.

9.2. DExMA: An R Package for Performing Gene Expression Meta-Analysis with Missing Genes

Este artículo fue publicado el 17 de septiembre de 2022 en la revista *Mathematics* volumen 10, número 18, página 3376, DOI: **https://doi.org/10.3390/math10183376** bajo licencia *Open Access*, bajo los términos y condiciones de Creative Commons Attribution (CC BY). Esta es la versión aceptada del artículo. De acuerdo con la editorial esta versión del artículo tiene permiso de reutilización sin restricciones.

DExMA: An R package for performing gene expression meta-analysis with missing genes

Juan Antonio Villatoro-García[1,2], Jordi Martorell-Marugán[1,2,3], Daniel Toro-Domínguez[4], Yolanda Román-Montoya[1], Pedro Femia[1] and Pedro Carmona-Sáez[1,2,*]

[1]Department of Statistics and Operational Research, University of Granada, 18071 Granada, Spain
[2]Bioinformatics Unit, Centre for Genomics and Oncological Research Pfizer/University of Granada/Andalusian Regional Government, 18016 Granada, Spain
[3]Data Science for Heath Research Unit, Fondazione Bruno Kessler, 38123 Trento, Italy
[4]Medical Genomics, Centre for Genomics and Oncological Research Pfizer/University of Granada/Andalusian Regional Government, 18016 Granada, Spain
[*]Author to whom correspondence should be addressed.

Abstract

Meta-analysis techniques allow researchers to jointly analyse different studies to determine common effects. In the field of transcriptomics, these methods have gained popularity in re-cent years due to the increasing number of datasets that are available in public repositories. De-spite this, there is a limited number of statistical software packages that implement proper me-ta-analysis functionalities for this type of data. This article describes **DExMA**, an R package that provides a set of functions for performing gene expression meta-analyses, from data downloading to results visualization. Additionally, we implemented functions to control the number of missing genes, which can be a major issue when comparing studies generated with different analytical platforms. DExMA is freely available in the Bioconductor repository.

1. Introduction

In recent years, due to the widespread use of high-throughput gene expression technologies, the amount of gene expression data stored in public databases such as GEO [1] has grown drastically [2]. Gene expression is the process by which a product (usually a protein) is generated from the information encoded in genes. Gene expression studies usually measure the expression levels of thousands of genes simultaneously and generate a gene expression matrix with thousands of variables (genes) and tens or hundreds of samples. Each element of the matrix represents the amount of mRNA of a gene in a sample. One of the main types of analyses carried out in these studies is finding genes that are differentially expressed among groups of samples by means of hypothesis testing of mean differences (case–control studies). Therefore, these databases are invaluable resources to help researchers perform new analyses and gain new scientific insights.

A meta-analysis is a statistical technique that has achieved considerable popularity during the last few years for the integration of gene expression studies and making inferences

about a population of interest. Gene expression meta-analyses have been widely applied for different purposes such as increasing statistical power for the identification of biomarkers [3,4], the discovery of common gene expression patterns be-tween different diseases [5,6], or the search for inverse gene expression patterns be-tween different conditions [7].

An important step of a meta-analysis is to carry out prior quality control to reduce bias, check the homogeneity of data, detect unmeasured values, etc. This also helps to select the most appropriate method to be applied and avoid inaccurate results. The publication of inconsistent results and misinterpretations has been severely criticized in recent years [8,9]. Therefore, it is necessary to implement dedicated software that allows users to apply the different meta-analysis methods properly.

R packages have been previously developed for gene expression meta-analyses such as MetaIntegrator [10] or MetaVolcanoR [11]. Surprisingly, most of the available packages discard the genes not available in all the datasets included in the me-ta-analysis. Nevertheless, these missing genes may lead to the omission of relevant in-formation, losing relevant patterns, and this can cause different results to be obtained between studies [12]. In other contexts, a typical solution to deal with missing values is to impute the missing values from the samples with available data within the same study. However, this approach is not applicable to gene expression meta-analyses, since expression values are absent for all the samples. The use of models to impute these missing genes from the correlation with other non-missing genes has been pro-posed [12]. Furthermore, methods that perform imputation from the samples of other studies have also been proposed, obtaining fewer errors when comparing the imputed and real values [13]. Nevertheless, none of these methods have been previously implemented in gene expression meta-analysis packages.

The DExMA package has been implemented to perform all the steps of gene expression meta-analyses, providing two functions to treat missing genes across datasets. The first approach is based on imputing missing genes from the samples of other studies using the k-nearest neighbours (sampkeKNN method) [13]. The second approach consists of considering those genes with available values in a minimum proportion of datasets selected by the user in the meta-analysis.

Moreover, DExMA allows users to download data from the GEO database simply by using the corresponding codes. In addition, it includes quality control steps and heterogeneity testing. This package is available in the Bioconductor repository (http://bioconductor.org/packages/release/bioc/html/DExMA.html, accessed on 31 August 2022).

In this article, we describe the main functions and functionalities of the DExMA package. In the first section, the different implemented meta-analysis methods are de-scribed. Next, we present the workflow through a use case with simulated data, and in the last section, we present results from the analysis of real expression datasets.

2. Materials and Methods

2.1. Meta-analysis Methods

A gene expression meta-analysis encompasses a set of statistical methods that al-low us to combine results from different gene expression studies to obtain a single result with greater statistical power and sample size. The most suitable method depends on the nature and characteristics of the analysed datasets [14]. The DExMA package includes most of the methods from the two main meta-analysis approaches: effect size combination and p-value combination.

2.1.1. Effect Size Combination Methods

A meta-analysis based on effect size combination aims to explain the strength of a measure (effect) between different groups (e.g., experimental and control groups). In the specific case of gene expression studies, the effect to be calculated is the difference in standardized means between the expression level of the experimental group and the control group. This model has the following assumptions [15]:

- There is independence between the experimental and the control group.
- Both the experimental and control groups are distributed according to a normal distribution with means $\mu_E$and μ_C,, respectively, and with the same σ^2variance.

Therefore, the effect size of a gene in the *i-th* dataset (T_i) is described as:

$$T_i = \frac{\mu_E - \mu_C}{\sigma} \quad (1)$$

The DExMA package internally calculates *Hedges' g* as an estimator of the effect size, which is obtained [15]:

$$T_i = c(m) \times \frac{\bar{\bar{y}}_E - \bar{\bar{y}}_C}{S} \quad (2)$$

where:

- $c(m) = 1 - \frac{8}{4(n_E + n_C) - 9}$, is a factor that corrects the positive bias. n_E and n_C are the sample sizes of the experimental and control groups, respectively.
- $\bar{\bar{y}}_E$ and $\bar{\bar{y}}_C$ are the gene expression means of the experimental and control group, respectively.
- $S = \sqrt{\frac{(n_E-1)S_E^2+(n_C-1)S_C^2}{n_E+n_C-2}}$ is the standard deviation between studies. S_E^2 and $S_C^2$are the variances in the experimental and control groups, respectively.

Moreover, the within-study variance of this estimator is calculated:

$$V_i = \frac{n_E + n_C}{n_E \times n_C} + \frac{T_i^2}{2 \times (n_E + n_C)} \quad (3)$$

Once effect sizes and their variances have been calculated for each gene in the different studies, the *combined effect size* must be calculated to determine if a gene is differentially expressed.

To obtain the *combined effect size* and its corresponding p-value, DExMA provides the application of two models: the Fixed Effects Model (FEM) and the Random Effects Model (REM).

The FEM assumes that all studies share a true common effect size, that is to say, studies with more information have greater weight in the combined effect size. Therefore, the combined effect size ($\bar{T}$) and its variance (V) for k studies are calculated [15]:

$$\bar{T} = \frac{\sum_{i=1}^{k} \omega_i T_i}{\sum_{i=1}^{k} \omega_i} \quad (4)$$

$$V = \frac{1}{\sum_{i=1}^{k} \omega_i} \quad (5)$$

Where:

- T_i is the effect size of the *i*-th study.
- ω_i is the weight assigned to the *i-th* study. In the case of a meta-analysis, the inverse of the variance is used as weights, $\omega_i = \frac{1}{V_i}$.

Since the FEM model assumes the existence of normality, the *z-value* (z) of the combined effect size for a standard normal:

$$z = \frac{\bar{T}}{\sqrt{V}} \quad (6)$$

This *z-value* is used to calculate the p-value. Furthermore, in the specific case of a gene expression meta-analysis, this z-value is used to determine if the gene is over-expressed $(z > 0)$or under-expressed $z < 0$).

The **REM** model considers that the true effect size varies from one study to another, that is, there is a distribution of the true effect sizes. The combined effect size *($\bar{T}^*$)* and its variance (V^*) are calculated:

$$\bar{T}^* = \frac{\sum_{i=1}^{k} \omega_i^* T_i}{\sum_{i=1}^{k} \omega_i^*} \quad (7)$$

$$V^* = \frac{1}{\sum_{i=1}^{k} \omega_i^*} \quad (8)$$

In this case, the calculation of the weights differs from the FEM, since it influences both the within-study variance (V_i) and between-study variance (τ^2) [15]. The between-study variance is obtained:

$$\tau^2 = \begin{cases} \frac{Q - df}{C}, & Q > df \\ 0, & Q \leq df \end{cases} \quad (9)$$

Where

- $Q = \sum_{i=1}^{k} \omega_i (T_i - T_.)$ represents the total variance, where:
- ω_i is the calculated weight for the Fixed Effects Model.
- $T_.$ is the combined effect size for the Fixed Effects Model (Equation (2)).
- $C = \sum_{i=1}^{k} \omega_i - \frac{\sum_{i=1}^{k} \omega_i^2}{\sum_{i=1}^{k} \omega_i}$is a scaling-related factor related to the fact that Q is a weighted sum of squares.

- $df = k - 1$ are the degrees of freedom for the meta-analysis.

Therefore, the weights for the REM are calculated:

$$\omega_i^* = \frac{1}{V_i + \tau^2} \tag{10}$$

As in the FEM, the *z-value* of the *combined effect size* for a standard normal is calculated:

$$z = \frac{\bar{T}^*}{\sqrt{V^*}} \tag{11}$$

As for the FEM, this *z-value* is used to determine if the gene is over-expressed ($z > 0$) or under-expressed ($z < 0$).

2.1.1. p-values Combination methods

A meta-analysis based on p-values combination methods aims to merge all the p-values from different hypothesis tests into a single *p-value*. *p-value* combination methods have the following assumptions [16]:

- $p_1, \dots, p_k$ are the p-values from the *k* independent studies.
- The $t_1, \dots, t_k$ test statistics have absolute continuous probability distributions under their corresponding null hypotheses.

In the specific case of a gene expression meta-analysis, it seeks to obtain a combined p-value for each of the genes. The *p-values* are obtained from performing a differential expression analysis for each of the datasets. The DExMA package internally uses the *limma* Bioconductor package [17] in order to obtain the individual p-values. Afterward, to merge the individual p-values, DExMA implements five different p-value combination methods: Fisher's method, Stouffer's method, Tippett's method, Wilkinson's method, and the Aggregated Cauchy Association Test method (ACAT).

Fisher's method calculates a statistic (S_F) as the sum of the logarithm of the *p-values*, $S_F = -2 \times \sum_{i=1}^{k} \ln(p_i)$ [18]. Under the null hypothesis, S_F is distributed as x^2 with $2 \times k$ degrees of freedom [16].

Stouffer's method assumes that $Z_i = \phi^{-1}(1 - p_i)$ [16], where ϕ is the standard normal cumulative distribution function. Then, for *k* independent studies, the statistic is calculated as the sum of the Z_i values divided by the square root of the number of studies, $S_S = \frac{\sum_{i=1}^{k} Z_i}{\sqrt{k}}$. Under the null hypothesis, S_S is distributed as a standard normal distribution [16]. Moreover, Stouffer's method allows the inclusion of each of the datasets, $\frac{\sum_{i=1}^{k} \omega_i Z_i}{\sqrt{\sum_{i=1}^{k} \omega_i^2}}$. The DExMA package implements the square roots of sample sizes as weights [19].

Tippett's method (also called the minimum of p-values method) and **Wilkinson's method** (also called the maximum of p-values method) use the minimum of *p-values* and the maximum of *p-values*, respectively, as statistics, that is to say, $S_T = min(p_1, p_2, \dots, p_i, \dots, p_k)$ and $S_W = max(p_1, p_2, \dots, p_i, \dots, p_k)$. Under the null hypothesis, S_T is distributed as a *Beta(1,K)*, while S_W is distributed as a *Beta(K,1)*.

Finally, the **ACAT method** uses a weighted sum of the Cauchy transformation of individual *p-values*, $S_{ACAT} = \sum_{i=1}^{k} \omega_i \tan[(0.5 - p_i)\pi]$, as a statistic, where the weights

ω_i are non-negative and $\sum_{i=1}^{k} \omega_i = 1$. Under the null hypothesis, S_{ACAT} is distributed as a standard Cauchy distribution [20,21].

2.2. Control of Missing Genes

DExMA contains two different approaches to control the possible existence of missing genes: (i) the selection of the minimum number of datasets in which a gene must appear and (ii) missing genes imputation.

The first approach consists of performing a meta-analysis by only considering those genes contained in a minimum number (or proportion) of datasets. For example, if a gene is in 2 of 4 datasets and the user-defined threshold is that the gene should be contained in 75% of the studies, this gene will be discarded. In the final results, a vari-able is shown with the proportion of studies in which the gene is contained to help us-ers to correctly interpret the results obtained.

The second approach applies the *sampleKNN* method described by *Mancuso et al*. [13]. This method imputes the gene expression of a gene by applying the KNN imputa-tion in the space of samples. Firstly, to impute the expression value of missing genes, the k samples of datasets without missing genes and with the most similar expression are chosen. Then, the gene expression of these missing genes is imputed by calculating the weighted average of the expression in the k selected samples.

3. Results

3.1. The DExMA package

The DExMA package includes the main methods for gene expression meta-analyses described previously. The DExMA workflow consists of five main steps (Figure 1): meta-analysis object creation, gene annotation, quality control, gene expression meta-analysis, and visualization. The DExMA package provides a set of functions that provide additional information. Table 1 contains a summary of all available functions.

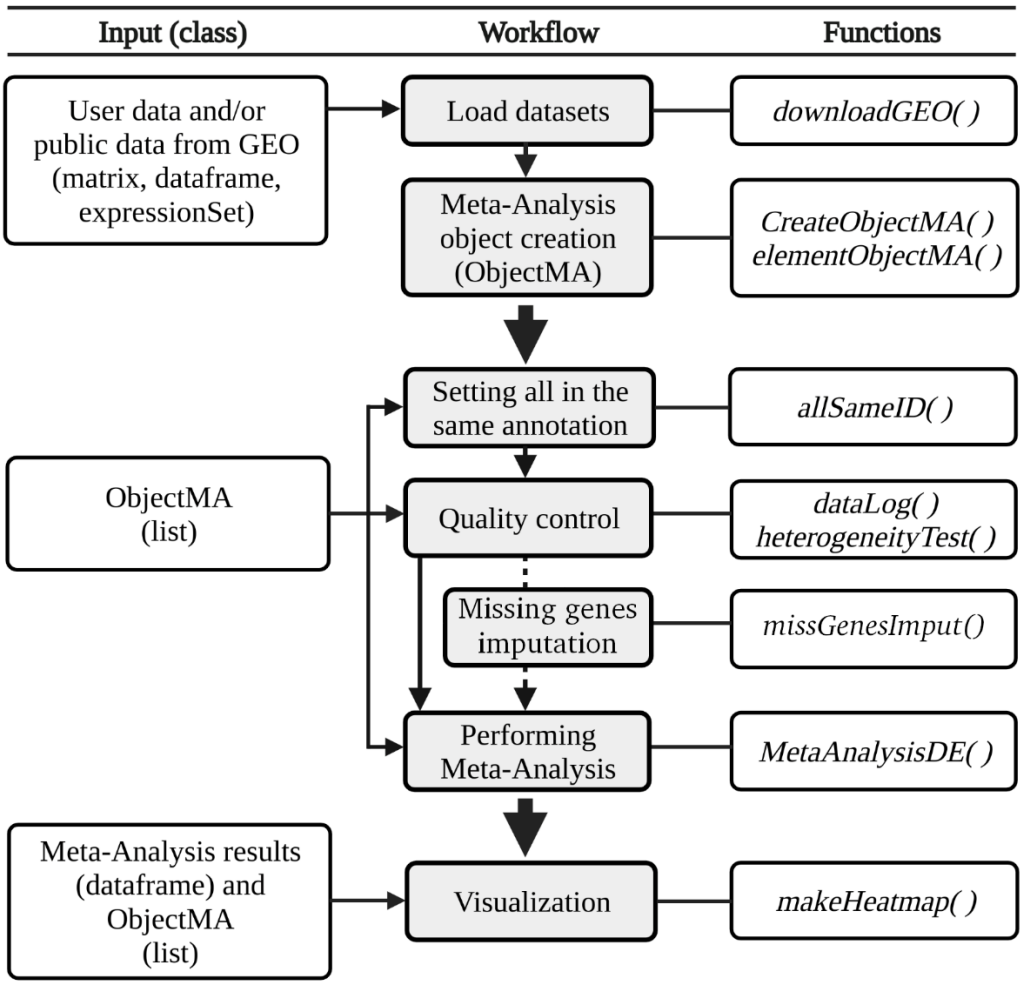


Figure 1. DExMA workflow. The figure shows the main steps of the DExMA package workflow: (1) data load and meta-analysis object creation, (2) gene annotation, (3) quality control, (4) miss-ing gene imputation (optional), (5) gene expression meta-analysis, and (6) visualization.

Table 1. Functions implemented in DExMA. Brief description of the functions developed in DExMA.

Function	Description
allsameID	Sets all datasets of objectMA in the same annotation (Official Gene Symbol, Entrez, or Ensembl)
batchRemove	Reduces the effects of batch or bias through the use of covariates
calculateES	Calculates the effects sizes and their variances for each gene and each dataset using Hedges' g estimator
createObjectMA	Creates the meta-analysis object (*objectMA*)
dataLog	Checks if data are log transformed and transforms them if they are not
downloadGEOData	Downloads ExpressionSets objects from GEO database
elementObjectMA	Creates an object that can be added to a meta-analysis object (*objectMA*)
heterogeneityTest	Shows a QQ-plot of Cochran's test and the quantiles of I^2statistic values to measure heterogeneity
makeHeatmap	Shows a heatmap with the expression of significant genes along samples
metaAnalysisDE	Performs a meta-analysis using the selected method
pvalueIndAnalysis	Performs a differential expression analysis in each of the studies to obtain the p-values
missGenesImput	Imputes missing genes using the *sampleKNN* method

In this section, the main steps to perform the gene expression meta-analysis are described. For this purpose, simulated gene expression data contained in the package itself, called

DExMAExampleData, were analysed. The data DExMAExampleData contain six different objects:

- "*listMatrixEX*": a list of four expression matrices.
- "*listPhenodatas*": a list of the four phenodata dataframes corresponding to four expression matrices.
- "*listExpressionSets*": a list of four *ExpressionSet* objects. It contains the same information as *listMatrixEX* and *listPhenodatas*.
- "*ExpressionSetStudy5*": an ExpressionSet object similar to the *ExpressionSets* objects of listExpressionSets.
- "*maObjectDif*": the meta-analysis object (*objectMA*) created from the listMatrixEx and listPhenodatas objects.
- "*maObject*": the meta-analysis object (*objectMA*) after setting all the studies in Official Gene Symbol annotation.
- Specifically, the *listMatrixEX* and *listPhenodatas* objects are used in the examples.

3.1.1. Meta-Analysis Object Creation

The first step in the analysis is the data entry. To this end, DExMA uses an *objectMA* object which is a list of nested lists where each one contains two elements: a gene expression matrix (with genes in rows and samples in columns) and a vector of 0 and 1 that indicates the group to which each sample belongs (0 represents the control group and 1 represents the experimental group).

DExMA provides the function *createObjectMA()* to facilitate the *objectMA* creation (details are provided in the package documentation).

When datasets with different gene names are used, it is necessary to convert them to a common gene identifier (*ID*). DExMA provides the *allSameID()* function, which allows us to translate genes to a common ID. Supported Gene IDs are official Gen Symbol or standard *IDs* from Entrez or Ensembl databases.

3.1.2. Quality Control

Quality control is a crucial step to conduct a proper meta-analysis and avoid misinterpretation. DExMA implements standard pre-processing steps in gene expression data analysis, such as data normalization and the analysis of heterogeneity [14].

Specifically, the *datalog()* function can be used to check if the data are in log scale or to perform log transformation, which is important when p-value combination methods are applied. To analyse data heterogeneity, DExMA provides the *heterogeneityTest()* function that implements two ways of measuring heterogeneity.

On the one hand, it returns a QQ-plot of Cochran's heterogeneity test (Figure 2) [22]. In the case of homogeneity, it is expected that the majority of the values will be close to the expected distribution (the central line of the graph).

In the case of homogeneity, it is expected that the majority of the values will be close to the expected distribution (the central line of the graph). On the contrary, if these points are distant from the central line, this is an indicator of heterogeneity.

On the other hand, the *heterogeneityTest()* function returns the quantiles of the I^2 statistic. The I^2 statistic measures the inconsistency, that is, the percentage of variation across studies due to heterogeneity [23]. When interpreting the I^2 results, it is considered that there is low heterogeneity when the I^2 value is less than 0.25 [23]. Therefore, to consider homogeneity, most of the I^2 values must be less than 0.25.

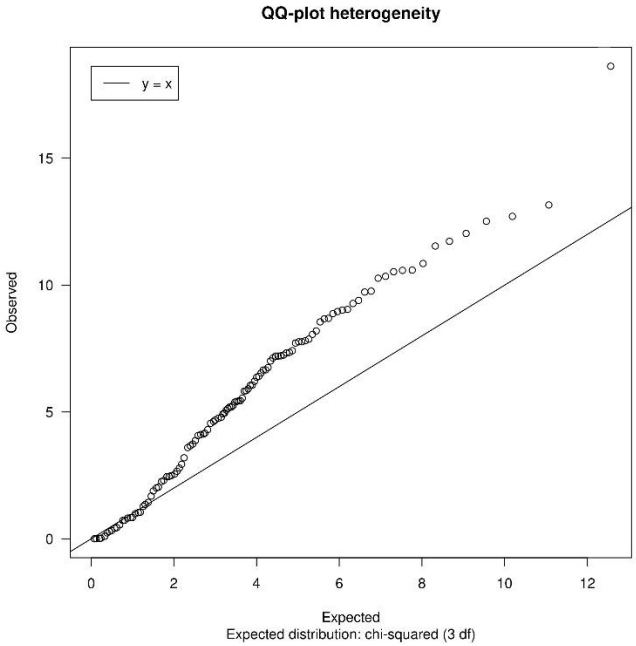


Figure 2. Heterogeneity QQ-plot. QQ-plot of Cochran's heterogeneity test values. The further the values are from the reference distribution (central line), the more heterogeneity there is.

3.1.3. Missing Gene Imputation

DExMA allows users to impute the expression of missing genes with the *missGenesImput()* function, which imputes the unmeasured expression using the k-nearest neighbours (KNN) in the space of samples (*sampleKNN method*). The function returns the objectMA with all the imputed studies.

Moreover, the *missGenesImput()* function returns an object (*imputIndicators*) with different indicators of the imputation. This item contains:

- *imputValuesSample*: the number of missing values imputed per sample.
- *imputPercentageSample*: the percentage of missing values imputed per sample.
- *imputValuesGene*: the number of missing values imputed per gene.
- *imputPercentageGene*: the percentage of missing values imputed per gene.

3.1.4. Performing Gene Expression Meta-Analysis

As it has been explained before, the main objective of the DExMA package is to perform the gene expression meta-analysis of several studies. For this purpose, DExMA includes the *metaAnalysisDE()* function. This function allows users to apply seven dif-ferent techniques of meta-analysis described in the methods sections: the Fixed Effect Model (FEM); the Random Effects Model (REM); Fisher's *p-value* combination method (Fisher); Stouffer's *p-value* combination method (Stouffer); Wilkinson's p-value com-

bination method (maxP); Tippett's *p-value* combination method (minP); and the Aggregated Cauchy Association Test method (ACAT).

If data imputation has not been previously applied, this function provides the option of considering genes that are in a minimum number of datasets. For example, if we have four datasets and we select that a gene must be in 75% of the datasets, those genes that are present in three or four datasets will be included in the meta-analysis. This allows users to control missing genes that are only present in a low number of datasets. Once the meta-analysis is complete, the function returns a table with the obtained results for both effect sizes and p-values based on the meta-analysis (see the package documentation for more details).

The results are also provided as heatmaps of the significant differentially ex-pressed genes (see Figure 3). DExMA provides the *makeHeatmap()* function for that purpose, which implements four types of scaling options:

- "*rscales*": this applies *rescale* function of the *scales* package [24]. Therefore, values will be between -1 and 1.
- "*zscor*": this calculates a z-score value for each gene and sample.
- "*swr*": this scales relative to a reference dataset approach [25].
- "*none*": no scaling approach is performed.

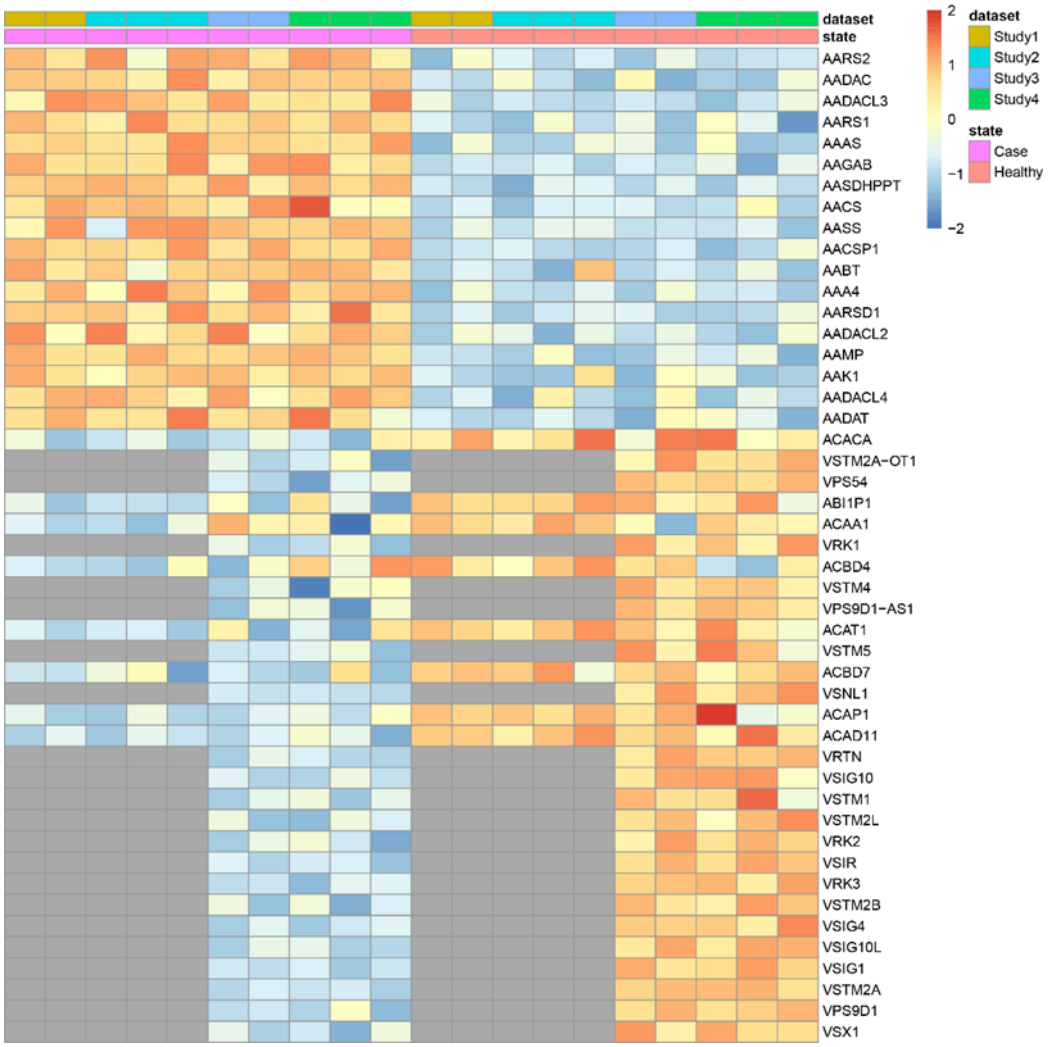


Figure 3. Synthetic data heatmap. Heatmap of the meta-analysis results for the 40 most significant genes. The red colour indicates that a gene is overexpressed in that sample, blue that it is un-der-expressed, and grey that is not present.

3.1.5. Other Useful Functions

DExMA provides some functions that allow users to speed up the analysis, correct the batch effect, or complete the results. With regard to accelerating the meta-analysis process, the function *downloadGEOData()* allows users to download multiple ExpressionSets objects from the GEO database [1]. In addition, the function *elementObjectMA()* can be used to create an element which can be added directly to a previously created *objectMA*, which avoids the user having to re-create the object from scratch. Regarding the batch effect correction, DExMA contains the function *batchremove*. The *batchRemove* function eliminates the effects of different covariates in the data variability. Finally, the functions *calculateES()* and *pvalueIndAnalysis()* return the effect sizes or the individual p-values of each study, respectively. This can help the user to better understand the results obtained.

3.2. Applying DExMA to Real Data

To illustrate the benefits of the DexMA package, it was applied to three real datasets. These data belong to systemic lupus erythematosus (SLE) gene expression studies, and they were extracted from the ADEX database [26]. Specifically, the identifiers of the selected studies were: *GSE24706* [27], *GSE50772* [28], and *GSE82221_GPL10558* [29]. These studies were chosen because of their samples were generated from the same cell tissue, peripheral blood mononuclear cells (PBMCs). In this way, a greater homogeneity between datasets was ensured than if they were extracted from different cell types. The code used for the data preparation for the use case is available in Appendix A.

In this case, it was not necessary to apply the *allSameID()* function, since all the datasets are annotated in the Official Gene Symbol. In the study of heterogeneity, a QQ-plot (Figure 4) was obtained in which most of the points were quite far from the reference line. In addition, 25% of the genes had an I^2 greater than 0.71, so it was concluded that there was heterogeneity between the different datasets. Therefore, as all the studies belonged to the same tissue, and there was heterogeneity between them, we decided to apply a Random Effects Model (REM).

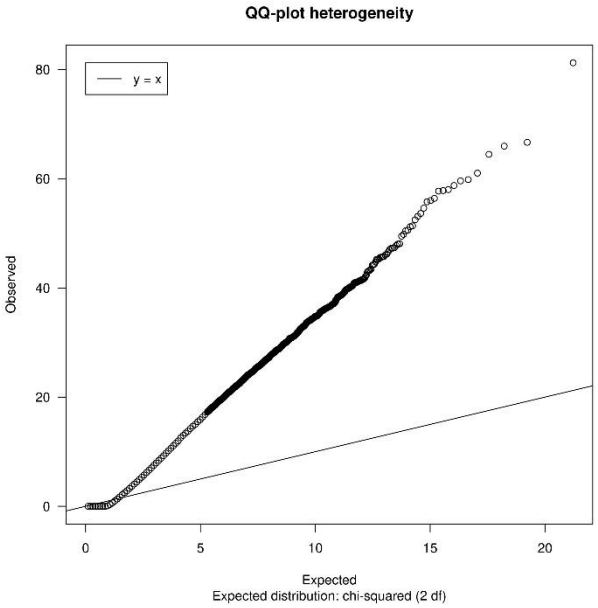


Figure 4. Heterogeneity QQ-plot of SLE data. QQ-plot of Cochran's heterogeneity test values for the SLE case study data.

To demonstrate the usefulness of the package, the meta-analysis was applied from three different approaches:

1. Using only common genes (*common genes approach*).
2. Considering the genes that are present in at least two of the studies (66%) (*minimum proportion approach*).
3. Performing a previous imputation of missing genes before accomplishing the meta-analysis (called the *imputing missing genes approach*).

The meta-analysis of only common genes took into account 11,298 genes, which only represented 49.5% of the total available genes, of which 1896 were found to be significant (16.8% of genes considered) (adjusted p-value (FDR) < 0.05). The *minimum proportion approach* worked with 14,548 genes, which represented 63.8% of the total available genes, of which 2444 were found to be significant (16.8% of genes considered). Finally, the meta-analysis of imputed missing genes considered all available genes, 22,807 genes (22,807), of which 4830 were found to be significant (21.1% of genes considered).

These results suggest that if only common genes were considered, an important part of the information would be lost (in this use case, more than 50% of the available genes would not influence the final result). Moreover, to verify this loss of information, the heatmap of the 50 most significant genes obtained by the *minimum proportion approach* was generated (Figure 5). This heatmap revealed that several of the most significant genes would have disappeared from the final result if the *common genes approach* was applied (missing values are marked in grey).

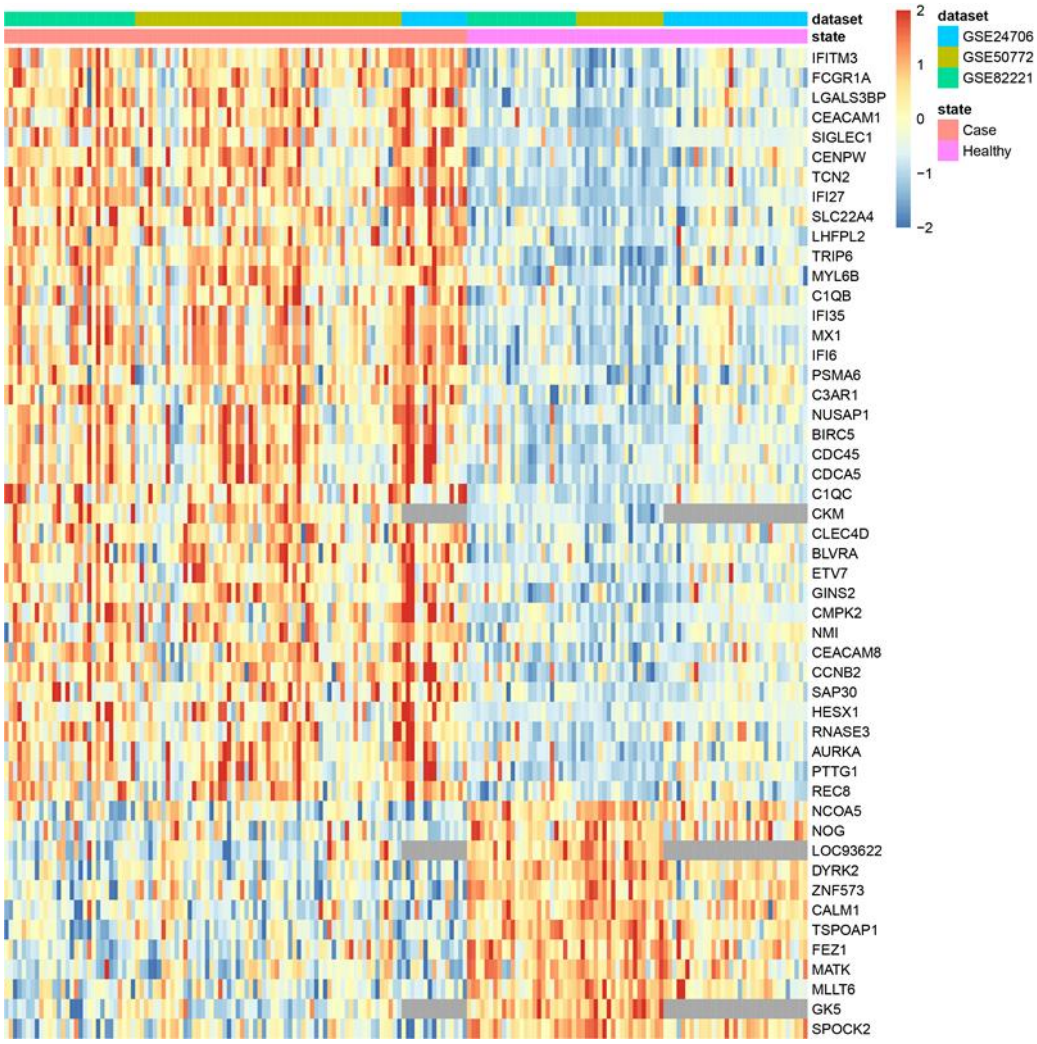


Figure 5. Heatmap of the 50 most significant genes in SLE data without imputation.

Finally, a functional enrichment analysis of the over-expressed genes was per-formed using GeneCodis4 [30,31] to validate the biological significance of the obtained results. The *systemic lupus erythematosus* pathway did not appear among the 10 most enriched pathways using the Bioplanet 2019 database [32] when the *common genes approach* and the *minimum proportion approach* were applied (Figure 6). Nevertheless, when the missing genes imputation was applied, the *systemic lupus erythematosus* biological pathway became the most significant pathway. Moreover, genes belonging to this pathway were recovered if the imputation of the missing genes was applied before the *common genes approach* and the *minimum proportion approach*.

These results highlight the impact of discarding missing genes in a gene expression meta-analysis, which can bias the final results.

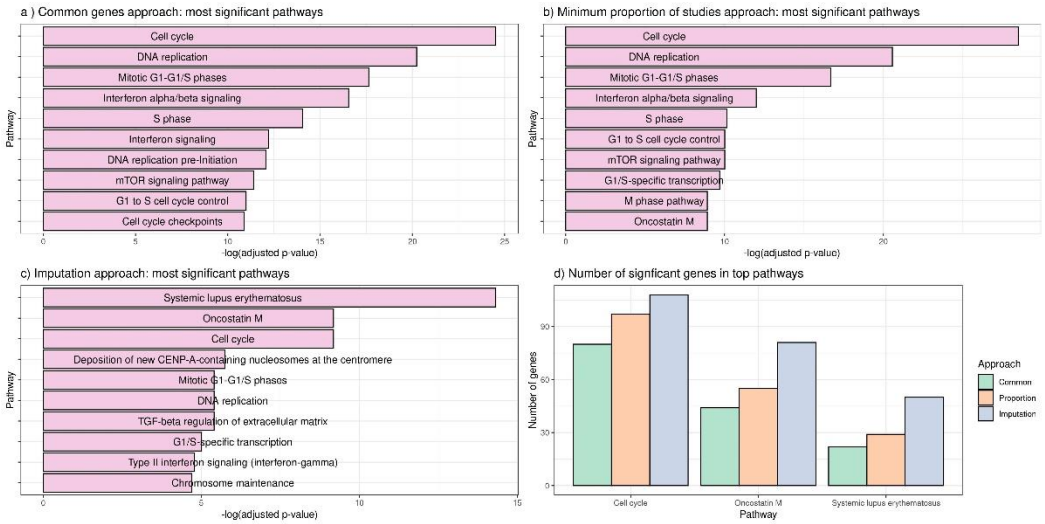


Figure 6. Graphical representations of the most significant pathways for each of the meta-analysis approaches. (a) Ten most significant pathways in common genes approach. (b) Ten most significant pathways considering genes that are contained in at least two studies. (c) Ten most significant pathways in the missing genes imputation approach. (d) Number of significant genes in top pathways in each approach.

3.3. Comparison to Other Available R Package

Currently, apart from DExMA, there are eight R packages available in CRAN or Bioconductor repositories that allow one to perform gene expression meta-analyses: *metahdep* [33], *GeneMeta* [34], *metaRNASeq* [35], *metaSeq* [36], *metaMA* [37], *crossmeta* [38], *MetaIntegrator* [10], and *MetaVolcanoR* [11]. Table 2 shows a summary of the main features of the packages currently available for performing gene expression me-ta-analyses.

Table 2. Comparison of the main features of gene expression meta-analysis packages. *Input*: "User data" means that the user can enter their own data, while "GEO data" means that the user can include GEO database codes. *QC* (quality control): "Yes" if the package has implemented func-tions for performing quality controls. *ES*: "Yes" if the package performs effect sizes combination methods. *PV*: "Yes" if the package performs p-value combination methods. *Considers Missing Genes*: "Yes" if the package somehow considers the unmeasured genes. *Imputes Missing genes*: "Yes" if the package somehow imputes the unmeasured genes. *Visualization*: "Yes" if the package has implemented a function to visualize the results.

Package	Input	QC	ES	PV	Considers Missing Genes	Imputes Missing Genes	Visualization
DExMA	GEO/User data	Yes	Yes	Yes	Yes	Yes	Yes
MetaIntegrator [10]	User data	Yes	Yes	Yes	Yes	No	Yes
GeneMeta [34]	User data	No	Yes	No	No	No	Yes
Metahdep [33]	User data	No	Yes	No	No	No	No
Crossmeta [38]	User data	No	Yes	No	Yes	No	No
metaMA [37]	User data	No	Yes	Yes	No	No	No
metaRNASeq [35]	User data	No	No	No	No	No	Yes
metaSeq [36]	User data	No	No	No	No	No	No
MetaVolcanoR [11]	User data	No	Yes	Yes	Yes	No	Yes

Most implemented packages usually use previously curated data by the user as input, while crossmeta allows the use of data downloaded from the GEO database. DExMA has the advantage that it admits users to work with both user data as well as with GEO-

downloaded datasets. Furthermore, it provides a function that facilitates the creation of the object needed to perform the meta-analysis from the information entered by the user.

Regarding the different steps of a gene expression meta-analysis, DExMA, unlike the rest of the packages, contains functions to perform quality control before the me-ta-analysis and help in the decision of which meta-analysis method to apply. Several packages mention the importance of these previous steps; only the *MetaIntegrator* package implements a function related to quality control, but it does not include any-thing about the heterogeneity of the studies.

Moreover, as previously referenced, most of these packages only perform the analyses with the genes common to all datasets. Only *crossmeta*, *MetaVolcvanoR*, and *MetaInetgrator* consider the possible existence of missing genes but do not make any imputation of them, nor do they show their possible effect on the final result.

4. Discussion

The accumulation and availability of experimental data in public repositories has fuelled the development of meta-analysis techniques as important tools to integrate heterogeneous datasets. In the field of transcriptomics, these techniques have been applied to jointly analyse gene expression for biomarker discovery or drug-repurposing applications, among others. The number of scientific publications with meta-analysis studies is growing exponentially, and as have been reported [8,9], a high proportion of these published analyses are misleading meta-analyses or have serious methodological flaws. In this context, it is important for the scientific community that software pack-ages that implement proper statistical methods and dedicated workflows are available.

This article introduces DExMA, an R package that implements the main steps and methods for gene expression meta-analyses. Moreover, to avoid the loss of information due to the use of only common genes, DExMA allows users to deal with missing genes with two approaches: selecting the proportion of datasets that must contain a gene or imputing the missing genes by using the KNN imputation method in the space of samples (*sampleKNN*). To the best of our knowledge, DExMA is the first gene expression meta-analysis package that controls missing genes. Although there are other packages that also consider the possible existence of unmeasured genes (*crossmeta*, *MetaIntegrator*, and *MetaVolcanoR*), none of them perform the imputation of these missing genes, nor do they show the possible effect of this lack of information in the results.

DExMA also offers the possibility of using both GEO codes and one's own data as well as performing the different steps of a gene expression meta-analysis (homogenizing gene annotation, quality control, and results visualization), which contributes to making appropriate use of these methods.

Supplementary Materials

The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/math10183376/s1.

Author Contributions

Conceptualization, J.A.V.-G., J.M.-M. and P.C.-S.; methodology, J.A.V.-G. and D.T.-D.; software, J.A.V.-G., J.M.-M. and D.T.-D.; validation, J.A.V.-G. and J.M.-M.; formal analysis, J.A.V.-G.; investigation, J.A.V.-G.; resources, Y.R.-M. and P.F.; data curation, J.A.V.-G. and J.M.-M.; writing—original draft preparation, J.A.V.-G.; writing—review and editing, J.A.V.-G., J.M.-M., D.T.-D., Y.R.-M., P.F. and P.C.-S.; visualization, J.A.V.-G. and D.T.-D.; supervision, P.C.-S.; project administration, P.C.-S.; funding acquisition, P.C.-S. All authors have read and agreed to the published version of the manuscript.

Funding

This research has been funded by the Teaching Staff Programme, implemented by the Ministerio de Universidades (grant number FPU19/01999). This work is funded by grants PID2020-119032RB-I00, MCIN/AEI/10.13039/501100011033, and FEDER/Junta de Andalucía-Consejería de Transformación Económica, Industria, Conocimiento y Universidades (Grants P20_00335 and B-CTS-40-UGR20). Toro-Domínguez. is supported through the aid granted of the 'Consejería de Transformación Económica, Industria, Conocimiento y Universidades' (CTEICU), in the 2020 call, being co-financed by the European Union through the European Social Fund (ESF) named 'Andalucía se mueve con Europa", within the framework of the Andalusian ESF Operational Program 2014–2020. Martorell-Marugán is funded by European Union—NextGenerationEU, Ministerio de Universidades (Spain's Government) and Recovery, Transformation and Resilience Plan, through a call from the University of Granada.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in the "SLE_Data.RData" supplementary file.

Acknowledgments : This work is part of Juan Antonio Villatoro-García's Ph.D. results. Juan Antonio Villatoro-García is enrolled in the Mathematical and Applied Statistics Ph.D. program.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

KNN	K-nearest neighbours
FEM	Fixed Effects Model
REM	Random Effects Model
ID	Identifier
SLE	systemic lupus erythematosus

Appendix A

Loading and Preparing the Case Study Data Directly from the ADEX Data-base

The data used in the case study are available in the ADEx database [26] (https://adex.genyo.es/ , accessed on 15 August 2022). Specifically, we downloaded the following datasets: GSE24706, GSE50772, and GSE82221_GLP10558. Once the studies were downloaded, four files were obtained:

- GSE24706.tsv: gene expression matrix of the study GSE24706.
- GSE50772.tsv: gene expression matrix of the study GSE50772.
- GSE82221_GPL10558.tsv: gene expression matrix of the study GSE82221.
- metadata.tsv: dataframe with the information from the different samples of the studies (phenodata).

We loaded these files and prepared them for the use case:

```
R> #Loading gene expression matrix
R> GSE24706Ex <- as.matrix(read.delim("GSE24706.tsv", header = TRUE,
+       row.names = 1))
R> GSE50772Ex <- as.matrix(read.delim("GSE50772.tsv", header = TRUE,
+       row.names = 1))
R> GSE82221Ex <- as.matrix(read.delim("GSE82221_GPL10558.tsv",
+       header = TRUE, row.names = 1))
R> #Preparing studies phenodatas
R> Pheno <- read.delim("metadata.tsv", header = T, row.names = 1)
R> GSE24706Pheno <- Pheno[colnames(GSE24706Ex),]
R> GSE50772Pheno <- Pheno[colnames(GSE50772Ex),]
R> GSE82221Pheno <- Pheno[colnames(GSE82221Ex),]
```

Once the expression matrices were loaded and the phenodata were obtained for each of the studies, the data were ready for the case study.

References

1. Barrett, T.; Wilhite, S.E.; Ledoux, P.; Evangelista, C.; Kim, I.F.; Tomashevsky, M.; Marshall, K.A.; Phillippy, K.H.; Sherman, P.M.; Holko, M.; et al. NCBI GEO: Archive for Functional Genomics Data Sets—Update. Nucleic Acids Res. 2013, 41, D991–D995. https://doi.org/10.1093/nar/gks1193.

2. Perez-Riverol, Y.; Zorin, A.; Dass, G.; Vu, M.-T.; Xu, P.; Glont, M.; Vizcaíno, J.A.; Jarnuczak, A.F.; Petryszak, R.; Ping, P.; et al. Quantifying the Impact of Public Omics Data. Nat. Commun. 2019, 10, 3512. https://doi.org/10.1038/s41467-019-11461-w.

3. Song, G.G.; Kim, J.-H.; Seo, Y.H.; Choi, S.J.; Ji, J.D.; Lee, Y.H. Meta-Analysis of Differentially Expressed Genes in Primary Sjogren's Syndrome by Using Microarray. Hum. Immunol. 2014, 75, 98–104. https://doi.org/10.1016/j.humimm.2013.09.012.

4. Afroz, S.; Giddaluru, J.; Vishwakarma, S.; Naz, S.; Khan, A.A.; Khan, N. A Comprehensive Gene Expression Meta-Analysis Identifies Novel Immune Signatures in Rheumatoid Arthritis Patients. Front. Immunol. 2017, 8, 74. https://doi.org/10.3389/fimmu.2017.00074.

5. Badr, M.T.; Häcker, G. Gene Expression Profiling Meta-Analysis Reveals Novel Gene Signatures and Pathways Shared between Tuberculosis and Rheumatoid Arthritis. PLoS ONE 2019, 14, e0213470. https://doi.org/10.1371/journal.pone.0213470.

6. Kelly, J.; Moyeed, R.; Carroll, C.; Albani, D.; Li, X. Gene Expression Meta-Analysis of Parkinson's Disease and Its Relationship with Alzheimer's Disease. Mol. Brain 2019, 12, 16. https://doi.org/10.1186/s13041-019-0436-5.

7. Ibáñez, K.; Boullosa, C.; Tabarés-Seisdedos, R.; Baudot, A.; Valencia, A. Molecular Evidence for the Inverse Comorbidity between Central Nervous System Disorders and Cancers Detected by Transcriptomic Meta-Analyses. PLoS Genet. 2014, 10, e1004173. https://doi.org/10.1371/journal.pgen.1004173.

8. Ioannidis, J.P.A. The Mass Production of Redundant, Misleading, and Conflicted Systematic Reviews and Meta-Analyses. Milbank Q. 2016, 94, 485–514. https://doi.org/10.1111/1468-0009.12210.

9. Park, J.H.; Eisenhut, M.; van der Vliet, H.J.; Shin, J.I. Statistical Controversies in Clinical Research: Overlap and Errors in the Meta-Analyses of MicroRNA Genetic Association Studies in Cancers. Ann. Oncol. 2017, 28, 1169–1182. https://doi.org/10.1093/annonc/mdx024.

10. Haynes, W.A.; Vallania, F.; Liu, C.; Bongen, E.; Tomczak, A.; Andres-Terrè, M.; Lofgren, S.; Tam, A.; Deisseroth, C.A.; Li, M.D.; et al. Empowering Multi-Cohort Gene Expression Analysis to Increase Reproducibility. Pac. Symp. Biocomput 2016, 22, 144–153.

11. Prada, C.; Lima, D.; Nakaya, H. MetaVolcanoR: Gene Expression Meta-Analysis Visualization Tool; 2022.; https://www.bioconductor.org/packages/release/bioc/html/MetaVolcanoR.html (accessed on 1 July 2022).

12. Bobak, C.A.; McDonnell, L.; Nemesure, M.D.; Lin, J.; Hill, J.E. Assessment of Imputation Methods for Missing Gene Expression Data in Meta-Analysis of Distinct Cohorts of Tuberculosis Patients. Pac. Symp. Biocomput. 2020, 25, 307–318.

13. Mancuso, C.A.; Canfield, J.L.; Singla, D.; Krishnan, A. A Flexible, Interpretable, and Accurate Approach for Imputing the Expression of Unmeasured Genes. Nucleic Acids Res. 2020, 48, e125, https://doi.org/10.1093/nar/gkaa881.

14. Toro-Domínguez, D.; Villatoro-García, J.A.; Martorell-Marugán, J.; Román-Montoya, Y.; Alarcón-Riquelme, M.E.; Carmona-Sáez, P. A Survey of Gene Expression Meta-Analysis: Methods and Applications. Brief. Bioinform. 2021, 22, 1694–1705. https://doi.org/10.1093/bib/bbaa019.

15. Borenstein, M.; Hedges, L.V.; Higgins, J.P.T.; Rothstein, H.R. Introduction to Meta-Analysis; John Wiley & Sons: Hoboken, NJ, USA, 2021; ISBN 978-1-119-55838-5.

16. Heard, N.A.; Rubin-Delanchy, P. Choosing between Methods of Combining p-Values. Biometrika 2018, 105, 239–246. https://doi.org/10.1093/biomet/asx076.

17. Ritchie, M.E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C.W.; Shi, W.; Smyth, G.K. Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies. Nucleic Acids Res. 2015, 43, e47. https://doi.org/10.1093/nar/gkv007.

18. Li, J.; Tseng, G.C. An Adaptively Weighted Statistic for Detecting Differential Gene Expression When Combining Multiple Transcriptomic Studies. Ann. Appl. Stat. 2011, 5, 994–1019. https://doi.org/10.1214/10-AOAS393.

19. Zaykin, D.V. Optimally Weighted Z-Test Is a Powerful Method for Combining Probabilities in Meta-Analysis. J. Evol. Biol. 2011, 24, 1836–1841. https://doi.org/10.1111/j.1420-9101.2011.02297.x.

20. Liu, Y.; Chen, S.; Li, Z.; Morrison, A.C.; Boerwinkle, E.; Lin, X. ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. Am. J. Hum. Genet. 2019, 104, 410–421. https://doi.org/10.1016/j.ajhg.2019.01.002.

21. Liu, Y.; Xie, J. Cauchy Combination Test: A Powerful Test with Analytic p-Value Calculation under Arbitrary Dependency Structures. J. Am. Stat. Assoc. 2020, 115, 393–402. https://doi.org/10.1080/01621459.2018.1554485.

22. Higgins, J.P.T.; Thompson, S.G. Quantifying Heterogeneity in a Meta-Analysis. Stat. Med. 2002, 21, 1539–1558. https://doi.org/10.1002/sim.1186.

23. Higgins, J.P.T.; Thompson, S.G.; Deeks, J.J.; Altman, D.G. Measuring Inconsistency in Meta-Analyses. BMJ 2003, 327, 557–560. https://doi.org/10.1136/bmj.327.7414.557.

24. Wickham, H.; Seidel, D. RStu dio Scales: Scale Functions for Visualization; 2020. ; https://cran.r-project.org/web/packages/scales/index.html (accessed on 30 June 2022).

25. Lazar, C.; Meganck, S.; Taminau, J.; Steenhoff, D.; Coletta, A.; Molter, C.; Weiss-Solís, D.Y.; Duque, R.; Bersini, H.; Nowé, A. Batch Effect Removal Methods for Microarray Gene Expression Data Integration: A Survey. Brief. Bioinform. 2013, 14, 469–490. https://doi.org/10.1093/bib/bbs037.

26. Martorell-Marugán, J.; López-Domínguez, R.; García-Moreno, A.; Toro-Domínguez, D.; Villatoro-García, J.A.; Barturen, G.; Martín-Gómez, A.; Troule, K.; Gómez-López, G.; Al-Shahrour, F.; et al. A Comprehensive Database for Integrated Analysis of Omics Data in Autoimmune Diseases. BMC Bioinform. 2021, 22, 343. https://doi.org/10.1186/s12859-021-04268-4.

27. Li, Q.-Z.; Karp, D.R.; Quan, J.; Branch, V.K.; Zhou, J.; Lian, Y.; Chong, B.F.; Wakeland, E.K.; Olsen, N.J. Risk Factors for ANA Positivity in Healthy Persons. Arthritis Res. Ther. 2011, 13, R38. https://doi.org/10.1186/ar3271.

28. Kennedy, W.P.; Maciuca, R.; Wolslegel, K.; Tew, W.; Abbas, A.R.; Chaivorapol, C.; Morimoto, A.; McBride, J.M.; Brunetta, P.; Richardson, B.C.; et al. Association of the Interferon Signature Metric with Serological Disease Manifestations but Not Global Activity Scores in Multiple Cohorts of Patients with SLE. Lupus Sci. Med. 2015, 2, e000080. https://doi.org/10.1136/lupus-2014-000080.

29. Zhu, H.; Mi, W.; Luo, H.; Chen, T.; Liu, S.; Raman, I.; Zuo, X.; Li, Q.-Z. Whole-Genome Transcription and DNA Methylation Analysis of Peripheral Blood Mononuclear Cells Identified Aberrant Gene Regulation Pathways in Systemic Lupus Erythematosus. Arthritis Res. Ther. 2016, 18, 162. https://doi.org/10.1186/s13075-016-1050-x.

30. Carmona-Saez, P.; Chagoyen, M.; Tirado, F.; Carazo, J.M.; Pascual-Montano, A. GENECODIS: A Web-Based Tool for Finding Significant Concurrent Annotations in Gene Lists. Genome Biol. 2007, 8, R3. https://doi.org/10.1186/gb-2007-8-1-r3.

31. Garcia-Moreno, A.; López-Domínguez, R.; Villatoro-García, J.A.; Ramirez-Mena, A.; Aparicio-Puerta, E.; Hackenberg, M.; Pascual-Montano, A.; Carmona-Saez, P. Functional Enrichment Analysis of Regulatory Elements. Biomedicines 2022, 10, 590. https://doi.org/10.3390/biomedicines10030590.

32. Huang, R.; Grishagin, I.; Wang, Y.; Zhao, T.; Greene, J.; Obenauer, J.C.; Ngan, D.; Nguyen, D.-T.; Guha, R.; Jadhav, A.; et al. The NCATS BioPlanet—An Integrated Platform for Exploring the Universe of Cellular Signaling Pathways for Toxicology, Systems Biology, and Chemical Genomics. Front. Pharmacol. 2019, 10, 445. https://doi.org/10.3389/fphar.2019.00445.

33. Stevens, J.R.; Nicholas, G. Metahdep: Hierarchical Dependence in Meta-Analysis; 2022. ; https://www.bioconductor.org/packages/release/bioc/html/metahdep.html (accessed on 25 June 2022).

34. Lara Lusa <lusa at; Gentleman, R.; Ruschhaupt, M. GeneMeta: MetaAnalysis for High Throughput Experiments 2021. https://www.bioconductor.org/packages/release/bioc/html/GeneMeta.html (accessed on 25 June 2022).

35. Marot, G.; Rau, A.; Jaffrezic, F.; Blanck, S. MetaRNASeq: Meta-Analysis of RNA-Seq Data 2021.; https://cran.r-project.org/web/packages/metaRNASeq/index.html (accessed on 27 June 2022).

36. Tsuyuzaki, K.; Nikaido, I. MetaSeq: Meta-Analysis of RNA-Seq Count Data in Multiple Studies 2022.; https://www.bioconductor.org/packages/release/bioc/html/metaSeq.html (accessed on 27 June 2022).

37. Marot, G. MetaMA: Meta-Analysis for MicroArrays 2022;. https://cran.r-project.org/web/packages/metaMA/index.html (accessed on 27 June 2022).

38. Pickering, A. Crossmeta: Cross Platform Meta-Analysis of Microarray Data 2022;. https://www.bioconductor.org/packages/release/bioc/html/crossmeta.html (accessed on 27 June 2022).

9.3. DatAC: A visual analytics platform to explore climate and air quality indicators associated with the COVID-19 pandemic in Spain

Este artículo fue publicado el 1 de enero de 2021 en la revista *Science of the Total Environment* volumen 750, página 141424, DOI: https://doi.org/10.1016/j.scitotenv.2020.141424 bajo modelo de suscripción. Esta es la versión aceptada del artículo. De acuerdo con la editorial (Elsevier B.V.) esta versión del artículo tiene un permiso de reutilización e inclusión en una tesis doctoral siempre que no se publique comercialmente.

DatAC: A visual analytics platform to explore climate and air quality indicators associated with the COVID-19 pandemic in Spain

Jordi Martorell-Marugán [a,b,†], Juan Antonio Villatoro-García [a,†], Adrián García-Moreno [a], Raúl López-Domínguez [a], Francisco Requena [c], Juan Julián Merelo [d], Marina Lacasaña [e,f,g], Juan de Dios Luna [h], Juan J. Díaz-Mochón [a], Jose A. Lorente [a], Pedro Carmona-Sáez [a,h,*]

[a] Bioinformatics Unit. GENYO. Centre for Genomics and Oncological Research: Pfizer / University of Granada / Andalusian Regional Government, PTS Granada, 18016, Granada, Spain.
[b] Atrys Health S.A., Barcelona, Spain.
[c] Imagine Institute of Genetic Diseases, INSERM, 75015, Paris, France.
[d] Department of Computer Architecture and Technology, Universidad de Granada, 18071, Granada, Spain.
[e] Andalusian School of Public Health (EASP), 18011, Granada, Spain
[f] Ciber de Epidemiología y Salud Pública (CIBERESP), Spain
[g] Instituto de Investigación Biosanitaria ibs.GRANADA, Granada, Spain.
[h] Department of Statistics. University of Granada, 18071, Granada, Spain.

*Corresponding author (pedro.carmona@genyo.es)
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Highlights (Word file)

- DatAC integrates spatio-temporal data of weather, air quality and COVID-19.
- NO2, CO, SO2, PM2.5 and PM10 declined after lockdown, while O3 levels rose.
- The lockdown impact on rural air quality is smaller than in urban environments.
- Current data does not support climatic factors as driving factors of the pandemic.

Abstract

The coronavirus disease 2019 (COVID-19) pandemic has caused an unprecedented global health crisis, with several countries imposing lockdowns to control the coronavirus spread. Important research efforts are focused on evaluating the association of environmental factors with the survival and spread of the virus and different works have been published, with contradictory results in some cases. Data with spatial and temporal information is a key factor to get reliable results and, although there are some data repositories for monitoring the disease both globally and locally, an application that integrates and aggregates data from meteorological and air quality variables with COVID-19 information has not been described so far to the best of our knowledge.

Here, we present DatAC (Data Against COVID-19), a data fusion project with an interactive web frontend that integrates COVID-19 and environmental data in Spain. DatAC is provided with powerful data analysis and statistical capabilities that allow users to explore and analyze individual trends and associations among the provided data.

Using the application, we have evaluated the impact of the Spanish lockdown on the air quality, observing that NO_2, CO, $PM_{2.5}$, PM_{10} and SO_2 levels decreased drastically in the entire territory, while O_3 levels increased. We observed similar trends in urban and rural areas, although the impact has been more important in the former. Moreover, the application allowed us to analyze correlations among climate factors, such as ambient temperature, and the incidence of COVID-19 in Spain. Our results indicate that temperature is not the driving factor and without effective control actions, outbreaks will appear and warm weather will not substantially limit pandemic growth. DatAC is available at https://covid19.genyo.es.

Keywords

SARS-CoV-2, pollution, weather variables, RStudio Shiny framework, Spatio-temporal analysis

Abbreviations

CFR: Case fatality rate, **CO**: Carbon monoxide, **CRR**: Case recovery rate, **COVID-19**: Coronavirus disease 2019, **FDR**: False discovery rate, **ICU**: Intensive care unit, **NO_2**: Nitrogen dioxide, **O_3**: Ozone, **PCR**: Polymerase chain reaction, **$PM_{2.5}$**: Particulate matter 2.5 micrometers or less in diameter, **PM_{10}**: Particulate matter 10 micrometers or less in diameter, **SARS-CoV-2**: Severe acute respiratory syndrome coronavirus 2, **SD**: Standard deviation, **SO_2**: Sulfur dioxide, **VOC**: Volatile organic compound

1. Introduction

In December 2019, Coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was described in Wuhan, China (Guan et al., 2020). The spread of the disease has presented an extreme challenge to the international community and different countries have implemented different strategies depending on social, economic and political factors. Coronavirus infection spreads in clusters and some of the oldest and most effective containment measures such as social distancing, quarantine and isolation have been adopted to control the disease outbreak. In Spain, the Government declared the state of alarm and strict lockdown on March 14th, 2020. This lockdown was even stricter during the period March 30th to April 8th, when non-essential activities were totally prohibited. Between April 9th and May 3rd, the initial lockdown conditions were restored. Since May 4th, lockdown restrictions have been relaxed asymmetrically depending on the pandemic indicators of each region. On June 21st, the alarm finished along with the majority of the restrictions in most of the country.

Early detection of new cases and the identification of factors associated with the spread of SARS-CoV-2 are important aspects to control the pandemic. In this context, a research focus is on studying the role that environment factors play in the propagation of the disease. Previous studies have reported significant associations among air quality and meteorological variables and the amount or severity of new cases. In fact, Dr. Coccia

reported the poor air quality in the North of Italy as one of the factors for the quick diffusion of SARS-CoV-2 in this region (Coccia, 2020). Furthermore, particulate matter 2.5 micrometers or less in diameter ($PM_{2.5}$), particulate matter 10 micrometers or less in diameter (PM_{10}), sulfur dioxide (SO_2), nitrogen dioxide (NO_2), carbon monoxide (CO) or ozone (O_3) have been also associated with COVID-19 incidence (Bashir et al., 2020b; Fronza et al., 2020; Jiang et al., 2020; Ogen, 2020; Zhu et al., 2020), but it remains unclear if these correlations are actually related to causation (Riccò et al., 2020). Regarding weather conditions, there are several studies that have been published during recent months reporting negative correlation between temperature and COVID-19 cases (see for example Luo et al., 2020; Pequeno et al., 2020; Wang et al., 2020), humidity and death counts (Ma et al., 2020) or rainfall and new daily cases (Menebo, 2020). A recent study followed almost 7,000 hospitalized patients from Europe and China and reported that the increase in ambient temperature is linked to less severe symptoms (Kifer et al., 2020).

However, most of these studies are still preliminary and they are focused on specific regions. There are also some contradictory findings (e.g. a study of 122 chinese cities reported that temperature was positively correlated with cases (Xie and Zhu, 2020)), and the analyses are based on a short period of time, which generally covers the first peak of cases. More data is needed to derive more conclusive results and to elucidate the actual impact of ambient factors on COVID-19 pandemic.

A more robust evidence supports that lockdowns imposed by the governments in order to fight the SARS-CoV-2 spreading resulted in an improvement of air quality in major urban areas like Barcelona (Tobías et al., 2020), São Paulo (Nakada and Urban, 2020) or Northern China (Bao and Zhang, 2020), where SO_2, $PM_{2.5}$, PM_{10}, NO_2, and CO air concentration dropped. On the contrary, O_3 levels have increased significantly in these regions. This rise has been observed also in other Southern European cities (Rome, Valencia, Turín and Nice) as well as Wuhan (Sicard et al., 2020). Such phenomenon can be explained by a combination of factors, such as the higher volatile organic compounds (VOCs)-NO_x ratio, the reduction of O_3 titration due to the drop of NO_x or the reduction of $PM_{2.5}$ and PM_{10} (Sicard et al., 2020).

In this work we have processed and curated data from COVID-19 cases, meteorological and air quality data in Spain since January 1st, 2020. We have implemented a web based software, named DatAC (Data Against COVID-19), that integrates all these data and provides spatial-temporal aggregation of all these sources of information. The application includes visual analytics capabilities that allow users to explore temporal and regional evolution of variables as well as an easy interactive exploration of data relationships and associations. Using this application, we have analyzed the impact of the lockdown in Spain on the air quality in urban, suburban and rural areas. In addition, we have evaluated the relationship between meteorological variables and COVID-19 incidence in the entire Spanish territory.

We are confident that DatAC would be very useful to assess how all these variables are interacting and the actual impact of environmental factors on COVID-19 spread. DatAC has a free license and it is available at https://covid19.genyo.es.

2. Methods

2.1. Data collection

Daily COVID-19 total cases and cases diagnosed with polymerase chain reaction (PCR) from the autonomous communities and provinces have been obtained from the Ministry of Health of Spain (MISAN, 2020). Dates of these data refer to the onset of symptoms.

The number of daily deaths, cumulative hospitalized patients, cumulative patients translated to intensive care units (ICUs) and cumulative recovered patients from the autonomous communities were obtained from Datadista repository (Datadista, 2020). For provinces data, we obtained these variables from the Andalusian Institute of Statistic and Cartography (IECA, 2020) for the Andalusian provinces and from Escovid19data repository (Escovid19data, 2020) for the rest of the provinces.

Climatological information was downloaded from the Spanish State Meteorological Agency (AEMET, 2020). The daily data for each of the monitoring stations in Spain was downloaded. These data were processed to obtain the average daily temperature, rainfall, wind speed and hours of solar radiation for each province and community.

Air quality data from the different Andalusian monitoring stations were obtained from the Andalusian Office of Agriculture, Livestock, Fisheries and Sustainable Development (Junta de Andalucía, 2020). The air quality data from the rest of the Spanish monitoring stations were downloaded from the European Air Quality Portal (European Environment Agency, 2020). All this information was processed to obtain the daily mean concentrations of NO_2, CO, $PM_{2.5}$, PM_{10}, SO_2 and O_3 for each province. Furthermore, this information was stratified according to three types of monitoring stations: urban, suburban and rural.

Finally, population information was obtained from the Municipal Register compiled by the Spanish Statistics Institute, updated on January 1st, 2020 (INE, 2020).

2.2. Metrics, data aggregation and statistical analysis

The collected information was processed to calculate other variables and carry out different analyses. Specifically, when daily data was available, cumulative data was calculated by making a cumulative sum of the data from the previous days. In the same way, if cumulative information was available, it was used to obtain daily data by subtracting the value of the previous day from the value of the reference day. Cumulative mortality rates and cumulative incidence rates were calculated for each day dividing the number of cumulative cases or cumulative deaths by the total population of the territory. Case fatality rates (CFRs) have been also calculated in order to assess the lethality of the disease over time, dividing the number of deaths by cases. However, this approach has been criticized due to the possible bias produced by the rapid expansion of the number of infected by COVID-19 (Baud et al., 2020; Spychalski et al., 2020). To take this potential bias into account we have also obtained case recovery rates (CRRs), which are calculated as the division of the cumulative recovered patients by cumulative cases. Furthermore, a 7-days and 14-days cumulative incidence was calculated as well as the percentage of daily increase of cases. Finally, 3-days, 7-days and 14-days rolling averages were obtained for

daily variables in order to improve the observation of the trends. These rolling averages are calculated as the average of the value of a day and the previous n-1 days.

All the analyses described in this manuscript were performed using the DatAC data and analytical functionalities. For correlation analysis we chose the Spearman coefficient because **t**here may not be a linear relationship between the analyzed variables. In order to take into account the lockdown effect on the COVID-19 incidence, partial correlation was applied (Ahmadi et al., 2020) correcting by the number of lockdown days. We used the 7-day rolling average for ~~temperature~~ climatic variables and a 7-day lag for the number of lockdown days, because the time between the infection and the appearance of the first symptoms is usually 5-6 days (up to 14 days) (Lauer et al., 2020; Xie and Zhu, 2020). Correlation P-values were adjusted to correct for multiple testing with false discovery rate (FDR) method (Benjamini Y and Hochberg Y, 1995).

2.3. DatAC tool for easy data exploration and interactive analysis

DatAC has been developed with the RStudio Shiny framework. Internally, the application uses R packages to perform all the plots and calculations. Leaflet package (Cheng et al., 2019) is used to generate the interactive map. Interactive plots are generated with plotly package (Sievert, 2020). Partial correlations are calculated with ppcor package (Kim, 2015). The tool runs on a dedicated server with Ubuntu 18.04 operating system, 16 processors and 32 Gb of RAM memory. The source code as well as all the data contained in the application is available with a free license through GENyO Bioinformatics Unit GitHub repository (https://github.com/GENyO-BioInformatics/DatAC).

We collected and curated data since January 1st, 2020 and it is being updated daily with the new data reported by the different sources. As epidemiological data we collected total and PCR+ cases, deaths, recovered patients, hospitalized patients and patients transferred to ICU, all of them as cumulative and daily data. In addition, we calculated the incidence and deaths rates, the percentage of recovered and deaths, the cumulative incidence rate per 100,000 habitants in 14 and 7 days and the percentage increase in the number of daily positive cases. Regarding meteorological data, we collected daily data for temperature, rainfall, wind speed and solar radiation. For air quality data, we included daily measures for NO_2, CO, $PM_{2.5}$, PM_{10}, SO_2 and O_3 pollutants. There are different levels of geographical aggregation, including autonomous communities and provinces. Environmental data is also available for urban, suburban and rural monitoring stations.

The application is structured in three main modules, including a Map, Trend analysis and Time trends panels. The Map panel aggregates and visualizes data with spatial information (Figure 1A). Different variables can be selected and the map and visualization panel are updated dynamically to show and generate reactive plots (Figure 1A and 1B). More advanced analyses can be performed in the Trend analysis tab, where different models (polynomial models and correlation with adjustable parameters) can be applied to study the relationship between two variables (Figure 1C). A lag between the two variables can be applied in order to explore potential relationships between variables with some time difference. Finally, the temporal trends of the different variables can be represented at the Time trends page, comparing two variables for all the desired regions (Figure 1D).

In the following sections we cover the results from analyses carried out with the application to explore the evolution of contaminants during the lockdown and the effect of environment variables on the COVID-19 cases.

3. Results

The data and visual analytics capabilities included in DatAC can be very useful to explore trends and associations among climate and air quality variables and COVID-19 indicators, with a layer of spatio-temporal information. In this work we have used the tool to analyze the air quality evolution during the lockdown as well as the potential association of climatic factors with COVID-19 data. In the following sections we provide a detailed analysis of the obtained results.

3.1. Air quality improved after lockdown in urban and suburban environments

Taking the different phases of the COVID-19 pandemic in Spain (see Introduction) into account, we analyzed the pollutants levels in 3 periods: prior to lockdown (January 23rd to March 13th, 50 days), strict lockdown (March 14th to May 3rd, 50 days) and relaxed lockdown (May 4th to June 20th, 47 days).

We analyzed all the Spanish territory except Ceuta and Melilla autonomous cities, where no official air quality monitoring stations are installed. We also excluded the Canary Islands from the analysis because we detected some outlier values, likely because of measuring errors. The results for the rest of the autonomous communities in urban environments are compiled in Supplementary Table 1. Average values for Spain are shown in Table 1 and Figure 2. As expected, all the analyzed pollutants except O_3 dropped significantly during the strict lockdown, specially NO_2 and PM_{10}. O_3 levels rose more than 50 % in this same period. During the relaxed lockdown, NO_2 and PM_{10} levels increased compared to the strict lockdown period, but the levels of all pollutants except O_3 are still significantly lower than prior to lockdown. Interestingly, CO, SO_2 and $PM_{2.5}$ levels continued to fall moderately in spite of the relaxation of the lockdown measures, while O_3 continued to rise. The trends for suburban environments are similar (Supplementary Tables 2 and 3).

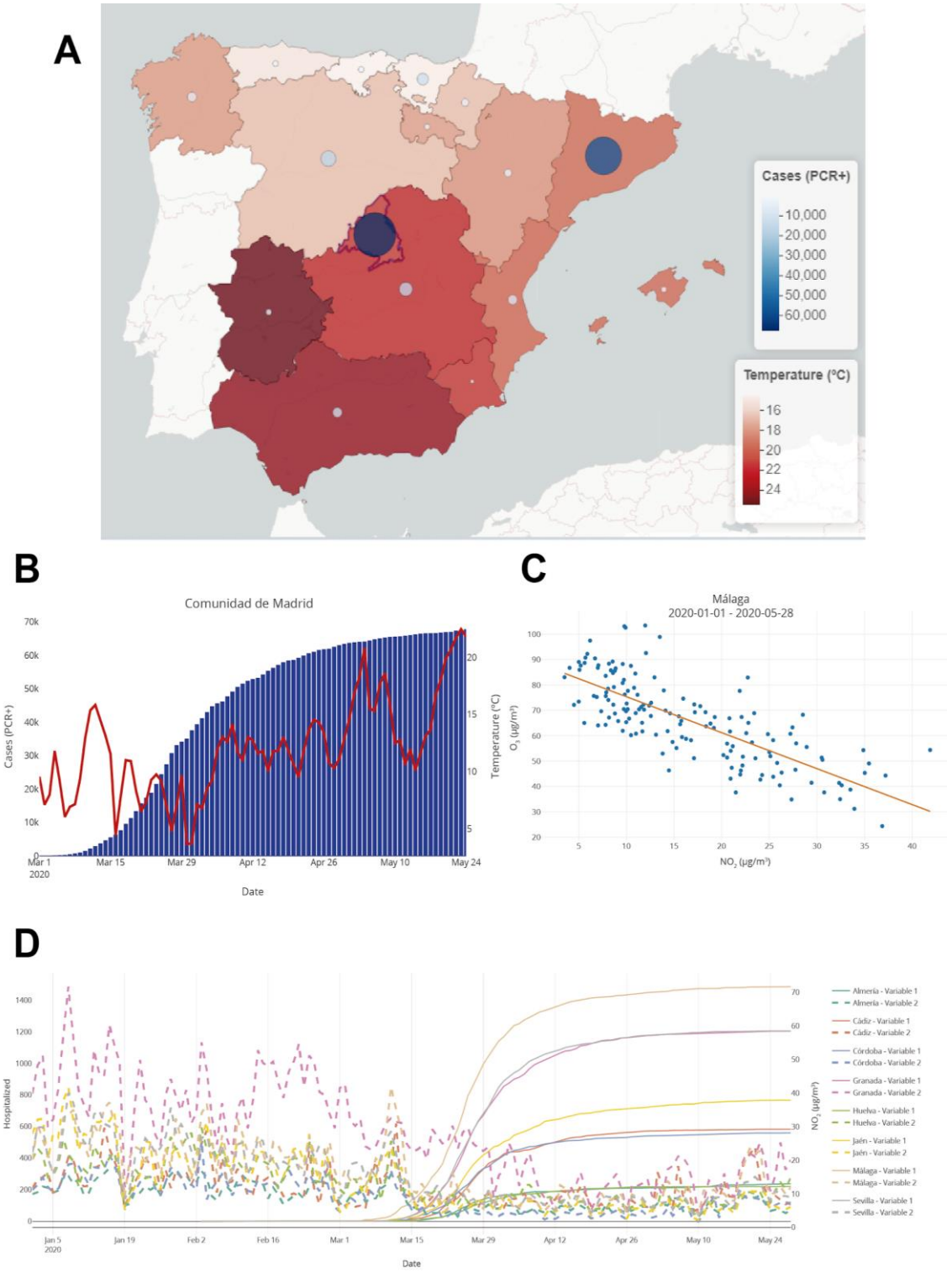


Figure 1. DatAC sample outputs, including: **A)** Map with the COVID-19 cases confirmed by PCR test (blue circles) and mean temperature (background color) for the autonomous communities of Spain on May 24th, 2020. **B)** Longitudinal plot with the same variables as A) for the Comunidad de Madrid region from March 1st to May 24th, 2020, representing the cases with bars and the temperature with the red line. **C)** Correlation plot between NO_2 and O_3 concentrations for Málaga province from January 1st to May 28th, 2020. **D)** Longitudinal plot representing hospitalized patients (solid lines) and NO_2 concentration (semi continuous lines) for all the Andalusian provinces from January 1st to May 24th, 2020.

Table 1. Average air pollutants levels in urban environments of Spain during the 3 periods and the variation between periods.

Pollutant	Prior to lockdown mean (SD)	Strict lockdown mean (SD)	Relaxed lockdown mean (SD)	Difference between strict lockdown and prior to lockdown (% change)	Difference between relaxed lockdown and prior to lockdown (% change)	Difference between relaxed lockdown and strict lockdown (% change)
NO_2 ($\mu g/m^3$)	23.8 (5.67)	8.95 (2.4)	9.93 (2.55)	-14.85 (-62.39 %)	-13.88 (-58.31 %)	0.97 (10.86 %)
CO (mg/m^3)	0.33 (0.04)	0.26 (0.02)	0.23 (0.01)	-0.08 (-22.88 %)	-0.1 (-30 %)	-0.02 (-9.24 %)
$PM_{2.5}$ ($\mu g/m^3$)	12.06 (4.13)	8.48 (2.47)	8.05 (2.16)	-3.58 (-29.67 %)	-4.01 (-33.24 %)	-0.43 (-5.09 %)
PM_{10} ($\mu g/m^3$)	24.9 (10.91)	15.14 (3.93)	16.33 (3.12)	-9.75 (-39.18 %)	-8.57 (-34.41 %)	1.19 (7.84 %)
SO_2 ($\mu g/m^3$)	3.72 (0.36)	3.15 (0.24)	2.97 (0.23)	-0.57 (-15.38 %)	-0.75 (-20.04 %)	-0.17 (-5.51 %)
O_3 ($\mu g/m^3$)	40.22 (10.97)	60.37 (6.87)	62.88 (5.73)	20.15 (50.09 %)	22.66 (56.33 %)	2.51 (4.16 %)

3.2. Lockdown had a significantly lower impact on air quality in rural areas

We analyzed the pollutants trends in Spain in rural monitoring stations (see Table 2 for a summary and Supplementary Table 4 for the complete results). As can be observed, the trends for strict and relaxed lockdowns periods are similar to those observed in urban stations, but the variations between periods are smaller in rural environments. The major differences are found in NO_2 (-62.39 % and -38.16 % variation between strict lockdown and prior to lockdown in urban and rural stations respectively) and in O_3 (+50.09 % and +15.58 % variation between strict lockdown and prior to lockdown in urban and rural stations respectively). Many differences can be found in the variations between relaxed and strict lockdowns: CO, SO_2 and $PM_{2.5}$ levels dropped in both urban and rural environments, but in rural stations the differences are greater. On the other hand, both PM_{10} and O_3 levels have risen during the relaxed lockdown but more in urban areas than in rural stations. Regarding NO_2 we observed a difference in the trend: it raised 10.86 % in urban stations during the relaxed lockdown but dropped 4.34 % in rural stations. All these trends can be observed in Figure 2.

Table 2. Average air pollutants levels in rural environments of Spain during the 3 periods and the variation between periods.

Pollutant	Prior to lockdown mean (SD)	Strict lockdown mean (SD)	Relaxed lockdown mean (SD)	Difference between strict lockdown and prior to lockdown (% change)	Difference between relaxed lockdown and prior to lockdown (% change)	Difference between relaxed lockdown and strict lockdown (% change)
NO_2 ($\mu g/m^3$)	5.26 (1.26)	3.25 (0.63)	3.11 (0.43)	-2.01 (-38.16 %)	-2.15 (-40.84 %)	-0.14 (-4.34 %)
CO (mg/m^3)	0.31 (0.02)	0.28 (0.04)	0.23 (0.02)	-0.03 (-9.8 %)	-0.08 (-25.22 %)	-0.05 (-17.09 %)
$PM_{2.5}$ ($\mu g/m^3$)	8.35 (3.64)	7.47 (2.27)	6.34 (1.65)	-0.88 (-10.57 %)	-2 (-24.01 %)	-1.12 (-15.04 %)
PM_{10} ($\mu g/m^3$)	17.46 (12.95)	12.66 (4.34)	13.49 (3.43)	-4.8 (-27.49 %)	-3.96 (-22.69 %)	0.84 (6.61 %)
SO_2 ($\mu g/m^3$)	2.01 (0.22)	1.86 (0.16)	1.72 (0.14)	-0.15 (-7.34 %)	-0.29 (-14.63 %)	-0.15 (-7.86 %)
O_3 ($\mu g/m^3$)	58.86 (9.47)	68.02 (7.01)	70.15 (7.55)	9.17 (15.58 %)	11.3 (19.2 %)	2.13 (3.13 %)

Figure 2. Average pollutants levels in Spain across time during the three periods of the COVID-19 pandemic**.** Green, red and blue backgrounds represent prior to lockdown, strict lockdown and relaxed lockdown periods respectively.

3.3. Analysis of the association among climatic variables and COVID-19 incidence

There are previous studies that associate temperature, solar radiation, wind speed or rainfall with COVID-19 cases or deaths in different regions (Bashir et al., 2020a; Guasp et al., 2020; Kifer et al., 2020; Liu et al., 2020; Luo et al., 2020; Pequeno et al., 2020; Rosario et al., 2020; Tosepu et al., 2020; J. Wang et al., 2020). However, generally there are contradictory and/or inconclusive findings. We calculated the Spearman correlation among temperature, solar radiation, wind speed, rainfall and daily cases (Supplementary Table 5). These correlations were analyzed in all Spanish communities during the period from March 7th to June 20th (end of the state of alarm in Spain). As can be observed, correlations among wind speed and rainfall with cases are very heterogeneous and non-significant for most of the Spanish communities, indicating that these variables were not correlated with the COVID-19 incidence in Spain. On the other hand, temperature and solar radiation are negatively correlated with cases and these correlations are significant for the majority of communities.

However, considering that a strict lockdown was imposed at the beginning of the analyzed period, it is expected that social distancing measures were the actual factor causing the cases decreasing. Therefore, when we calculated partial correlations between temperature, solar radiation and daily detected COVID-19 cases controlling the influence

of the lockdown. (Table 3), correlation coefficients were low and non significant for most of the communities.

In order to check if lockdown was linked to the decrease in daily cases regardless of temperature and solar radiation, we calculated the correlation between lockdown days and daily cases controlling for the effect of these two variables (Table 3). As can be observed, days of lockdown were very negatively correlated with the daily cases and these correlations are significant for the entire territory. These results indicate that although temperature and solar radiation can have a role in COVID-19 incidence, they are not the main factors. and long-term data is required in order to have conclusive results. The evolution of the pandemic during this year will be very important to really understand climatic factors that can be important for the spread, incidence and severity of the virus.

Table 3. Partial correlations among temperature, solar radiation, lockdown days and daily cases during the period March 7th to June 20th for the Spanish autonomous communities. Lockdown correlation was corrected for both temperature and solar radiation variables.

	Temperature vs. cases controlling lockdown			Solar radiation vs. cases controlling lockdown		
Autonomous communnity	Spearman partial correlation	P-value	FDR	Spearman partial correlation	P-value	FDR
Andalucía	-0.2321	0.0172	0.0544	-0.1791	0.0675	0.1603
Aragón	-0.2520	0.0095	0.0452	-0.3455	0.0003	0.0058
Canarias	0.0527	0.5934	0.6392	-0.3269	0.0007	0.0063
Cantabria	-0.3078	0.0014	0.0133	-0.0314	0.7508	0.7925
Castilla-La Mancha	-0.1262	0.1996	0.2917	-0.0852	0.3875	0.5663
Castilla y León	-0.1144	0.2452	0.2953	-0.2217	0.0231	0.0864
Cataluña	-0.2055	0.0355	0.0963	-0.2155	0.0273	0.0864
Ciudad de Ceuta	-0.1210	0.2188	0.2953	-0.1558	0.1125	0.2376
Ciudad de Melilla	-0.0463	0.6392	0.6392	0.1251	0.2034	0.3513
Comunidad de Madrid	-0.1541	0.1166	0.2215	-0.0763	0.4393	0.5961
Comunidad Foral de Navarra	-0.1814	0.064	0.1351	0.0086	0.9303	0.9303
Comunitat Valenciana	-0.1426	0.1468	0.2340	-0.1810	0.0646	0.1603
Extremadura	-0.1422	0.1478	0.2340	-0.2907	0.0026	0.0166
Galicia	-0.2387	0.0142	0.0539	-0.0338	0.7320	0.7925
Illes Balears	-0.2579	0.0079	0.0452	-0.2755	0.0044	0.0211
La Rioja	-0.3769	0.0001	0.0014	-0.1110	0.2597	0.4111
País Vasco	-0.1883	0.0544	0.1291	-0.1500	0.1268	0.2408
Principado de Asturias	-0.0488	0.6209	0.6392	-0.0489	0.6202	0.7364
Región de Murcia	-0.1136	0.2487	0.2953	-0.0608	0.5380	0.6815
	Lockdown vs. cases controlling temperature			Lockdown vs. cases controlling solar radiation		
Autonomous communnity	Spearman partial correlation	P-value	FDR	Spearman partial correlation	P-value	FDR
Andalucía	-0.7216	<0.0001	<0.0001	-0.8404	<0.0001	<0.0001

Aragón	-0.5837	<0.0001	<0.0001	-0.7800	<0.0001	<0.0001
Canarias	-0.6385	<0.0001	<0.0001	-0.8253	<0.0001	<0.0001
Cantabria	-0.5224	<0.0001	<0.0001	-0.8168	<0.0001	<0.0001
Castilla-La Mancha	-0.6711	<0.0001	<0.0001	-0.8179	<0.0001	<0.0001
Castilla y León	-0.6499	<0.0001	<0.0001	-0.8335	<0.0001	<0.0001
Cataluña	-0.5500	<0.0001	<0.0001	-0.8194	<0.0001	<0.0001
Ciudad de Ceuta	-0.279	0.0039	0.0039	-0.3638	0.0001	0.0001
Ciudad de Melilla	-0.4719	<0.0001	<0.0001	-0.7144	<0.0001	<0.0001
Comunidad de Madrid	-0.7518	<0.0001	<0.0001	-0.8489	<0.0001	<0.0001
Comunidad Foral de Navarra	-0.7700	<0.0001	<0.0001	-0.8994	<0.0001	<0.0001
Comunitat Valenciana	-0.6896	<0.0001	<0.0001	-0.8349	<0.0001	<0.0001
Extremadura	-0.6078	<0.0001	<0.0001	-0.7672	<0.0001	<0.0001
Galicia	-0.6863	<0.0001	<0.0001	-0.8915	<0.0001	<0.0001
Illes Balears	-0.3692	0.0001	0.0001	-0.6951	<0.0001	<0.0001
La Rioja	-0.6312	<0.0001	<0.0001	-0.8897	<0.0001	<0.0001
País Vasco	-0.8306	<0.0001	<0.0001	-0.9161	<0.0001	<0.0001
Principado de Asturias	-0.8376	<0.0001	<0.0001	-0.9202	<0.0001	<0.0001
Región de Murcia	-0.5420	<0.0001	<0.0001	-0.7492	<0.0001	<0.0001

4. Discussion and conclusions

During COVID-19 pandemic, real-time data availability and accurate quality data repositories are essential in order to get insights into the possible effect of different factors in the SARS-CoV-2 spread and disease incidence. This might help to assess government decisions, for early temporal and geographic detection of new focuses of infection or to make predictions about the evolution of the pandemic.

In this context, large efforts have been made to develop software tools to collect COVID-19 global pandemic data, like the dashboard developed by the John Hopkins University (Dong et al., 2020) or HealthMap (Xu et al., 2020). Nevertheless, to the best of our knowledge, DatAC is the first application that integrates epidemiological data with meteorological and air quality information. Although the first release of the application is based on Spain, we have made the code publicly available so it can be adapted for other regions.

Using the data and analyses implemented in DatAC we evaluated the impact of the lockdown measures in Spain on the air quality in urban and rural environments. NO_2, CO, $PM_{2.5}$, PM_{10} and SO_2 declined after lockdown in all the Spanish territory, especially in urban environments. This observation is coherent with previous local studies in Spain and other countries (Bao and Zhang, 2020; Nakada and Urban, 2020; Tobías et al., 2020). NO_2 is the pollutant with the major reduction in both urban and rural areas. This is expected due to outdoor NO_2 main source is traffic (IARC Working Group on the Evaluation of Carcinogenic Risk to Humans, 2016), which was very limited during

lockdown. For the other pollutants, although the decrease is also significant, other natural and anthropogenic sources may be maintaining certain emissions even during lockdown. For instance, SO_2 anthropogenic emissions main sources are industry and power sectors (IARC Working Group on the Evaluation of Carcinogenic Risk to Humans, 2016). On the other hand, O_3 is the only analyzed pollutant with higher concentration during lockdown. This was also observed in other regions (Sicard et al., 2020) and can be explained by the reduction of NO_x, $PM_{2.5}$ and PM_{10} and by a higher VOCs-NO_x ratio (Sicard et al., 2020). Interestingly, CO, $PM_{2.5}$ and SO_2 levels have continued decreasing in Spain after relaxation of the lockdown constraints. CO and $PM_{2.5}$ are produced in the incomplete combustion of carbon-containing fuels (Cheng et al., 2017; Elbayoumi et al., 2014). High temperatures facilitate the complete combustion of fuels, so the effect of the rising temperatures during the relaxed lockdown (which started in May) may be influencing more than the traffic back during this period (Rozante et al., 2017). In addition, usage of heating sources like stoves, fires, etc., which are another important source of CO and $PM_{2.5}$, drops with warm weather, reducing even more the concentration of these pollutants when temperature rises. Regarding SO_2, its main source is the coal combustion in electrical power plants (MITECO, 2020; Schreifels et al., 2012), given that the main gaseous residue produced by coal burning is SO_2 (Miller, 2017). During the relaxed lockdown the production of electricity from this source dropped 11 % (REData, 2020), so we hypothesize that the reduction of the main SO_2 production source is the cause of the SO_2 concentration decrease during this period.

We also used DatAC to explore the relationship between meteorological variables ~~temperature~~ and the amount of daily COVID-19 cases, finding heterogeneous and non-significant correlation for wind speed and rainfall, but large and significant negative correlation for temperature and solar radiation in almost all Spanish communities. After correcting these correlations for the lockdown effect on the pandemic, these are basically lost. On the contrary, we found that correlation between lockdown and cases is substantial and statistically significant after correcting for temperature and solar radiation effects. These results indicate that lockdown, and not temperature nor solar radiation, was the driving factor of the COVID-19 pandemic evolution in Spain. This is in agreement with previous studies which reported no correlation between temperature and cases in Spain (Briz-Redón and Serrano-Aroca, 2020).

More data and longer records are required to derive more conclusive results. DatAC will be also a very valuable resource in this context, as the application will be updated periodically and it will contain the historical registry since the appearance of the pandemic.

We are sure that DatAC will be a very valuable resource for monitoring possible future outbreaks, as well as trends in air quality data and weather. In addition, the inclusion of more data in the next months will provide more reliable results about evolution of environmental factors and their impact on the spread of the disease, by means of using the analytic functionalities provided within the application or downloading the data that it is also publicly available to use with third party software.

Acknowledgements

We would like to thank Alberto Ramírez and Manuel Orcera for their technical support during the implementation. This work is part of the Jordi Martorell-Marugán's and Juan Antonio Villatoro-García's PhD theses. Jordi Martorell-Marugán is enrolled in the PhD program in Biomedicine at the University of Granada, Spain. Juan Antonio Villatoro-García is enrolled in the PhD program in Mathematical and Applied Statistics at the University of Granada, Spain.

Funding sources

Jordi Martorell-Marugán is partially funded by Ministerio de Economía, Industria y Competitividad. This work was partially supported by Consejería de Salud, Junta de Andalucía [grant number PI-0173-2017].

Author contributions

Jordi Martorell-Marugán: Software development, Methodology, Writing - Original Draft. **Juan Antonio Villatoro-García:** Methodology, Formal analysis, Data Curation, Writing - Review & Editing. **Adrián García-Moreno:** Software development and testing, Writing - Review & Editing. **Raúl López-Domínguez:** Software testing, Writing - Review & Editing. **Francisco Requena:** Software development. **Juan Julián Merelo:** Analysis, Data Curation. **Juan J. Monchón:** Data curation, Writing - Review & Editing. **Marina Lacasaña:** Validation, Writing - Review & Editing. **Juan de Dios Luna:** Validation, Writing - Review & Editing. **José A. Lorente:** Validation, Writing - Review & Editing. **Pedro Carmona-Sáez:** Conceptualization, Supervision, Funding acquisition, Original Draft.

Supplementary data

Supplementary Tables 1_5.xlsx: Excel file with Supplementary Tables 1 to 5.

References

AEMET, 2020. AEMET OpenData [WWW Document]. URL https://opendata.aemet.es/centrodedescargas/inicio (accessed 6.9.20).

Ahmadi, M., Sharifi, A., Dorosti, S., Jafarzadeh Ghoushchi, S., Ghanbari, N., 2020. Investigation of effective climatology parameters on COVID-19 outbreak in Iran. Science of The Total Environment 729, 138705. https://doi.org/10.1016/j.scitotenv.2020.138705

Bao, R., Zhang, A., 2020. Does lockdown reduce air pollution? Evidence from 44 cities in northern China. Sci. Total Environ. 731, 139052. https://doi.org/10.1016/j.scitotenv.2020.139052

Bashir, M.F., Ma, B., Bilal, Komal, B., Bashir, M.A., Tan, D., Bashir, M., 2020a. Correlation between climate indicators and COVID-19 pandemic in New York, USA. Science of The Total Environment 728, 138835. https://doi.org/10.1016/j.scitotenv.2020.138835

Bashir, M.F., Ma, B.J., Bilal, null, Komal, B., Bashir, M.A., Farooq, T.H., Iqbal, N., Bashir, M., 2020b. Correlation between environmental pollution indicators and COVID-19 pandemic: A brief study in Californian context. Environ. Res. 187, 109652.

https://doi.org/10.1016/j.envres.2020.109652
Baud, D., Qi, X., Nielsen-Saines, K., Musso, D., Pomar, L., Favre, G., 2020. Real estimates of mortality following COVID-19 infection. Lancet Infect Dis. https://doi.org/10.1016/S1473-3099(20)30195-X
Benjamini Y, Hochberg Y, 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society 57, 289–300.
Briz-Redón, Á., Serrano-Aroca, Á., 2020. A spatio-temporal analysis for exploring the effect of temperature on COVID-19 early evolution in Spain. Sci. Total Environ. 728, 138811. https://doi.org/10.1016/j.scitotenv.2020.138811
Cheng, J., Karambelkar, B., Xie, Y., 2019. leaflet: Create Interactive Web Maps with the JavaScript "Leaflet" Library.
Cheng, N., Zhang, D., Li, Y., Xie, X., Chen, Z., Meng, F., Gao, B., He, B., 2017. Spatio-temporal variations of PM2.5 concentrations and the evaluation of emission reduction measures during two red air pollution alerts in Beijing. Sci Rep 7, 8220. https://doi.org/10.1038/s41598-017-08895-x
Coccia, M., 2020. Factors determining the diffusion of COVID-19 and suggested strategy to prevent future accelerated viral infectivity similar to COVID. Science of The Total Environment 729, 138474. https://doi.org/10.1016/j.scitotenv.2020.138474
Datadista, 2020. Datadista [WWW Document]. GitHub. URL https://github.com/datadista (accessed 6.9.20).
Dong, E., Du, H., Gardner, L., 2020. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis 20, 533–534. https://doi.org/10.1016/S1473-3099(20)30120-1
Elbayoumi, M., Ramli, N.A., Md Yusof, N.F.F., Madhoun, W.A., 2014. The effect of seasonal variation on indoor and outdoor carbon monoxide concentrations in Eastern Mediterranean climate. Atmospheric Pollution Research 5, 315–324. https://doi.org/10.5094/APR.2014.037
Escovid19data, 2020. Escovid19data: Capturando datos de COVID-19 por provincias en España [WWW Document]. URL https://github.com/montera34/escovid19data (accessed 6.9.20).
European Environment Agency, 2020. European Air Quality Portal – e-Reporting. URL https://aqportal.discomap.eea.europa.eu/ (accessed 6.9.20).
Fronza, R., Lusic, M., Schmidt, M., Lucic, B., 2020. Spatial-Temporal Variations in Atmospheric Factors Contribute to SARS-CoV-2 Outbreak. Viruses 12. https://doi.org/10.3390/v12060588
Guan, W.-J., Ni, Z.-Y., Hu, Y., Liang, W.-H., Ou, C.-Q., He, J.-X., Liu, L., Shan, H., Lei, C.-L., Hui, D.S.C., Du, B., Li, L.-J., Zeng, G., Yuen, K.-Y., Chen, R.-C., Tang, C.-L., Wang, T., Chen, P.-Y., Xiang, J., Li, S.-Y., Wang, J.-L., Liang, Z.-J., Peng, Y.-X., Wei, L., Liu, Y., Hu, Y.-H., Peng, P., Wang, J.-M., Liu, J.-Y., Chen, Z., Li, G., Zheng, Z.-J., Qiu, S.-Q., Luo, J., Ye, C.-J., Zhu, S.-Y., Zhong, N.-S., China Medical Treatment Expert Group for Covid-19, 2020. Clinical Characteristics of Coronavirus Disease 2019 in China. N. Engl. J. Med. https://doi.org/10.1056/NEJMoa2002032
Guasp, M., Laredo, C., Urra, X., 2020. Higher Solar Irradiance Is Associated With a Lower Incidence of Coronavirus Disease 2019. Clinical Infectious Diseases. https://doi.org/10.1093/cid/ciaa575
IARC Working Group on the Evaluation of Carcinogenic Risk to Humans, 2016. Sources of air pollutants, Outdoor air pollution. International Agency for Research on Cancer.
IECA, 2020. Instituto de Estadística y Cartografía de Andalucía [WWW Document]. URL http://www.juntadeandalucia.es/institutodeestadisticaycartografia (accessed 6.9.20).
INE, 2020. INEbase / Demography and population [WWW Document]. INE. URL https://www.ine.es/dyngs/INEbase/en/categoria.htm?c=Estadistica_P&cid=1254734710984 (accessed 6.22.20).
Jiang, Y., Wu, X.-J., Guan, Y.-J., 2020. Effect of ambient air pollutants and meteorological variables on COVID-19 incidence. Infect Control Hosp Epidemiol 1–11. https://doi.org/10.1017/ice.2020.222

Junta de Andalucía, 2020. Informes diarios de calidad del aire :: Red de Información Ambiental de Andalucía :: Consejería de Medio Ambiente y Ordenación del Territorio :: Junta de Andalucía :: [WWW Document]. URL http://www.juntadeandalucia.es/medioambiente/site/rediam/ (accessed 6.17.20).

Kifer, D., Bugada, D., Villar-Garcia, J., Gudelj, I., Menni, C., Sudre, C.H., Vuckovic, F., Ugrina, I., Lorini, L.F., Bettinelli, S., Ughi, N., Maloberti, A., Epis, O., Giannattasio, C., Rossetti, C., Kalogjera, L., Persec, J., Ollivere, L., Ollivere, B., Yan, H., Cai, T., Aithal, G., Steves, C., Kantele, A., Kajova, M., Vapalahti, O., Sajantila, A., Wojtowicz, R., Wierzba, W., Krol, Z., Zaczynski, A., Zycinska, K., Postula, M., Luksic, I., Civljak, R., Markotic, A., Mahnkopf, C., Markl, A., Brachmann, J., Murray, B., Ourselin, S., Pascual, J., Valdes, A.M., Posso, M., Horcajada, J., Castells, X., Allegri, M., Primorac, D., Spector, T., Barrios, C., Lauc, G., 2020. Effects of environmental factors on severity and mortality of COVID-19. medRxiv 2020.07.11.20147157. https://doi.org/10.1101/2020.07.11.20147157

Kim, S., 2015. ppcor: An R Package for a Fast Calculation to Semi-partial Correlation Coefficients. Commun Stat Appl Methods 22, 665–674. https://doi.org/10.5351/CSAM.2015.22.6.665

Lauer, S.A., Grantz, K.H., Bi, Q., Jones, F.K., Zheng, Q., Meredith, H.R., Azman, A.S., Reich, N.G., Lessler, J., 2020. The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. Ann Intern Med. https://doi.org/10.7326/M20-0504

Liu, J., Zhou, J., Yao, J., Zhang, X., Li, L., Xu, X., He, X., Wang, B., Fu, S., Niu, T., Yan, J., Shi, Y., Ren, X., Niu, J., Zhu, W., Li, S., Luo, B., Zhang, K., 2020. Impact of meteorological factors on the COVID-19 transmission: A multi-city study in China. Sci. Total Environ. 726, 138513. https://doi.org/10.1016/j.scitotenv.2020.138513

Luo, W., Majumder, M.S., Liu, D., Poirier, C., Mandl, K.D., Lipsitch, M., Santillana, M., 2020. The role of absolute humidity on transmission rates of the COVID-19 outbreak. medRxiv 2020.02.12.20022467. https://doi.org/10.1101/2020.02.12.20022467

Ma, Y., Zhao, Y., Liu, J., He, X., Wang, B., Fu, S., Yan, J., Niu, J., Zhou, J., Luo, B., 2020. Effects of temperature variation and humidity on the death of COVID-19 in Wuhan, China. Science of The Total Environment 724, 138226. https://doi.org/10.1016/j.scitotenv.2020.138226

Menebo, M.M., 2020. Temperature and precipitation associate with Covid-19 new daily cases: A correlation study between weather and Covid-19 pandemic in Oslo, Norway. Sci Total Environ 737, 139659. https://doi.org/10.1016/j.scitotenv.2020.139659

Miller, B.G., 2017. 3 - The Effect of Coal Usage on Human Health and the Environment, in: Miller, B.G. (Ed.), Clean Coal Engineering Technology (Second Edition). Butterworth-Heinemann, pp. 105–144. https://doi.org/10.1016/B978-0-12-811365-3.00003-X

MISAN, 2020. COVID-19 Panel [WWW Document]. URL https://cnecovid.isciii.es/covid19/ (accessed 6.17.20).

MITECO, 2020. Informative Inventory Report. Edition 2020.

Nakada, L.Y.K., Urban, R.C., 2020. COVID-19 pandemic: Impacts on the air quality during the partial lockdown in São Paulo state, Brazil. Sci Total Environ. https://doi.org/10.1016/j.scitotenv.2020.139087

Ogen, Y., 2020. Assessing nitrogen dioxide (NO2) levels as a contributing factor to coronavirus (COVID-19) fatality. Science of The Total Environment 726, 138605. https://doi.org/10.1016/j.scitotenv.2020.138605

Pequeno, P., Mendel, B., Rosa, C., Bosholn, M., Souza, J.L., Baccaro, F., Barbosa, R., Magnusson, W., 2020. Air transportation, population density and temperature predict the spread of COVID-19 in Brazil. PeerJ 8, e9322. https://doi.org/10.7717/peerj.9322

Pollán, M., Pérez-Gómez, B., Pastor-Barriuso, R., Oteo, J., Hernán, M.A., Pérez-Olmeda, M., Sanmartín, J.L., Fernández-García, A., Cruz, I., Fernández de Larrea, N., Molina, M., Rodríguez-Cabrera, F., Martín, M., Merino-Amador, P., León Paniagua, Jose, Muñoz-Montalvo, J.F., Blanco, F., Yotti, R., Blanco, F., Gutiérrez Fernández, R., Martín, M., Mezcua Navarro, S., Molina, M., Muñoz-Montalvo, J.F., Salinero Hernández, M.,

Sanmartín, J.L., Cuenca-Estrella, M., Yotti, R., León Paniagua, José, Fernández de Larrea, N., Fernández-Navarro, P., Pastor-Barriuso, R., Pérez-Gómez, B., Pollán, M., Avellón, A., Fedele, G., Fernández-García, A., Oteo Iglesias, J., Pérez Olmeda, M.T., Cruz, I., Fernandez Martinez, M.E., Rodríguez-Cabrera, F.D., Hernán, M.A., Padrones Fernández, S., Rumbao Aguirre, J.M., Navarro Marí, J.M., Palop Borrás, B., Pérez Jiménez, A.B., Rodríguez-Iglesias, M., Calvo Gascón, A.M., Lou Alcaine, M.L., Donate Suárez, I., Suárez Álvarez, O., Rodríguez Pérez, M., Cases Sanchís, M., Villafáfila Gomila, C.J., Carbo Saladrigas, L., Hurtado Fernández, A., Oliver, A., Castro Feliciano, E., González Quintana, M.N., Barrasa Fernández, J.M., Hernández Betancor, M.A., Hernández Febles, M., Martín Martín, L., López López, L.-M., Ugarte Miota, T., De Benito Población, I., Celada Pérez, M.S., Vallés Fernández, M.N., Maté Enríquez, T., Villa Arranz, M., Domínguez-Gil González, M., Fernández-Natal, I., Megías Lobón, G., Muñoz Bellido, J.L., Ciruela, P., Mas i Casals, A., Doladé Botías, M., Marcos Maeso, M.A., Pérez del Campo, D., Félix de Castro, A., Limón Ramírez, R., Elías Retamosa, M.F., Rubio González, M., Blanco Lobeiras, M.S., Fuentes Losada, A., Aguilera, A., Bou, G., Caro, Y., Marauri, N., Soria Blanco, L.M., del Cura González, I., Hernández Pascual, M., Alonso Fernández, R., Merino-Amador, P., Cabrera Castro, N., Tomás Lizcano, A., Ramírez Almagro, C., Segovia Hernández, M., Ascunce Elizaga, N., Ederra Sanz, M., Ezpeleta Baquedano, C., Bustinduy Bascaran, A., Iglesias Tamayo, S., Elorduy Otazua, L., Benarroch Benarroch, R., Lopera Flores, J., Vázquez de la Villa, A., 2020. Prevalence of SARS-CoV-2 in Spain (ENE-COVID): a nationwide, population-based seroepidemiological study. The Lancet. https://doi.org/10.1016/S0140-6736(20)31483-5

Prata, D.N., Rodrigues, W., Bermejo, P.H., 2020. Temperature significantly changes COVID-19 transmission in (sub)tropical cities of Brazil. Sci. Total Environ. 729, 138862. https://doi.org/10.1016/j.scitotenv.2020.138862

REData, 2020. Red Eléctrica de España [WWW Document]. URL https://www.ree.es/es/datos/generacion/estructura-generacion (accessed 7.16.20).

Riccò, M., Ranzieri, S., Balzarini, F., Bragazzi, N.L., Corradi, M., 2020. SARS-CoV-2 infection and air pollutants: Correlation or causation? Sci Total Environ. https://doi.org/10.1016/j.scitotenv.2020.139489

Rosario, D.K.A., Mutz, Y.S., Bernardes, P.C., Conte-Junior, C.A., 2020. Relationship between COVID-19 and weather: Case study in a tropical country. International Journal of Hygiene and Environmental Health 229, 113587. https://doi.org/10.1016/j.ijheh.2020.113587

Rozante, J.R., Rozante, V., Souza Alvim, D., Ocimar Manzi, A., Barboza Chiquetto, J., Siqueira D'Amelio, M.T., Moreira, D.S., 2017. Variations of carbon monoxide concentrations in the megacity of São Paulo from 2000 to 2015 in different time scales. Atmosphere 8, 81. https://doi.org/10.3390/atmos8050081

Schreifels, J.J., Fu, Y., Wilson, E.J., 2012. Sulfur dioxide control in China: policy evolution during the 10th and 11th Five-year Plans and lessons for the future. Energy Policy 48, 779–789. https://doi.org/10.1016/j.enpol.2012.06.015

Sicard, P., De Marco, A., Agathokleous, E., Feng, Z., Xu, X., Paoletti, E., Rodriguez, J.J.D., Calatayud, V., 2020. Amplified ozone pollution in cities during the COVID-19 lockdown. Sci. Total Environ. 735, 139542. https://doi.org/10.1016/j.scitotenv.2020.139542

Sievert, C., 2020. Interactive Web-Based Data Visualization with R, plotly, and shiny, 1 edition. ed. Chapman and Hall/CRC, Boca Raton, FL.

Spychalski, P., Błażyńska-Spychalska, A., Kobiela, J., 2020. Estimating case fatality rates of COVID-19. Lancet Infect Dis. https://doi.org/10.1016/S1473-3099(20)30246-2

Tobías, A., Carnerero, C., Reche, C., Massagué, J., Via, M., Minguillón, M.C., Alastuey, A., Querol, X., 2020. Changes in air quality during the lockdown in Barcelona (Spain) one month into the SARS-CoV-2 epidemic. Sci Total Environ. https://doi.org/10.1016/j.scitotenv.2020.138540

Tosepu, R., Gunawan, J., Effendy, D.S., Ahmad, L.O.A.I., Lestari, H., Bahar, H., Asfian, P., 2020. Correlation between weather and Covid-19 pandemic in Jakarta, Indonesia. Science of The Total Environment 725, 138436. https://doi.org/10.1016/j.scitotenv.2020.138436

Triplett, M., 2020. Evidence that higher temperatures are associated with lower incidence of COVID-19 in pandemic state, cumulative cases reported up to March 27, 2020. medRxiv 2020.04.02.20051524. https://doi.org/10.1101/2020.04.02.20051524

Wang, J., Tang, K., Feng, K., Lv, W., 2020. High Temperature and High Humidity Reduce the Transmission of COVID-19. https://doi.org/10.2139/ssrn.3551767

Wang, M., Jiang, A., Gong, L., Luo, L., Guo, W., Li, Chuyi, Zheng, J., Li, Chaoyong, Yang, B., Zeng, J., Chen, Y., Zheng, K., Li, H., 2020. Temperature significant change COVID-19 Transmission in 429 cities. medRxiv 2020.02.22.20025791. https://doi.org/10.1101/2020.02.22.20025791

Xie, J., Zhu, Y., 2020. Association between ambient temperature and COVID-19 infection in 122 cities from China. Science of The Total Environment 724, 138201. https://doi.org/10.1016/j.scitotenv.2020.138201

Xu, B., Kraemer, M.U.G., Open COVID-19 Data Curation Group, 2020. Open access epidemiological data from the COVID-19 outbreak. Lancet Infect Dis 20, 534. https://doi.org/10.1016/S1473-3099(20)30119-5

Zhu, Y., Xie, J., Huang, F., Cao, L., 2020. Association between short-term exposure to air pollution and COVID-19 infection: Evidence from China. Science of The Total Environment 727, 138704. https://doi.org/10.1016/j.scitotenv.2020.138704

9.4. Exploring the interplay between climate, population immunity and SARS-CoV-2 transmission dynamics in Mediterranean countries

Este artículo fue publicado el 13 de julio de 2023 en la revista *Science of the Total Environment* volumen 897, página 165487, DOI: 10.1016/j.scitotenv.2023.165487 bajo licencia *Open Access*, bajo los términos y condiciones de Creative Commons Attribution (CC BY-NC). Esta es la versión aceptada del artículo. De acuerdo con la editorial (Elsevier B.V.) esta versión del artículo tiene un permiso de reutilización no comercial.

Exploring the Interplay between Climate, Population Immunity and SARS-CoV-2 Transmission Dynamics in Mediterranean Countries

Juan Antonio Villatoro-García[a,b], Raúl López-Domínguez[a,b], Jordi Martorell-Marugán[b,c], Juan de Dios Luna[a], José Antonio Lorente[b,d], Pedro Carmona-Sáez[a,b,*]

[a] *Department of Statistics and Operations Research, University of Granada, Granada, Spain.*
[b] *GENYO. Centre for Genomics and Oncological Research: Pfizer / University of Granada / Andalusian Regional Government, PTS Granada, 18016, Granada, Spain.*
[c] *Fundación para la Investigación Biosanitaria de Andalucía Oriental-Alejandro Otero (FIBAO)*
[d] *Department of Legal Medicine and Toxicology, Faculty of Medicine, University of Granada, PTS Granada, 18016 Granada, Spain*

Highlights

- During the pandemic, there were inconsistent findings regarding the impact of meteorological factors on SARS-CoV-2 transmission.
- Population immunity is a confounding factor in the study of the seasonality of COVID-19.
- Temperature and specific humidity did not affect transmission of the disease when the population lacked immunity.
- When there is enough immunization coverage, there is a slight decrease in the transmission of SARS-CoV-2 during warmer periods

Keywords

COVID-19, coronavirus, pandemic, vaccination, seasonality

Abbreviations

COVID-19: Coronavirus disease 2019; **dfs**: degrees of freedom; **GAM**: Generalized Additive Model; **OxCGRT**: Oxford COVID-19 Government Response Tracker; **P1**: Period with low population immunity; **P2**: Period with a certain degree of population immunity due to vaccination; **R_e**: effective reproductive number; **RR**: Relative Risk; **SARS-CoV-2**: severe acute respiratory syndrome coronavirus 2; **SH**: Specific Humidity; **SI**: Stringency Index

Abstract

The relationship between SARS-CoV-2 transmission and environmental factors has been analyzed in numerous studies since the outbreak of the pandemic, resulting in heterogeneous results and conclusions. This may be due to differences in methodology, considered variables, confounding factors, studied periods and/or lack of adequate data. Furthermore, previous works have reported that the lack of population immunity is the fundamental driver in transmission dynamics and can mask the potential impact of environmental variables. In this study, we aimed to investigate the association between climate variables and COVID-19 transmission considering the influence of population immunity. We analyzed two different periods characterized by the absence of vaccination (low population immunity) and a high degree of vaccination (high level of population immunity), respectively. Although this study has some limitations, such us the restriction to a specific climatic zone and the omission of other environmental factors, our results indicate that transmission of SARS-CoV-2 may increase independently of temperature and specific humidity in periods with low levels of population immunity while a negative association is found under conditions with higher levels of population immunity in the analyzed regions.

1. Introduction

Since the onset of the COVID-19 pandemic, great efforts have focused on studying the influence of environmental factors on the transmission of the disease. One key area of research is the seasonal pattern of COVID-19 transmission, which exhibits increased transmission in cold and dry environments (Martinez, 2018), as observed in other respiratory viruses such us influenza and human coronaviruses (Baker et al., 2018; Moriyama et al., 2020).

Despite the extensive research on this topic, many of the studies were preliminary and showed inconsistent findings (Carlson et al., 2020). Specifically, regarding the influence of temperature, different works have reported a negative correlation between temperature and disease transmission, with some showing a considerable impact on the number of new cases and on the variation of the effective reproductive number (R_e) (D'Amico et al., 2022; Fontal et al., 2021; Hoogeveen et al., 2022; Ma et al., 2021; Yamasaki et al., 2021; Yin et al., 2022). However, other studies have reported small effects (Bashir et al., 2020; Briz-Redón and Serrano-Aroca, 2020; Meyer et al., 2020) or no association whatsoever between temperature and transmission (Kassem, 2020; Liu et al., 2022; O'Reilly et al., 2020; Pan et al., 2021).

The inconsistency between the different studies may be attributed to various factors, such as i) the analysis period selected, especially at the initial stages of the pandemic when a considerable number of cases and deaths were not reported (Chatterjee, 2020; Pifarré i Arolas et al., 2021); ii) the use of an inadequate and limited methodology that may have introduced biases in the results obtained (Dong et al., 2021; Nottmeyer et al., 2023; Villeneuve and Goldberg, 2020; Weaver et al., 2022) and iii) the omission of relevant variables that may significantly impact transmission, such as containment measures adopted by governments (Mecenas et al., 2020; Sera et al., 2021; Smit et al., 2020), among others.

Moreover, initial lack of population immunity is a critical factor in virus spread (Baker et al., 2020; Carlson et al., 2020) that has not been considered in previous works. In two previous studies, Baker et al. developed a climate-dependent epidemic model to simulate the SARS-CoV-2 pandemic using data from other human coronaviruses (HKU1 and HCoV-OC43) (Baker et al., 2021, 2020). They found that, while weather fluctuations may to some extent contribute to transmission, high levels of susceptibility (low population immunity) is the main driving factor for the pandemic and will mitigate the effect of environmental variables such as obstruction of spread of infection by high temperature. A limitation of these important studies is that they did not directly estimate the sensitivity of SARS-CoV-2 to climate, despite offering valuable insights into the possible role of weather in the pandemic. In this context, modeling the impact of climate factors in the scenario of population immunity could provide significant clues about the impact of temperature and humidity on virus transmission.

Nevertheless, after more than three years of pandemic, over 4 billion people have been vaccinated with at least two doses worldwide, according to Our World in Data (Mathieu et al., 2021). This available immunity data, together with larger epidemiological records, is an invaluable source for evaluating the effect of climate on the evolution of COVID-19 transmission and obtaining further insights into the potential seasonality pattern.

In this study, we investigated the association between SARS-CoV-2 transmission, temperature, and specific humidity (SH), considering the effect of population immunity. We analyzed data from two periods of the same duration: June to December 2020 (P1) and June to December 2021 (P2). During P1 there was a near absence of population immunity, whereas in P2, a notable level was achieved due to vaccination (approximately 80% of the population having received two or more doses at the end of the period).

In our analyses, we considered temperature and specific humidity as the environmental variables. In addition, to take into account potential confounding factors that could have an impact on our results, we also considered other variables, such as the dominant virus variants and the Stringency Index (SI), which measures the different government restrictions. Results from the analysis of data in Spain revealed that temperature and specific humidity only slightly influence transmission in the analyzed period with vaccination, which was also validated by modeling the meteorological effects in different European countries and Italian regions.

2. Methods

2.1. Data collection

We considered two periods for data collection: the first period (P1) ran from June 1, 2020 to December 31, 2020, and was characterized by a low level of population immunity. The second period (P2) ran from June 1, 2021 to December 31, 2021, during which a certain degree of population immunity was achieved due to vaccination efforts. We selected these specific time periods due to their equal duration and identical dates across two different years, making them more affordable. All the code and data and are available at Github: https://github.com/GENyO-BioInformatics/Covid19_Seasonality.

2.1.1. Data for Spanish Communities

The study focused on 16 autonomous communities (regions) in Spain, excluding the Canary Islands due to their subtropical climate and narrower temperature range, which makes it difficult to compare with other communities.

Data of the evolution of COVID-19 pandemic from the 16 autonomous communities of Spain were extracted from the DatAC project (Martorell-Marugán et al., 2021), a web application that contains data from the national COVID-19 pandemic that were collected from the Spanish Ministry of Health ("Spanish Ministry of Health," 2023) and the Datatista repository ("Datatista," 2022) from the beginning of the pandemic up to mid-2022, when the health authorities stopped publishing detailed data. Specifically, the number of daily COVID-19 cases, the number of daily COVID-19 deaths and the daily percentage of fully COVID-19 vaccinated individuals from the two different periods were considered. Moreover, from the daily COVID-19 cases, the effective reproductive number (Re) was estimated by applying the EpiEstim R package (Cori et al., 2013), assuming an uncertain serial interval with a mean of 4.7 and a standard deviation of 2.9 days, respectively (Baker et al., 2021). The effective reproductive number measures the number of secondary cases generated by a single infected individual. If R_e is greater than 1, the disease is spreading rapidly, but if R_e is less than 1, the spread of the disease is decreasing.

The daily mean of temperature foreach region was also extracted from the DatAC application, which took the data from the Spanish State Meteorological Agency (AEMET, 2023). Subsequently, the variable hourly 2-m specific humidity (kg/kg) was extracted from the European Centre for Medium-Range Weather Forecast ERA5 climate reanalysis (Hersbach et al., 2020). This variable refers to the specific humidity (SH) at two meters above the surface of the land. This information was then used to calculate the daily mean of SH for each region.

Finally, to take into account the effects of the different measures adopted by the governments during the course of the pandemic, the *Stringency Index* (SI) for Spain was obtained from the Oxford COVID-19 Government Response Tracker (OxCGRT) (Hale et al., 2021). The SI consists of a composite measure based on nine response indicators: workplace closures, cancellation of public events, restrictions on public gatherings, closures of public transport, stay-at-home requirements, public information campaigns, restrictions on internal movements and international travel controls.

2.1.1. Data for European countries and Italians regions

The European countries chosen for the study are countries on a similar latitude to Spain and mainly close to the Mediterranean Sea. Specifically, those considered were: France, Portugal, Italy, Greece, Slovenia, Croatia, Serbia and Montenegro. Other countries with similar latitudes to others, such as Albania and North Macedonia, were not chosen due to their inferior reported data quality. This is exemplified by the presence of daily cases equal to 0, which does not correspond to the reality of other countries.

The data for Italy were also analyzed in this study, where 19 out of the 20 regions were considered. Valle d'Aosta was excluded due to its significantly colder climate compared to the remaining regions, with an average temperature below 20ºC.

COVID-19 pandemic data were extracted from the John Hopkins University Coronavirus Resource Centre (Dong et al., 2020) and Our World in Data (Hannah Ritchie and Roser, 2020; Mathieu et al., 2021). As in the case of the communities of Spain, the number of daily COVID-19 cases, the number of daily COVID-19 deaths, the daily percentage of fully COVID-19 vaccinated individuals and the effective reproductive (R_e) from the two different periods were obtained.

Climatological data were downloaded from the European Centre for Medium-Range Weather Forecast ERA5 climate reanalysis (Hersbach et al., 2020). The same meteorological variables were obtained as in the case of Spanish communities.

Finally, again as in the case of Spanish communities, the SI of Italy and the different European countries was obtained from the OxCGRT (Hale et al., 2021).

2.2. Influence of meterological factor in COVID-19 transmission

The study of the influence of meteorological factors' on COVID-19 transmission involved two distinct stages. Firstly, the potential relationship between these factors and the reproduction number (R_e) was tested. Secondly, the significant relationships identified in the first step were quantified. These analyses were performed on the two different periods, P1 and P2.

2.2.1 Relationship between meteorological factors and evolution of R_e

The complex relationship between temperature and R_e was estimated using generalized additive mixed models (GAMs),applied independently for each Spanish region and European country included in the study.

This GAM model can be represented as:

$$\mathrm{Re}_{,i,t} = \alpha + s(MV_{i,t}) + \beta_1(VR_{i,t}) + \beta_2(SI_{i,t}) + f(V_t)$$

Where:

$Re_{i,t}$ is the effective reproductive number (R_e) on a day t in the region i.

$MV_{i,t}$ is the meteorological variable on a day t in the region i.

$VT_{i,t}$ is the vaccination rate on a day t in the region i.

$SI_{i,t}$ is the stringency index on a day t in the region i.

And $f(V_t)$ the dominant variant on a day t.

In this model, the meteorological variable (temperature or specific humidity) is incorporated as a natural cubic spline (s). To account for the potential effect of variants on transmission, a factor representing the dominant variant at each time period is also included in the model. The analysis was performed using the mgcv R package (Wood, 2022, 2017). The p-values obtained were corrected by the Benjamini & Hochberg method (Benjamini and Hochberg, 1995).

To control for the time interval between infection and detection, a lag was applied to the independent variables (lagged variables) included in the model. To account for the distinct

characteristics of the two periods under study, distinct lag days were considered for each period. In P1, the delay between infection and detection was substantial, as indicated by previous studies examining the seasonality of COVID-19 that reported a lag range from 10 to 15 days (Ma et al., 2021; Nottmeyer et al., 2023; Sera et al., 2021). As a result, a lag of 14 days (corresponding to 2 weeks) was employed for P1. In contrast, P2 saw a significant improvement in detection methods, leading to a reduced lag. Thus, a lag of 7 days (equivalent to 1 week) was applied for P2.

2.2.2 Quantification of effect of meteorological factors on risk of contagion

In addition to examining the influence of factors on the evolution of disease transmission, it is crucial to understand the magnitude of this influence.To quantify the evolution of the effect of environmental factors on the risk of infection, a two-stage analysis was conducted, encompassing individual analyses by region followed by an estimation of the global effect.

The first step involved the implementation of a distributed lag non-linear model (DLNM) (Gasparrini et al., 2010) for each region during the two distinct periods (P1 and P2). This models can be represented as:

$$R_{e,i,t} = CB(MV_{i,t}) + CB(SI_{i,t}) + f(V_t) + Ind(Vac) + \text{int} + NS(date, df = 2)$$

Where:

CB(MVi,t) is the cross-basics term for the meteorological variable on a day *t* in the region *i* . It incorporates a lag ranging from 7 to 14 days, allowing for consistent considerations comparable to previous (GAMs).

CB(SIi,t) is the cross-basics term for the SI on a day *t* in the region *i*. Similar to *CB(MVi,t)*, it considers a lag between 7 and 14 days to ensure consistency with the previous GAMs.

f(Vt) is the dominant variant on a day *t*.

Ind(Vac) is a binary variable that represents the lack or presence of vaccination

int is an interaction term for the pre and post vaccination period.

NS(date, df =2) is a term that modulates the intra-period trend of COVID-19 evolution. In this case a natural spline function of the date with 2 degrees of freedom (dfs) is considered, which equals approximately 1 df per three months.

The model was built using the *dlnm* R package (Gasparrini, 2011) and the residual variation of the $R_{e,i,t}$ was assumed to follow a quasi-Poisson distribution. For each region and period, the evolution of the effect of R_e was obtained (association curve), measured by the relative risk (RR) and taking as reference the mean of the meteorological variable.

Secondly, a meta-analysis of the different association curves was applied to obtain the evolution of the global effect by using the *mvmeta* R package with the estimation method of restricted maximum likelihood (REML) (Gasparrini et al., 2012). This enabled us to obtain the global association curve for each meteorological factor,

3. Results

3.1 Evolution of R_e and meteorological factors in Spain during analyzed periods

Throughout the pandemic, the Spanish government reported epidemiological data for various autonomous communities. Fig 1 illustrates the evolution of the effective reproduction number (R_e) across these communities from June 1, 2020 to December 31, 2021. This evolution remained consistent across the different communities. Furthermore, this consistency can also be observed in the distribution of temperature and specific humidity (SH). (Fig 1 and Supplementary Figure S1). Moreover, the distribution of meteorological factors remained homogeneous between the two periods under study, exhibiting similar patterns between the first period with low population immunity (P1) and the second period with higher population immunity due to vaccination (P2). However, no clear correlation can be observed between the evolution of these environmental factors and the evolution of R_e.

On the other hand, a certain relationship can be observed between the distribution of the vaccination rate and the measures adopted by the government. An increase in the vaccination rates concur with a decrease in the Stringency Index (SI) (Supplementary Figure S2). Furthermore, similar to the evolution of R_e and meteorological variables, a comparable distribution of vaccination rates can be observed among the different communities.

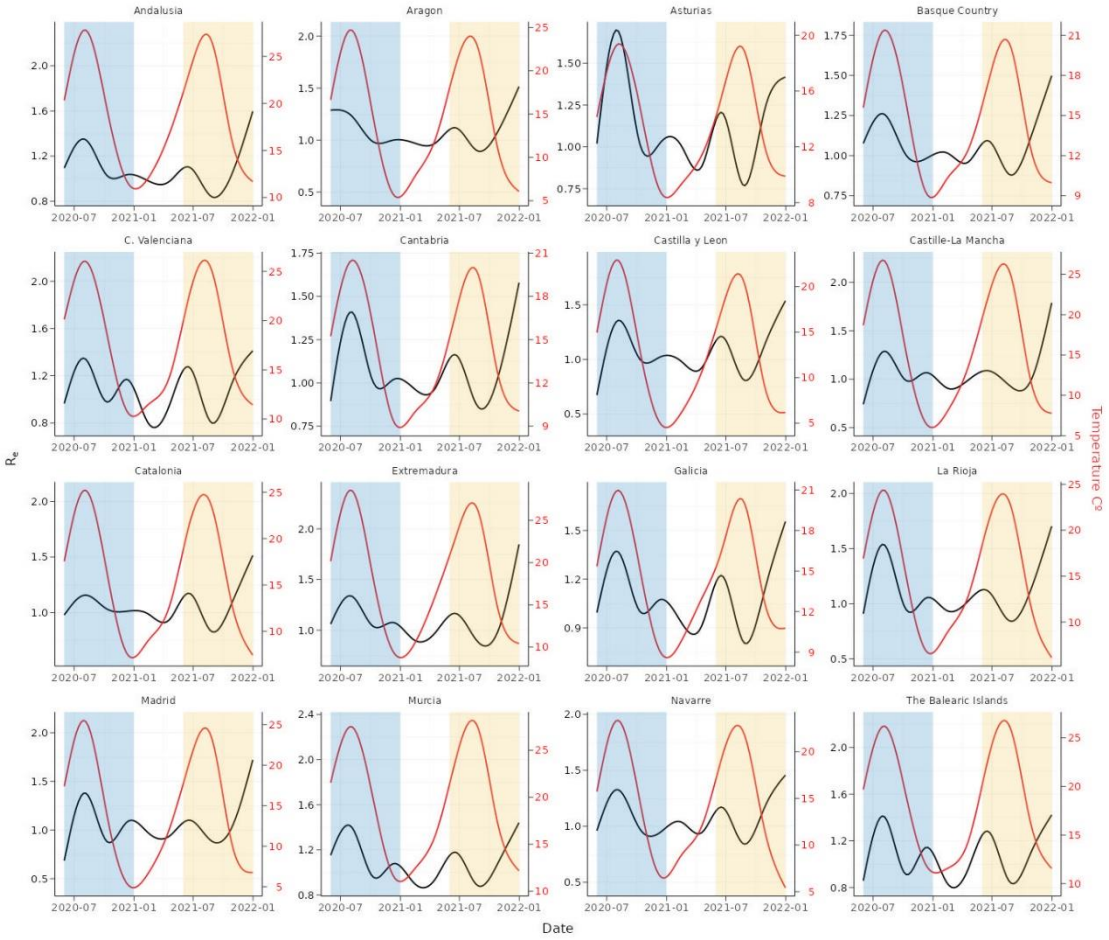


Fig 1. **Distribution of COVID-19 R_e and temperature (ºC)**. Longitudinal plot representing the daily COVID-19 R_e in black and the daily temperature in red line from June 1, 2020 and December 31, 2021 in the different Spanish communities The following periods considered in the study are represented with blue and yellow backgrounds respectively: a first period with a low level of population immunity (P1) and a second period with a high level of population immunity thanks to vaccination (P2). Smoothing has been applied to the R_e and temperature to facilitate visualization.

3.2 Influence of temperature and specific humidity varies depending on population immunity

To assess the impact of meteorological factors on the progression of disease transmission, two types of model were employed.

Firstly, GAMs were utilized to examine whether meteorological factors had any influence on the disease's evolution, differentiating between the two periods. Once the influence of these factors was confirmed based on the vaccinated population, DLNMs were employed to quantitatively determine their effect on the risk of contagion.

3.2.1 Relationship between meteorological factors and transmission of disease in different periods

The relationship between temperature and R_e was estimated using non-generalized additive mixed models (GAMs)., applied for each Spanish region included in the study in two different time periods: P1, with a low level of population immunity (1 June 2020 to 31 December 2020) and P2 (1 June 2021 to 31 December 2021), with a substantial

level of immunization due to the vaccination of the population. The results of the models can be observed in Table 1.

Table 1. Results of the GAM models with temperature as meteorological factor for Spanish communities

P1 (1 June 2020 to 31 December 2020)						
Autonomous community	**Deviance explained**	**R^2**	**Corrected P-values**			
			Temperature	**SI**	**Vaccination Rate**	**Variants**
Andalusia	38.54	0.36	<0.0001	<0.0001	NA	0.0731
Aragon	43.72	0.42	0.0050	<0.0001	NA	0.0413
Cantabria	13.32	0.12	0.0171	0.0002	NA	0.3109
Castilla y Leon	36.98	0.35	<0.0001	0.1917	NA	0.0731
Castille-La Mancha	47.41	0.45	<0.0001	0.7907	NA	0.0614
Catalonia	62.55	0.61	<0.0001	<0.0001	NA	0.1952
Madrid	61.21	0.60	<0.0001	0.3729	NA	0.9196
Navarre	46.07	0.45	0.6030	<0.0001	NA	0.0413
C. Valenciana	29.67	0.28	0.0095	<0.0001	NA	0.2570
Extremadura	23.42	0.21	0.0009	<0.0001	NA	0.0824
Galicia	20.24	0.17	0.0030	<0.0001	NA	0.9196
Balearic Islands	11.44	0.09	0.0046	0.6126	NA	0.0731
La Rioja	38.83	0.37	0.0001	<0.0001	NA	0.0739
Basque Country	35.95	0.35	0.1301	<0.0001	NA	0.9196
Asturias	27.32	0.26	0.5348	<0.0001	NA	0.5804
Murcia	40.01	0.38	0.0001	<0.0001	NA	0.0001
Median	**37.76**	**0.36**	**0.0019**	**<0.0001**	**NA**	0.0782
P2 (1 June 2021 to 31 December 2021)						
Autonomous community	**Deviance explained**	**R^2**	**Corrected p-values**			
			Temperature	**SI**	**Vaccination Rate**	**Variants**
Andalusia	81.79	0.81	<0.0001	<0.0001	<0.0001	<0.0001
Aragon	56.17	0.55	<0.0001	<0.0001	<0.0001	<0.0001
Cantabria	59.54	0.58	<0.0001	<0.0001	<0.0001	<0.0001
Castilla y Leon	60.01	0.59	<0.0001	<0.0001	<0.0001	0.0007
Castille-La Mancha	85.91	0.85	<0.0001	<0.0001	<0.0001	<0.0001
Catalonia	53.58	0.52	<0.0001	<0.0001	<0.0001	0.7814
Madrid	74.87	0.74	<0.0001	<0.0001	<0.0001	0.0001
Navarre	44.45	0.43	<0.0001	<0.0001	<0.0001	0.0011
C. Valenciana	68.00	0.67	<0.0001	<0.0001	<0.0001	<0.0001
Extremadura	76.66	0.76	<0.0001	<0.0001	<0.0001	0.0011
Galicia	72.83	0.72	<0.0001	<0.0001	<0.0001	0.7813
Balearic Islands	70.69	0.70	<0.0001	<0.0001	<0.0001	0.8998
La Rioja	66.19	0.65	<0.0001	<0.0001	<0.0001	0.0023
Basque Country	69.08	0.68	<0.0001	<0.0001	<0.0001	0.0873
Asturias	49.90	0.48	<0.0001	<0.0001	<0.0001	0.7848
Murcia	65.62	0.65	<0.0001	<0.0001	<0.0001	<0.0001
Median	**67.09**	**0.66**	**<0.0001**	**<0.0001**	**<0.0001**	**0.0009**

Note: Results of the GAM models with temperature as meteorological variable for the different Spanish autonomous communities. The first and the second columns represent the deviance explained and the adjusted R^2 of the different models. The remaining the columns represent the corrected p-values for the different variables included in the models.

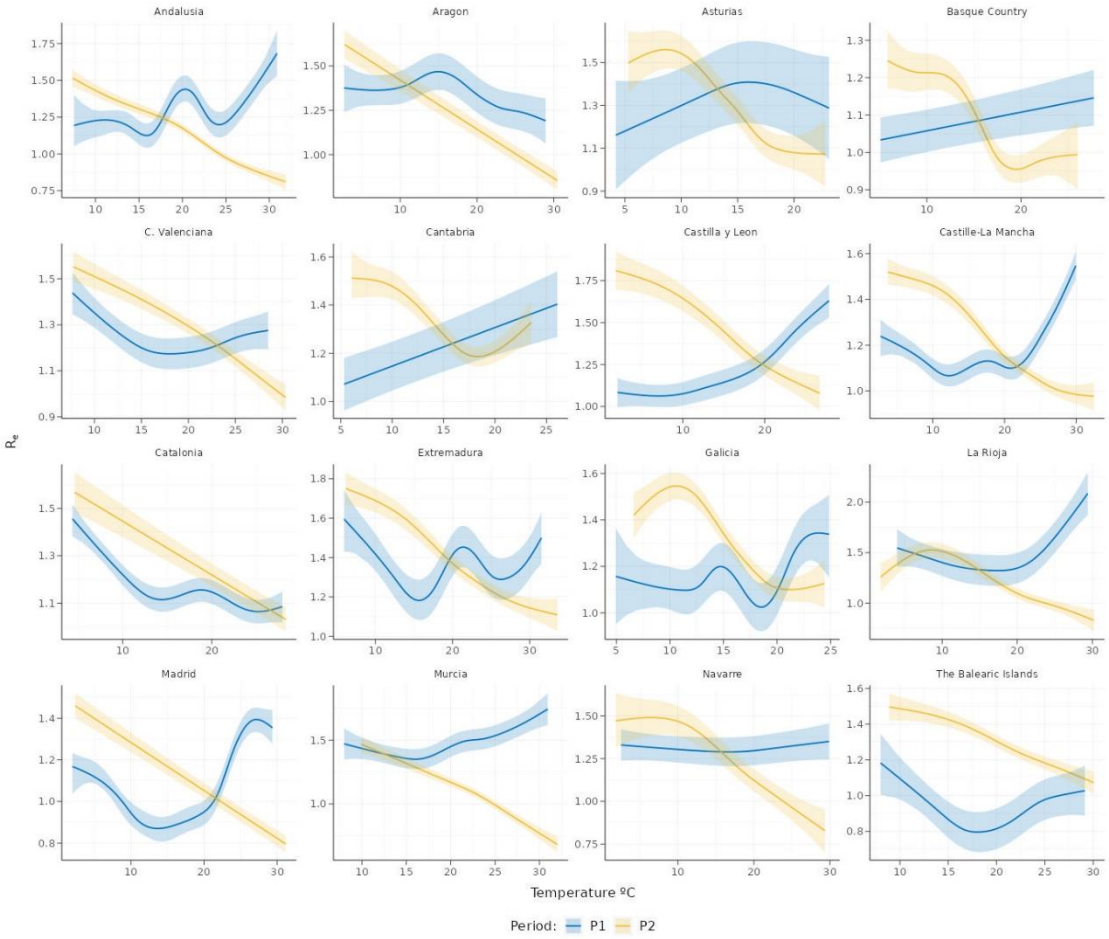


Fig 2. **Estimation of evolution R_e from influence of temperature (ºC) predicted by GAMs insolating rest of variables**. Graphic representation of the R_e predicted by the GAMs with temperature as meteorological variable and considering the rest of variables remain constant for the different Spanish communities. The blue color represents the period without vaccination (P1) and the yellow color represents the period with vaccination (P2).

The results of the models for P1 suggest an association between temperature and transmission in most communities, with the majority of corrected p-values below 0.05 (median of p-values 0.002, as shown in Table 1). However, analyzing the distribution of temperature and transmission for the GAMs. by isolating the influence of other variables, no clear association was observed before vaccination, with a large increase in R_e around 17-20ºC (Fig 2).

During P2, results the from GAM showed that temperature is also associated with transmission in the majority of communities, with a median of corrected p-values of less than 0.001 (Table 1). Unlike the previous period, in most communities, a general decrease in R_e was observed with increasing temperature (Fig 2). However, some northern regions (Cantabria, Basque Country, and Galicia) experienced an increase in transmission above 20ºC, likely due to their milder summers. Overall, virus transmission was lower at warmer temperatures compared to cooler temperatures (Fig 2).

Furthermore, the adjusted R^2 and percentage of explained deviance were considerably higher for P2 than for P1 in most communities (Table 1). This was also reflected in the read data estimation of the modes, which showed a better fit for P2 (Supplementary Figure

S3) than for P1 (Supplementary Figure S4). However, for both periods, the percentage of variability explained was moderate, with most adjusted R^2 not exceeding 0.75, indicating the presence of other factors beyond those included in the models that could affect virus transmission, such as population behavior. Nevertheless, these findings suggest that when a significant portion of the population is vaccinated, higher temperatures may lead to a slight reduction in the transmission of COVID-19.

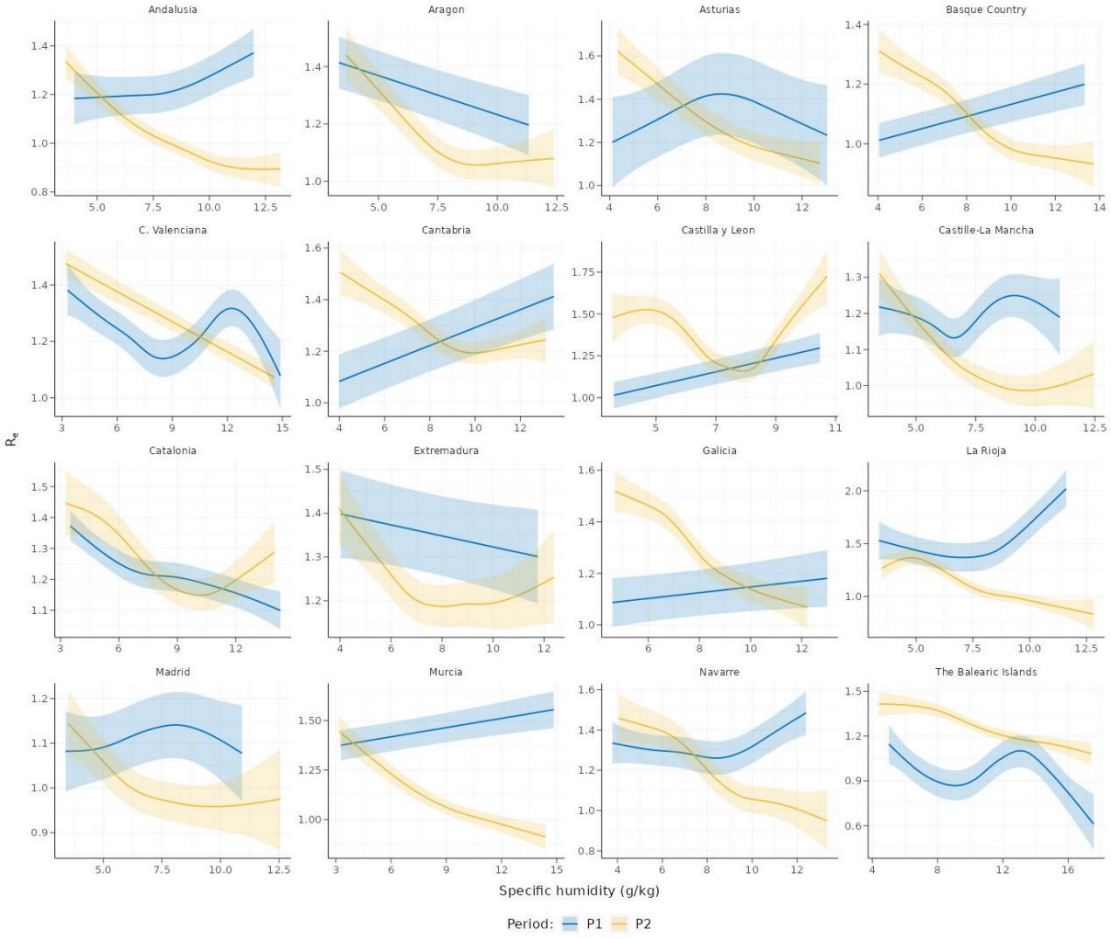


Fig 3. **Estimation of evolution R_e from influence of specific humidity (SH) predicted by GAMs insolating rest of variables**. Graphic representation of the R_e predicted by the GAMs with SH as meteorological variable and considering the rest of variables remain constant for the different Spanish communities. The blue color represents the period without vaccination (P1) and the yellow color represents the period with vaccination (P2).

Table 2. Results of GAMs with SH as meteorological factor for Spanish autonomous communities

P1 (1 June 2020 to 31 December 2020)						
Autonomous community	**Deviance explained**	**R^2**	**Corrected P-values**			
			SH	**SI**	**Vaccination Rate**	**Variants**
Andalusia	21.84	0.20	0.0502	<0.0001	NA	0.2942
Aragon	39.45	0.39	0.0308	<0.0001	NA	0.0669
Cantabria	14.49	0.13	0.0065	0.0008	NA	0.3418
Castilla y Leon	13.87	0.13	0.0011	0.0022	NA	0.2942
Castille-La Mancha	11.79	0.09	0.0400	0.0039	NA	0.0715

Catalonia	58.92	0.58	0.0001	<0.0001	NA	0.0057
Madrid	5.69	0.04	0.4886	0.0053	NA	0.5953
Navarre	48.54	0.47	0.0502	<0.0001	NA	0.0203
C. Valenciana	34.93	0.33	0.0005	<0.0001	NA	0.2229
Extremadura	11.53	0.10	0.3783	<0.0001	NA	0.0669
Galicia	9.41	0.08	0.3783	0.0003	NA	0.9958
Balearic Islands	13.61	0.11	0.0048	0.6563	NA	0.1716
La Rioja	39.98	0.39	<0.0001	<0.0001	NA	0.0376
Basque Country	38.04	0.37	0.0056	<0.0001	NA	0.8169
Asturias	27.78	0.26	0.4017	<0.0001	NA	0.5953
Murcia	31.64	0.31	0.0502	<0.0001	NA	0.0004
Median	24.81	0.23	**0.0354**	**<0.0001**	**NA**	**0.1972**

P2 (1 June 2021 to 31 December 2021)

Autonomous community	**Deviance explained**	**R^2**	**Corrected p-values**			
			SH	**SI**	**Vaccination Rate**	**Variants**
Andalusia	67.47	0.66	<0.0001	<0.0001	<0.0001	0.7620
Aragon	49.26	0.48	<0.0001	<0.0001	<0.0001	0.0146
Cantabria	56.28	0.55	<0.0001	<0.0001	<0.0001	<0.0001
Castilla y Leon	55.56	0.54	<0.0001	<0.0001	<0.0001	0.2524
Castille-La Mancha	75.46	0.75	<0.0001	<0.0001	<0.0001	0.0148
Catalonia	51.73	0.50	0.0003	<0.0001	<0.0001	0.2524
Madrid	64.85	0.64	0.02	<0.0001	<0.0001	0.7620
Navarre	41.92	0.40	<0.0001	<0.0001	<0.0001	0.0217
C. Valenciana	64.65	0.64	<0.0001	<0.0001	<0.0001	<0.0001
Extremadura	64.15	0.63	0.0145	<0.0001	<0.0001	0.7620
Galicia	68.56	0.67	<0.0001	<0.0001	<0.0001	0.7620
Balearic Islands	68.70	0.68	<0.0001	<0.0001	<0.0001	0.5464
La Rioja	61.22	0.60	<0.0001	<0.0001	<0.0001	0.3484
Basque Country	69.30	0.68	<0.0001	<0.0001	<0.0001	0.2037
Asturias	43.87	0.42	<0.0001	<0.0001	<0.0001	0.7620
Murcia	53.37	0.52	<0.0001	<0.0001	<0.0001	0.7620
Median	62.69	0.61	**<0.0001**	**<0.0001**	**<0.0001**	**0.3004**

Note: Results of the GAM models with SH as meteorological variable for the different Spanish autonomous communities. The first and the second columns represent the deviance explained and the adjusted R^2 of the different models. The remaining the columns represent the corrected p-values for the different variables included in the models.

The analysis of specific humidity (SH) revealed that it was a significant factor for most communities for P1 (Table 2). However, deviance explained and adjusted R^2 values were relatively low for most communities during this time period, with median deviance explained of 24.807 and a median adjusted R^2 of 0.2335. As a result, the explained variability in most communities is less than 25% in P1, and the Stringency Index (SI) was the most significant variable among all the communities.

Further investigation into the evolution of R_e from SH after eliminating the influence of other variables revealed no common pattern among communities, as in the case of temperature (Fig. 3). Some communities exhibited greater transmission during dry

periods, while others showed the opposite. Therefore, we did not observe a consistent relationship between SH and COVID-19 transmission during P1.

During P2, SH was also a significant factor for all communities (Table 2). Nevertheless, the explained deviance and adjusted R^2 showed a significant increase, with median values of 62.687 and 0.613, respectively. This suggested that, for this period, the model is able to explain over 60% of the variability for most communities. The predictions of the model also showed a better fit for this period (as seen in Supplementary Figure S5) than for the previous one (as seen in Supplementary Figure S6).

Furthermore, after removing the influence of other variables, a general decrease in transmission is observed in most communities as the humidity level increases. Therefore, when a certain level of population immunity exists, a slight increase in COVID-19 transmission is observed during drier periods.

3.2.2 Quantification of effect of environmental factors on risk of contagion in different periods

In the preceding section, we noted contrasting influences of temperature and humidity during the period with and without vaccination. Nonetheless, the true impact of these variables on the risk of contagion in the different periods remains uncertain. Therefore, to quantify the evolution of the effect of the different environmental factors on the risk of contagion, DLNMs were applied. to each region and period for each of the environmental factors.

Figure 4 shows the overall pooled association curve representing the relative risk (RR) of COVID-19 infection in relation to temperature. This curve was derived from a meta-analysis of the models across different regions. The analysis reveals significant disparities between the two periods examined. During P1, the risk of infection decreases at lower temperatures and increases as the temperature rises. Conversely, in P2, the opposite pattern emerges, with a higher risk of contagion at lower temperatures, which diminishes as temperatures increase. In the case of P2, the probability of contagion during lower temperatures, around 5ºC, is approximately 1.75 times higher (RR = 1.75, CI = [1.63,1.87]) than with respect to the mean global temperature used as a reference, 17.5 ºC. Likewise, at higher temperatures, around 28ºC, during P2, the probability of contagion is 1.61 times lower than with respect to the reference temperature (RR = 0.62, CI = [0.59,0.66]). This consistent trend is observable in the specific association curves for each Spanish region (Supplementary Figure S7).

In the case of SH, a similar trend is observed as with temperature (Supplementary Figure S8 and S9). The overall pooled association curve (Supplementary Figure S8) shows that during P1 the risk of contagion is greater in more humid environments. However, in accordance with the GAMs, in the period with vaccination (P2), it is observed that risk of contagion drops as humidity increases. Specifically, during P2, the probability of contagion during drier environments, around 3.3 g/kg, is approximately 1.72 times higher (RR = 1.72, CI = [1.62, 1.81]) than that in respect to the mean global SH used as a reference, 8.12 g/kg. Likewise, in humid environments, around 12 g/kg, during P2, the probability of contagion is 1.54 times lower than that regards to the reference SH (RR = 0.65, CI = [0.99, 1.31]. This same evolution is observed in the specific association curves

of each autonomous community, in which for P1 the RR increases or does not vary in P1 as a function of humidity, while in P2, in a general way, this risk decreases. to higher humidity.

These findings align with the results obtained from our previous GAM models, affirming that higher temperatures and humid conditions contribute to a marginal reduction in the risk of contagion when considerable percentage of people are immunized.

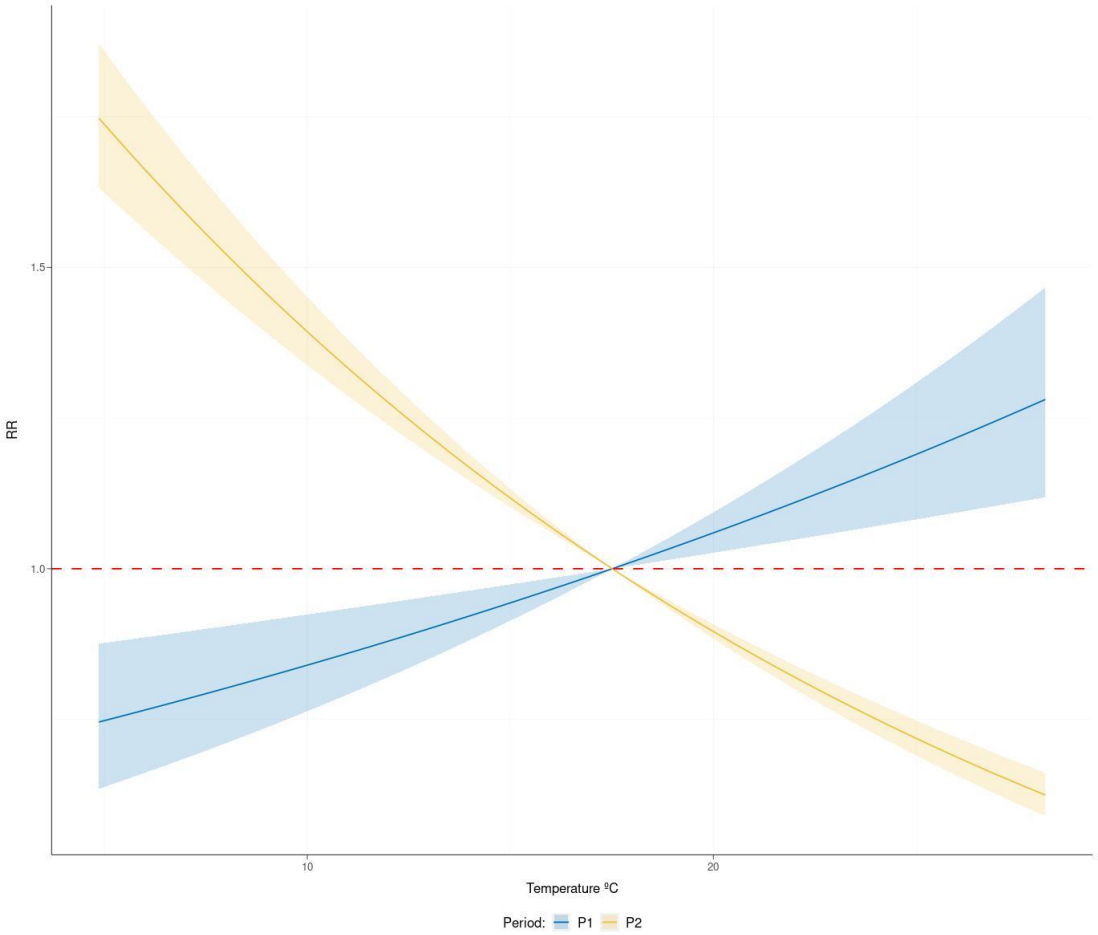


Fig 4. **Global pooled association curve for temperature.** Graphic representation of the evolution of Relative Risk (RR) of COVID-19 infection depending on the temperature obtained by the meta-analysis of the different region's models. The global mean temperature (17.5 ºC) is used as a reference. The blue color represents the period without vaccination (P1) and the yellow color represents the period with vaccination (P2).

3.3 Association of temperature and specific humidity with R_e in other countries

To validate our findings on the association of temperature and relative humidity with R_e we tested our models on countries with similar latitudes to Spain (Supplementary Tables S1 and S2, finding significant heterogeneity in the results (Supplementary Figures S10 and S11). When examining the influence of temperature and SH on R_e while controlling for other variables, we observed similar patterns to Spain in some countries, such as France, Portugal, Greece, and Italy. In contrast, in countries with lower vaccination rates in P2, such as Macedonia and Serbia, the observed patterns were almost the opposite. This difference may be mainly attributed to the much lower vaccination rate in P2 in these

countries than in Spain, France, Portugal, Greece, and Italy, where vaccination rates were similar (Supplementary Figure S12). These trends are also observed when we quantify the effect of the meteorological factors, finding that the specific association curves (Supplementary S13 and S14) for France, Portugal, Greece and Italy show similar tendency to the regions of Spain.

As Italy globally shows a similar pattern to that of Spain, to further validate our findings, we applied the same models to different regions in Italy, which have a similar climate and latitude to their Spanish counterpart. For temperature, we found comparable results in the GAMs between Italian and Spanish regions (Supplementary Table S3 and Supplementary Figure S15). While temperature is significant in the model for P1, we observed an increase in transmission as the temperature rose in several regions, with a relatively low percentage of variability explained during this time. However, for the period with vaccination, the temperature is significant in almost all regions, except for Molise. We also observed a decrease in transmission with increasing temperature for most regions in the model. In a similar way to Spain, the explained deviance and adjusted R^2 also increased considerably, with over 50% of the variability being explained for the majority of the regions. A similar pattern emerges from the DLNMs applied for each of the regions of Italy [Supplementary Figure S16]. Only in P2is a reduction in the risk of contagion observed as the temperature increases.

Regarding SH, we observed similar patterns compared to temperature (Supplementary Figure S17, S18 and Supplementary Table S4). SH is only significant for P2, and greater transmission of the disease is observed during drier periods.

Overall, the results for Italian regions confirmed our findings for Spain. To observe the potential seasonality of the disease, a certain population must be vaccinated. In this scenario, more risk of contagion is seen in cold and dry periods.

4. Discussion and conclusions

The impact of environmental factors on COVID-19 transmission has received significant attention among researchers due to its medical, political, and social implications. However, the findings of various studies exploring the relationship between the transmission of the virus and different environmental factors have been conflicting. Such changeability in results may stem from differences in methodology and variables considered, the presence of confounding factors, period studied, or insufficient data (Nottmeyer et al., 2023).

Being a respiratory disease, transmission of COVID-19 may increase in cool and dry conditions, similar to other respiratory diseases like influenza (Baker et al., 2018; Lowen et al., 2007; Lowen and Steel, 2014). Previous studies have suggested that population immunity is a significant confounding factor that may influence the impact of seasonality in SARS-CoV-2 transmission (Baker et al., 2020; Telenti et al., 2021). The initial absence of population immunity during the start of the pandemic may have resulted in unreliable data and findings in non-immunized populations, making it challenging to determine the seasonal patterns of the virus accurately. Thus, the time frame selected for studying the effect of environmental factors on the transmission of SARS-CoV-2 is crucial for the accuracy of results and the reliability of the conclusions.

To consider the impact of population immunity on the association between weather and virus transmission we analyzed two separate time periods. The first period, from June to December 2020, saw a low proportion of the population being vaccinated and hence, low population immunity. In contrast, the second period from June to December 2021 had a high proportion of the population vaccinated with two or more doses, resulting in a high level of population immunity. Additionally, the two periods were selected to include the same months of the year, allowing for a more accurate comparison of the weather conditions. Furthermore, to control other potential confounding factors that could impact the outcomes, the models incorporated variables that took into account population mobility and severity of measures implemented by governments.

Our analysis revealed that population immunity influences the relationships between temperature, SH and COVID-19 transmission. During the period of low population immunity (June-December 2020), our results indicate that an increase in temperature does not lead to a decrease in virus transmission and may even be associated with increased transmission. These findings are consistent with previous studies conducted early in the pandemic, which reported either a lack of correlation or a positive correlation between temperature and transmission of COVID-19 (Bashir et al., 2020; Briz-Redón and Serrano-Aroca, 2020, p.; Meyer et al., 2020).

However, during the period when a high proportion of the population were vaccinated, the results suggest a different trend, indicating that temperature and specific humidity do have an impact on SARS-CoV-2 transmission in accordance with prior studies (Baker et al., 2020; Telenti et al., 2021). Specifically, a slight reduction in the risk of contagion was noted with higher temperatures and less dry environment.

We wish to point out some limitations of this study. Firstly, the scope of our analysis is restricted to individuals who received a minimum of two doses of the COVID-19 vaccine, thereby disregarding those who may have acquired immunity from contracting the virus. Additionally, we have not accounted for the variable duration of vaccine-induced immunity, which may influence the definition of what constitutes effective immunization (Lopman et al., 2021). Nevertheless, it is notable that a decline in COVID-19 immunity has been reported starting from the first month after vaccination (Addo et al., 2022), with immunity largely diminishing by the sixth month, a timeframe which corresponds to the duration of our study periods.

Secondly, another limitation of our study is that it primarily concentrates on Spain, Italy, and other Mediterranean European countries, which are generally known for having warm summers and mild winters. Weather conditions, however, may have varying impacts on the transmission of the virus in other regions, such as tropical or colder areas, as occurs with other viruses such as influenza (Baker et al., 2019; Tamerius et al., 2013). Therefore, to gain a comprehensive understanding of how different climates affect the transmission of the disease, it is necessary to conduct a more extensive analysis in different climatic regions, in which the temperature and humidity do not follow seasonal patterns.

Thirdly, an additional potential limitation of our study is that it only considers SH and temperature, while there are other environmental factors that may affect the seasonality of the virus and that better explain the seasonal patterns in other climatic zones (Tamerius et al., 2013). For instance, certain factors like allergens or daily sunlight duration have

been explored in previous studies as potential explanations for this seasonal pattern (Abraham et al., 2021; Hoogeveen et al., 2022; Shah et at., 2021). In future investigations, it would be valuable to incorporate these variables for a more comprehensive analysis and to account for other potentially relevant factors associated with the seasonality of the disease.

Finally, it is important to note that in the case of our study it was not possible to analyze other variables due to the unavailability of comprehensive pandemic-related information in some countries. For instance, in Spain, the collection of pandemic-related data was halted in 2022, which limits the completeness of our analysis. Furthermore, due to the lack of these data, the periods analyzed do not include complete annual cycles, which may have influence on our results. To perform a more compressive analysis of the impact of environmental factors on the disease, it is critical that countries provide complete and up-to-date information on the progression of the pandemic.

Our results suggest that meteorological factors might affect COVID-19 transmission, which may be slightly reduced during warmer periods when there is a substantial proportion of the population immunized. Despite some limitations, this study represents a novel approach in exploring the relationship between COVID-19 seasonality and population immunity. Our results suggest that temperature and specific humidity have a differential effect on virus transmission, with the effects observed being a result of population immunity. Future studies incorporating more data and longer periods of immunity are expected to further clarify the relationship between seasonality and COVID-19 transmission. The insights gained from this study provide valuable information for public health and disease management strategies.

Acknowledgements

This work is part of the thesis of Juan Antonio Villatoro-García'. Juan Antonio Villatoro-García is enrolled on the PhD program in Mathematical and Applied Statistics at the University of Granada, Spain.

Funding sources

This work was funded by grant P20-00335 from Consejería de Universidad, Investigación e Innovación-Junta de Andalucía and FEDER-"Una manera de Hacer Europa" and MCIN/AEI/10.13039/501100011033 [grant number PID2020-119032RB-I00]. JAVG is funded by the Teaching Staff Programme, implemented by the Ministerio de Universidades [grant number FPU19/01999].

Author contributions

Juan Antonio Villatoro-García: Methodology, Formal analysis, Data Curation, Writing - Review & Editing. **Jordi Martorell-Marugán**: Validation, Writing - Review & Editing. **Raúl López Domínguez**: Data curation, Writing - Review & Editing. **Juan de Dios Luna**: Methodology, Writing - Review & Editing, **Jose Antonio Lorente:** Results interpretation, Writing - Review & Editing**, Pedro Carmona-Sáez**: Conceptualization, Supervision, Methodology, Funding acquisition, Original Draft, Writing - Review & Editing

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Pedro Carmona-Sáez.

Supplementary information

Supplementary_Information.docx: Word file with Supplementary Figures S1 to S11 and Supplementary Tables S1 to S4.

Data availability

The data employed in this article is available at: https://github.com/GENyO-BioInformatics/Covid19_Seasonality.

Code availability

Code for recreating the results is available at: https://github.com/GENyO-BioInformatics/Covid19_Seasonality.

References

Abraham, J., Dowling, K., Florentine, S., 2021. Can Optimum Solar Radiation Exposure or Supplemented Vitamin D Intake Reduce the Severity of COVID-19 Symptoms? Int. J. Environ. Res. Public. Health 18, 740. https://doi.org/10.3390/ijerph18020740

Addo, I.Y., Dadzie, F.A., Okeke, S.R., Boadi, C., Boadu, E.F., 2022. Duration of immunity following full vaccination against SARS-CoV-2: a systematic review. Archives of Public Health 80, 200. https://doi.org/10.1186/s13690-022-00935-x

AEMET, 2023. AEMET OpenData [WWW Document]. URL https://opendata.aemet.es/centrodedescargas/inicio (accessed 2.8.23).

Baker, R.E., Mahmud, A.S., Metcalf, C.J.E., 2018. Dynamic response of airborne infections to climate change: predictions for varicella. Climatic Change 148, 547–560. https://doi.org/10.1007/s10584-018-2204-4

Baker, R.E., Mahmud, A.S., Wagner, C.E., Yang, W., Pitzer, V.E., Viboud, C., Vecchi, G.A., Metcalf, C.J.E., Grenfell, B.T., 2019. Epidemic dynamics of respiratory syncytial virus in current and future climates. Nat Commun 10, 5512. https://doi.org/10.1038/s41467-019-13562-y

Baker, R.E., Yang, W., Vecchi, G.A., Metcalf, C.J.E., Grenfell, B.T., 2021. Assessing the influence of climate on wintertime SARS-CoV-2 outbreaks. Nat Commun 12, 846. https://doi.org/10.1038/s41467-021-20991-1

Baker, R.E., Yang, W., Vecchi, G.A., Metcalf, C.J.E., Grenfell, B.T., 2020. Susceptible supply limits the role of climate in the early SARS-CoV-2 pandemic. Science 369, 315–319. https://doi.org/10.1126/science.abc2535

Bashir, M.F., Ma, B., Bilal, Komal, B., Bashir, M.A., Tan, D., Bashir, M., 2020. Correlation between climate indicators and COVID-19 pandemic in New York, USA. Sci Total Environ 728, 138835. https://doi.org/10.1016/j.scitotenv.2020.138835

Benjamini, Y., Hochberg, Y., 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical

Society: Series B (Methodological) 57, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Briz-Redón, Á., Serrano-Aroca, Á., 2020. A spatio-temporal analysis for exploring the effect of temperature on COVID-19 early evolution in Spain. Science of The Total Environment 728, 138811. https://doi.org/10.1016/j.scitotenv.2020.138811

Carlson, C.J., Gomez, A.C.R., Bansal, S., Ryan, S.J., 2020. Misconceptions about weather and seasonality must not misguide COVID-19 response. Nat Commun 11, 4312. https://doi.org/10.1038/s41467-020-18150-z

Chatterjee, P., 2020. Is India missing COVID-19 deaths? The Lancet 396, 657. https://doi.org/10.1016/S0140-6736(20)31857-2

Cori, A., Ferguson, N.M., Fraser, C., Cauchemez, S., 2013. A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics. American Journal of Epidemiology 178, 1505–1512. https://doi.org/10.1093/aje/kwt133

D'Amico, F., Marmiere, M., Righetti, B., Scquizzato, T., Zangrillo, A., Puglisi, R., Landoni, G., 2022. COVID-19 seasonality in temperate countries. Environ Res 206, 112614. https://doi.org/10.1016/j.envres.2021.112614

Datatista [WWW Document], 2022. . GitHub. URL https://github.com/datadista/datasets (accessed 1.17.23).

Dong, E., Du, H., Gardner, L., 2020. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis 20, 533–534. https://doi.org/10.1016/S1473-3099(20)30120-1

Dong, Z., Fan, X., Wang, J., Mao, Y., Luo, Y., Tang, S., 2021. Data-related and methodological obstacles to determining associations between temperature and COVID-19 transmission. Environ. Res. Lett. 16, 034016. https://doi.org/10.1088/1748-9326/abda71

Fontal, A., Bouma, M.J., San-José, A., López, L., Pascual, M., Rodó, X., 2021. Climatic signatures in the different COVID-19 pandemic waves across both hemispheres. Nat Comput Sci 1, 655–665. https://doi.org/10.1038/s43588-021-00136-6

Gasparrini, A., 2011. Distributed Lag Linear and Non-Linear Models in R: The Package dlnm. J. Stat. Softw. 43, 1–20.

Gasparrini, A., Armstrong, B., Kenward, M.G., 2012. Multivariate meta-analysis for non-linear and other multi-parameter associations. Stat. Med. 31, 3821–3839. https://doi.org/10.1002/sim.5471

Gasparrini, A., Armstrong, B., Kenward, M.G., 2010. Distributed lag non-linear models. Stat. Med. 29, 2224–2234. https://doi.org/10.1002/sim.3940

Hale, T., Angrist, N., Goldszmidt, R., Kira, B., Petherick, A., Phillips, T., Webster, S., Cameron-Blake, E., Hallas, L., Majumdar, S., Tatlow, H., 2021. A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). Nat Hum Behav 5, 529–538. https://doi.org/10.1038/s41562-021-01079-8

Hannah Ritchie, D.B., Edouard Mathieu, Lucas Rodés-Guirao, Cameron Appel, Charlie Giattino, Esteban Ortiz-Ospina, Joe Hasell, Bobbie Macdonald, Roser, M., 2020. Coronavirus Pandemic (COVID-19). Our World in Data.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R.J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti,

G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., Thépaut, J.-N., 2020. The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society 146, 1999–2049. https://doi.org/10.1002/qj.3803

Hoogeveen, M.J., Kroes, A.C.M., Hoogeveen, E.K., 2022. Environmental factors and mobility predict COVID-19 seasonality in the Netherlands. Environ Res 211, 113030. https://doi.org/10.1016/j.envres.2022.113030

Kassem, A.Z.E., 2020. Does Temperature Affect COVID-19 Transmission? Frontiers in Public Health 8.

Liu, Mengyang, Li, Z., Liu, Mengmeng, Zhu, Y., Liu, Y., Kuetche, M.W.N., Wang, J., Wang, X., Liu, X., Li, X., Wang, W., Guo, X., Tao, L., 2022. Association between temperature and COVID-19 transmission in 153 countries. Environ Sci Pollut Res 29, 16017–16027. https://doi.org/10.1007/s11356-021-16666-5

Lopman, B.A., Shioda, K., Nguyen, Q., Beckett, S.J., Siegler, A.J., Sullivan, P.S., Weitz, J.S., 2021. A framework for monitoring population immunity to SARS-CoV-2. Ann Epidemiol 63, 75–78. https://doi.org/10.1016/j.annepidem.2021.08.013

Lowen, A.C., Mubareka, S., Steel, J., Palese, P., 2007. Influenza Virus Transmission Is Dependent on Relative Humidity and Temperature. PLOS Pathogens 3, e151. https://doi.org/10.1371/journal.ppat.0030151

Lowen, A.C., Steel, J., 2014. Roles of humidity and temperature in shaping influenza seasonality. J Virol 88, 7692–7695. https://doi.org/10.1128/JVI.03544-13

Ma, Y., Pei, S., Shaman, J., Dubrow, R., Chen, K., 2021. Role of meteorological factors in the transmission of SARS-CoV-2 in the United States. Nat Commun 12, 3602. https://doi.org/10.1038/s41467-021-23866-7

Martinez, M.E., 2018. The calendar of epidemics: Seasonal cycles of infectious diseases. PLOS Pathogens 14, e1007327. https://doi.org/10.1371/journal.ppat.1007327

Martorell-Marugán, J., Villatoro-García, J.A., García-Moreno, A., López-Domínguez, R., Requena, F., Merelo, J.J., Lacasaña, M., de Dios Luna, J., Díaz-Mochón, J.J., Lorente, J.A., Carmona-Sáez, P., 2021. DatAC: A visual analytics platform to explore climate and air quality indicators associated with the COVID-19 pandemic in Spain. Sci Total Environ 750, 141424. https://doi.org/10.1016/j.scitotenv.2020.141424

Mathieu, E., Ritchie, H., Ortiz-Ospina, E., Roser, M., Hasell, J., Appel, C., Giattino, C., Rodés-Guirao, L., 2021. A global database of COVID-19 vaccinations. Nat Hum Behav 5, 947–953. https://doi.org/10.1038/s41562-021-01122-8

Mecenas, P., Bastos, R.T. da R.M., Vallinoto, A.C.R., Normando, D., 2020. Effects of temperature and humidity on the spread of COVID-19: A systematic review. PLoS One 15, e0238339. https://doi.org/10.1371/journal.pone.0238339

Meyer, A., Sadler, R., Faverjon, C., Cameron, A.R., Bannister-Tyrrell, M., 2020. Evidence That Higher Temperatures Are Associated With a Marginally Lower Incidence of COVID-19 Cases. Frontiers in Public Health 8.

Moriyama, M., Hugentobler, W.J., Iwasaki, A., 2020. Seasonality of Respiratory Viral Infections. Annu Rev Virol 7, 83–101. https://doi.org/10.1146/annurev-virology-012420-022445

Nottmeyer, L., Armstrong, B., Lowe, R., Abbott, S., Meakin, S., O'Reilly, K.M., von Borries, R., Schneider, R., Royé, D., Hashizume, M., Pascal, M., Tobias, A., Vicedo-Cabrera, A.M., Lavigne, E., Correa, P.M., Ortega, N.V., Kynčl, J., Urban, A., Orru, H., Ryti, N., Jaakkola, J., Dallavalle, M., Schneider, A., Honda, Y., Ng, C.F.S., Alahmad, B., Carrasco-Escobar, G., Holobâc, I.H., Kim, H., Lee, W., Íñiguez, C., Bell, M.L., Zanobetti, A., Schwartz, J., Scovronick, N., Coélho, M. de S.Z.S., Saldiva, P.H.N., Diaz, M.H., Gasparrini, A., Sera, F., 2023. The

association of COVID-19 incidence with temperature, humidity, and UV radiation – A global multi-city analysis. Science of The Total Environment 854, 158636. https://doi.org/10.1016/j.scitotenv.2022.158636

O'Reilly, K.M., Auzenbergs, M., Jafari, Y., Liu, Y., Flasche, S., Lowe, R., 2020. Effective transmission across the globe: the role of climate in COVID-19 mitigation strategies. Lancet Planet Health 4, e172. https://doi.org/10.1016/S2542-5196(20)30106-6

Pan, J., Yao, Y., Liu, Z., Meng, X., Ji, J.S., Qiu, Y., Wang, Weidong, Zhang, L., Wang, Weibing, Kan, H., 2021. Warmer weather unlikely to reduce the COVID-19 transmission: An ecological study in 202 locations in 8 countries. Science of The Total Environment 753, 142272. https://doi.org/10.1016/j.scitotenv.2020.142272

Pifarré i Arolas, H., Vidal-Alaball, J., Gil, J., López, F., Nicodemo, C., Saez, M., 2021. Missing Diagnoses during the COVID-19 Pandemic: A Year in Review. International Journal of Environmental Research and Public Health 18, 5335. https://doi.org/10.3390/ijerph18105335

Sera, F., Armstrong, B., Abbott, S., Meakin, S., O'Reilly, K., von Borries, R., Schneider, R., Royé, D., Hashizume, M., Pascal, M., Tobias, A., Vicedo-Cabrera, A.M., Gasparrini, A., Lowe, R., 2021. A cross-sectional analysis of meteorological factors and SARS-CoV-2 transmission in 409 cities across 26 countries. Nat Commun 12, 5968. https://doi.org/10.1038/s41467-021-25914-8

Shah, R.B., Shah, R.D., Retzinger, D.G., Retzinger, A.C., Retzinger, D.A., Retzinger, G.S., 2021. Competing Bioaerosols May Influence the Seasonality of Influenza-Like Illnesses, including COVID-19. The Chicago Experience. Pathogens 10, 1204. https://doi.org/10.3390/pathogens10091204

Smit, A.J., Fitchett, J.M., Engelbrecht, F.A., Scholes, R.J., Dzhivhuho, G., Sweijd, N.A., 2020. Winter Is Coming: A Southern Hemisphere Perspective of the Environmental Drivers of SARS-CoV-2 and the Potential Seasonality of COVID-19. Int J Environ Res Public Health 17, 5634. https://doi.org/10.3390/ijerph17165634

Spanish Ministry of Health [WWW Document], 2023. URL https://cnecovid.isciii.es/covid19/ (accessed 1.17.23).

Tamerius, J.D., Shaman, J., Alonso, W.J., Bloom-Feshbach, K., Uejio, C.K., Comrie, A., Viboud, C., 2013. Environmental predictors of seasonal influenza epidemics across temperate and tropical climates. PLoS Pathog 9, e1003194. https://doi.org/10.1371/journal.ppat.1003194

Telenti, A., Arvin, A., Corey, L., Corti, D., Diamond, M.S., García-Sastre, A., Garry, R.F., Holmes, E.C., Pang, P.S., Virgin, H.W., 2021. After the pandemic: perspectives on the future trajectory of COVID-19. Nature 596, 495–504. https://doi.org/10.1038/s41586-021-03792-w

Villeneuve, P.J., Goldberg, M.S., 2020. Methodological Considerations for Epidemiological Studies of Air Pollution and the SARS and COVID-19 Coronavirus Outbreaks. Environmental Health Perspectives 128, 095001. https://doi.org/10.1289/EHP7411

Weaver, A.K., Head, J.R., Gould, C.F., Carlton, E.J., Remais, J.V., 2022. Environmental Factors Influencing COVID-19 Incidence and Severity. Annual Review of Public Health 43, 271–291. https://doi.org/10.1146/annurev-publhealth-052120-101420

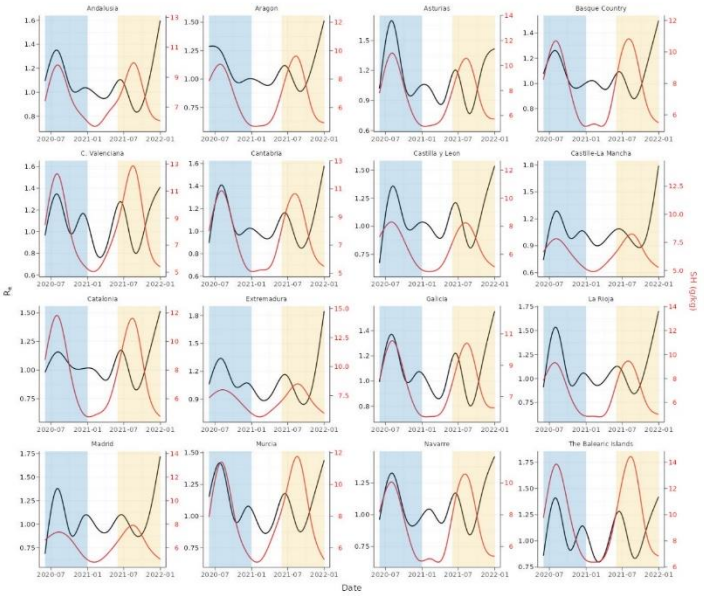
Wood, S., 2022. mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation.

Wood, S.N., 2017. Generalized Additive Models: An Introduction with R, Second Edition, 2nd ed. Chapman and Hall/CRC, New York. https://doi.org/10.1201/9781315370279

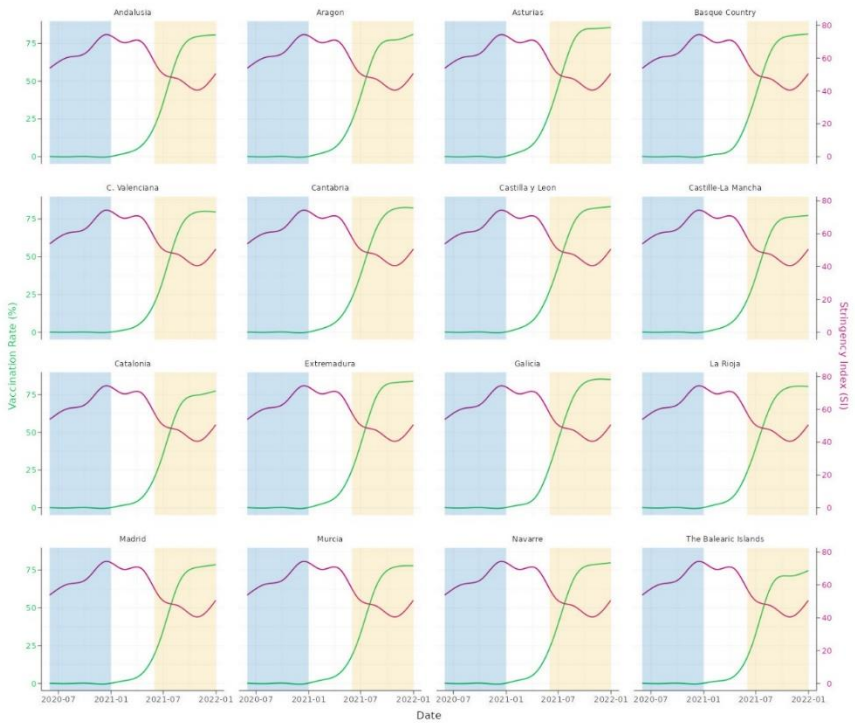
Yamasaki, L., Murayama, H., Hashizume, M., 2021. The impact of temperature on the transmissibility and virulence of COVID-19 in Tokyo, Japan. Sci Rep 11, 24477. https://doi.org/10.1038/s41598-021-04242-3

Yin, C., Zhao, W., Pereira, P., 2022. Meteorological factors' effects on COVID-19 show seasonality and spatiality in Brazil. Environ Res 208, 112690. https://doi.org/10.1016/j.envres.2022.112690

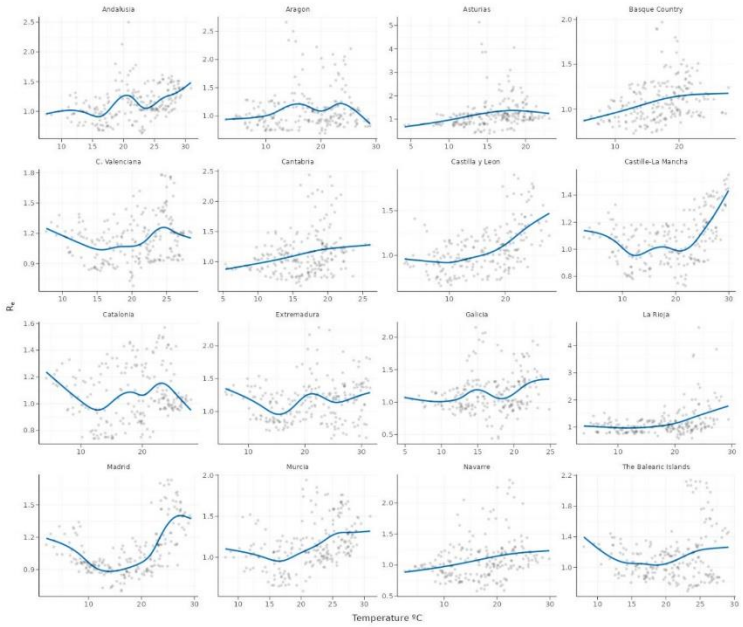
Supplementary Figures



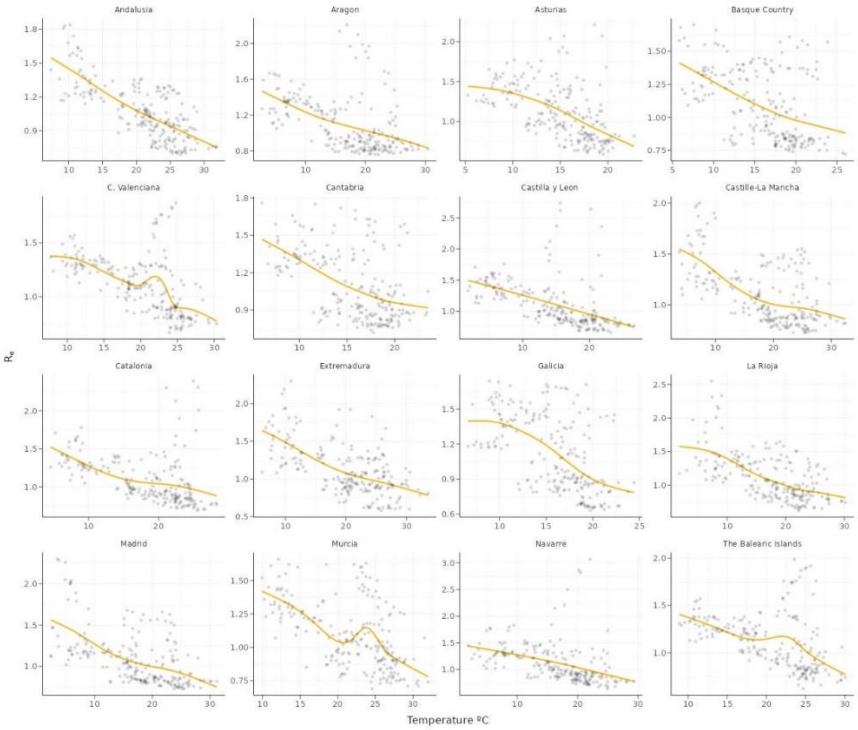
Supplementary Figure S1. **Distribution of COVID-19 R_e and the specific humidity (SH) (g/kg)**. Longitudinal plot representing the daily COVID-19 R_e in black and the daily SH in red line from June 1, 2020 and December 31, 2021 in the different Spanish communities. Blue and Yellow backgrounds represent the considered periods in the study: a first period with a low level of population immunity (P1) and a second period with an important level of population immunity thanks to the vaccination (P2). Smoothing has been applied to the R_e and SH to facilitate the visualization.



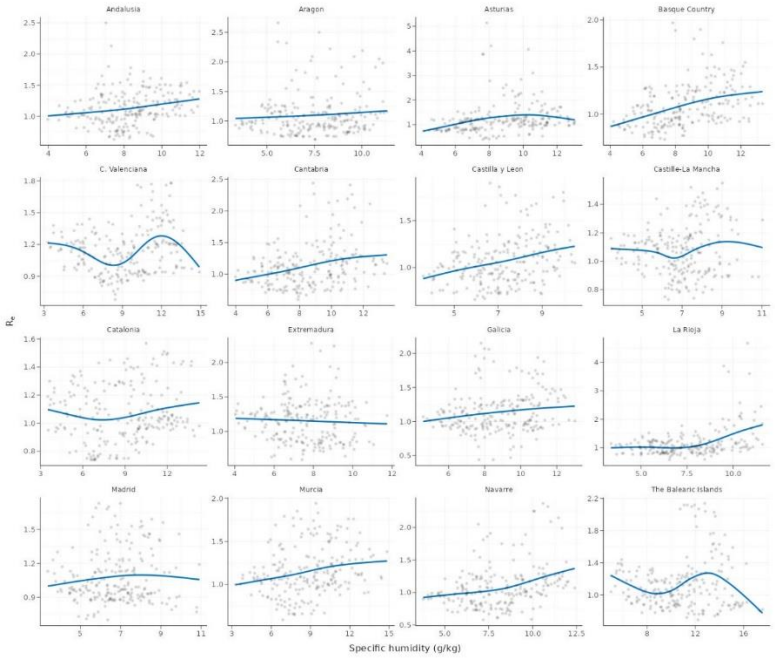
Supplementary Figure S2. **Distribution vaccination rates (%) and Stringency Index (SI) of the different Spanish communities**. Longitudinal plot representing the daily vaccination rates in black and the daily SI in red line from June 1, 2020 and December 31, 2021 in the different Spanish communities. Blue and Yellow backgrounds represent the considered periods in the study: a first period with a low level of population immunity (P1) and a second period with an important level of population immunity thanks to the vaccination (P2). Smoothing has been applied to the vaccination rates and SI to facilitate the visualization.



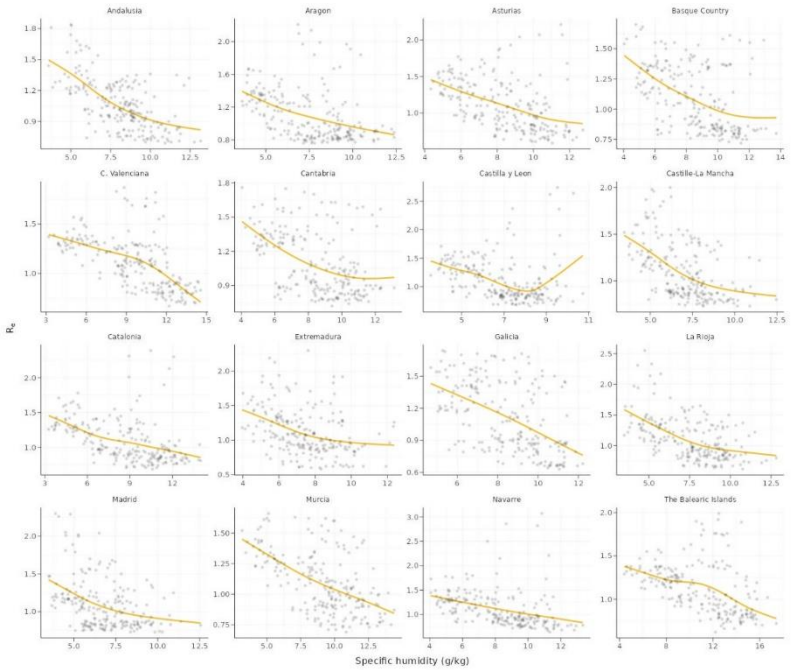
Supplementary Figure S3. **Adjustment of the GAM models with temperature as meteorological factor for the real data for each of the Spanish communities for P1**. Graphical representation of the estimation of R_e with temperature as a meteorological factor for the period without vaccination. The blue line represents the estimate for the R_e of the model. Each of the points (gray) is one of the real observations.



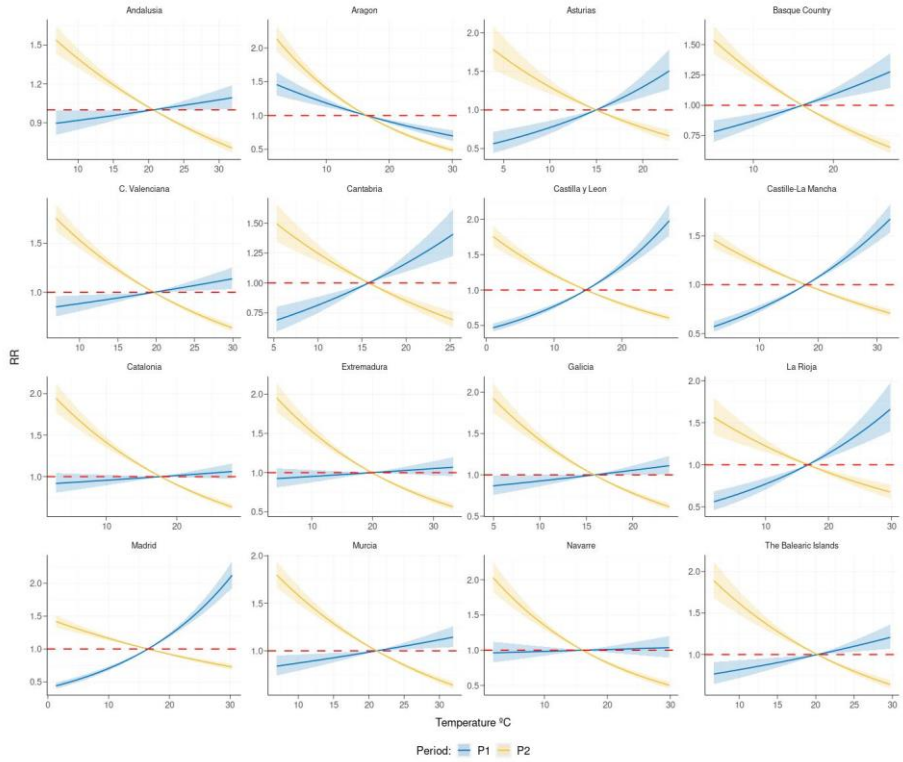
Supplementary Figure S4. Adjustment of the GAM models with temperature as meteorological factor for the real data for each of the Spanish communities for P2. Graphical representation of the estimation of R_e with temperature as a meteorological factor for the period with vaccination. The blue line represents the estimate for the R_e of the model. Each of the points (gray) is one of the real observations.



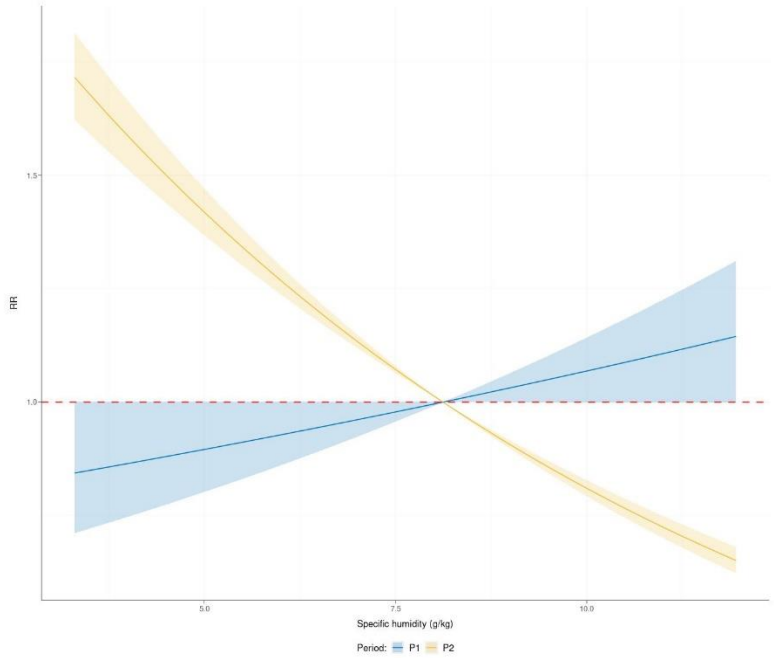
Supplementary Figure S5. **Adjustment of the GAM models with SH as meteorological factor for the real data for each of the Spanish communities for P1**. Graphical representation of the estimation of R_e with SH as a meteorological factor for the period without vaccination. The blue line represents the estimate for the R_e of the model. Each of the points (gray) is one of the real observations.



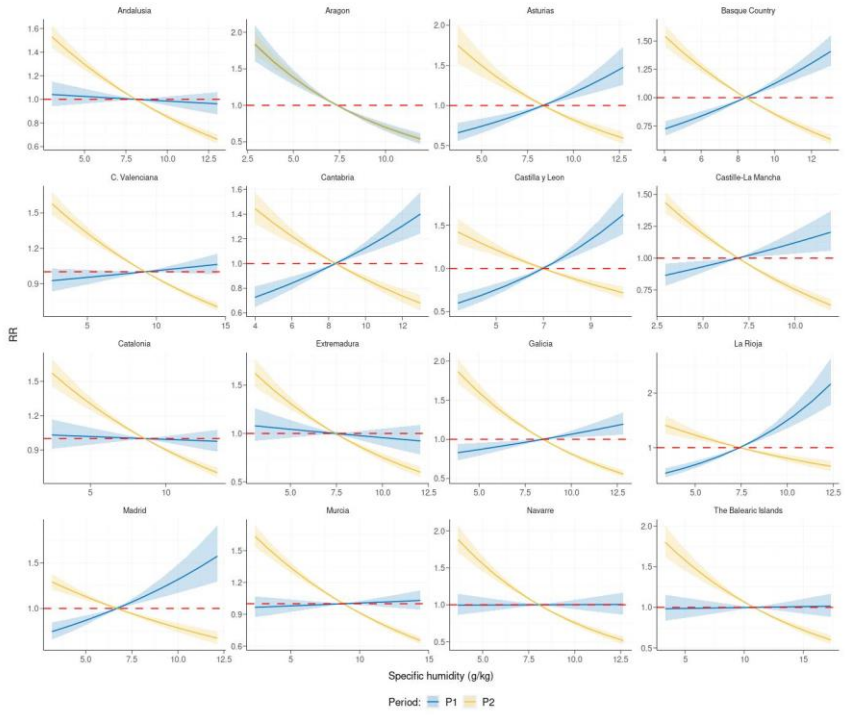
Supplementary Figure S6. Adjustment of the GAM models with SH as meteorological factor for the real data for each of the Spanish communities for P2. Graphical representation of the estimation of R_e with SH as a meteorological factor for the period with vaccination. The blue line represents the estimate for the R_e of the model. Each of the points (gray) is one of the real observations.



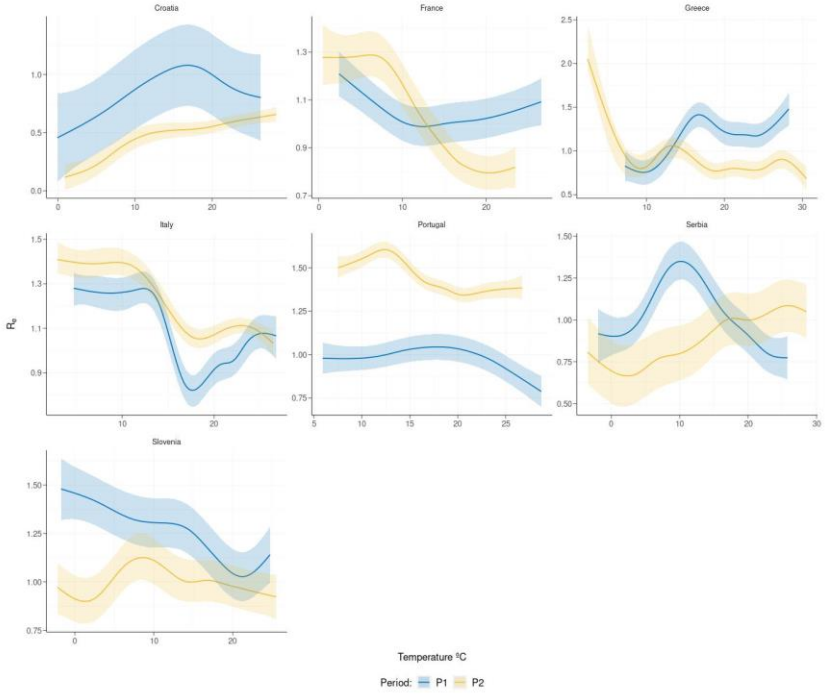
Supplementary Figure S7. Specific association curves for the temperature of the different Spanish regions. Graphic representation of the evolution of Relative Risk (RR) of COVID-19 infection depending on the temperature obtained for each region. The global mean temperature of each region is used as a reference. The blue color represents the period without vaccination (P1) and the yellow color represents the period with vaccination (P2).



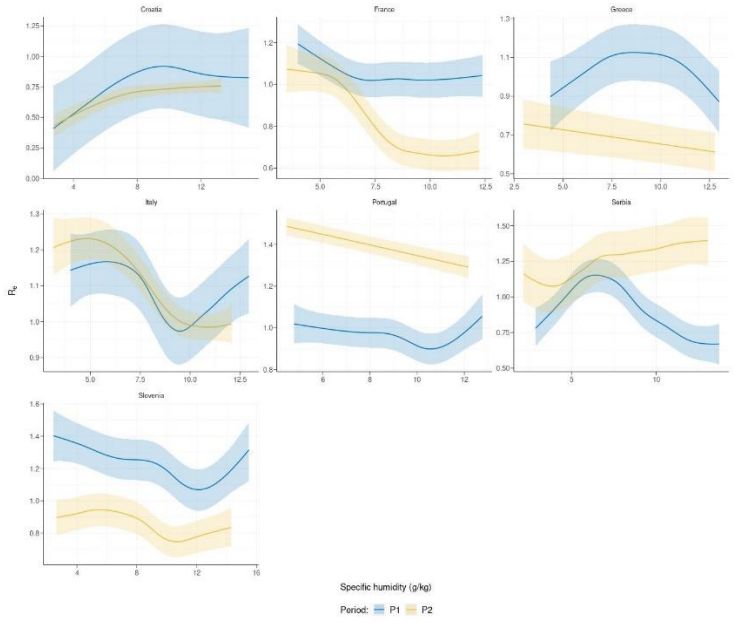
Supplementary Figure S8. Global pooled association curve for the SH. Graphic representation of the evolution of Relative Risk (RR) of COVID-19 infection depending on the SH obtained by the meta-analysis of the different regions models. The global mean SH is used as a reference. The blue color represents the period without vaccination (P1) and the yellow color represents the period with vaccination (P2).



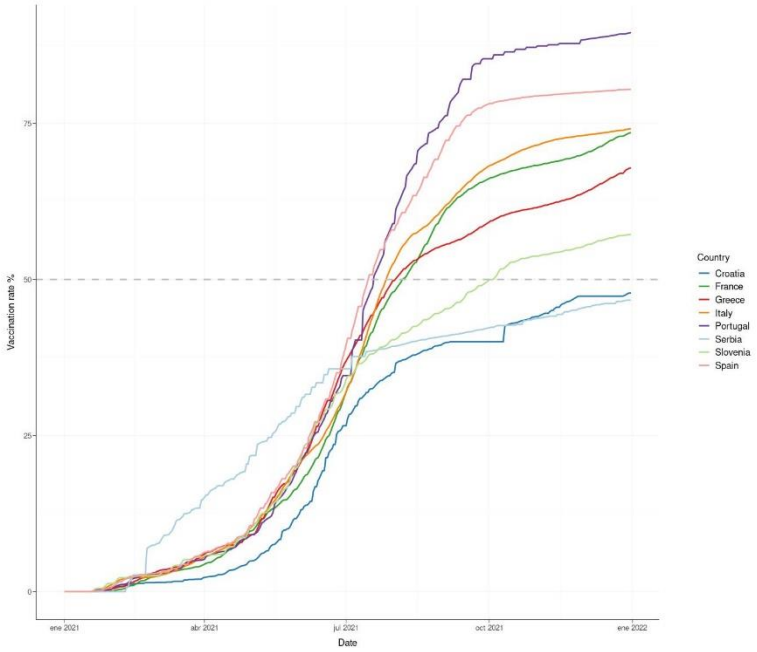
Supplementary Figure S9. Specific association curves for the SH of the different Spanish regions. Graphic representation of the evolution of Relative Risk (RR) of COVID-19 infection depending on the SH obtained for each region. The global mean SH of each region is used as a reference. The blue color represents the period without vaccination (P1) and the yellow color represents the period with vaccination (P2).



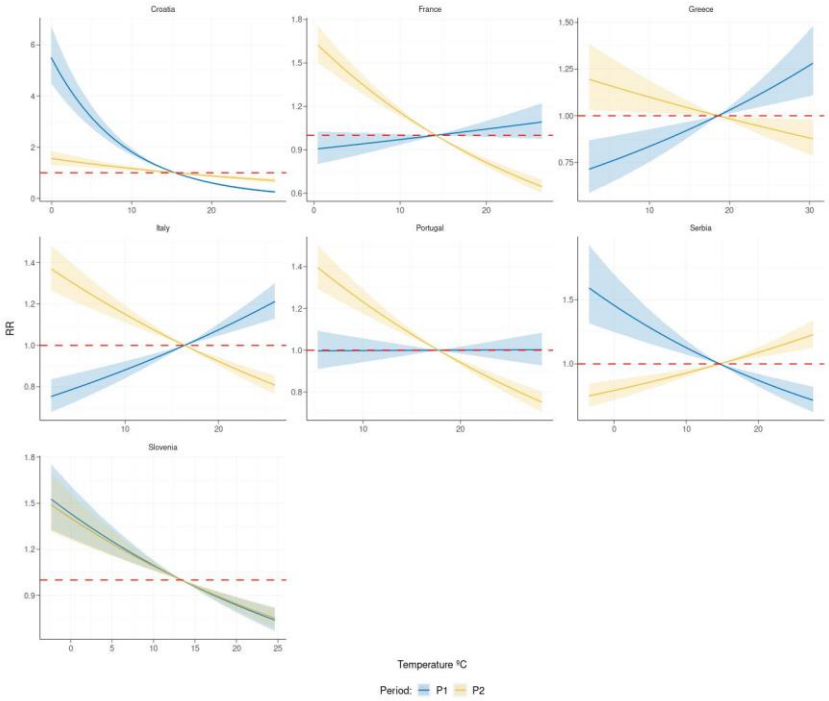
Supplementary Figure S10. Estimation of the evolution R_e from the influence of the temperature (ºC) predicted by the GAM models insolating the rest of variables for European countries. Graphic representation of the R_e predicted by the GAM models with temperature as meteorological variable and considering the rest of variables remain constant for the different European countries. The blue color represents the period without vaccination (P1) and the yellow color represents the period with vaccination (P2).



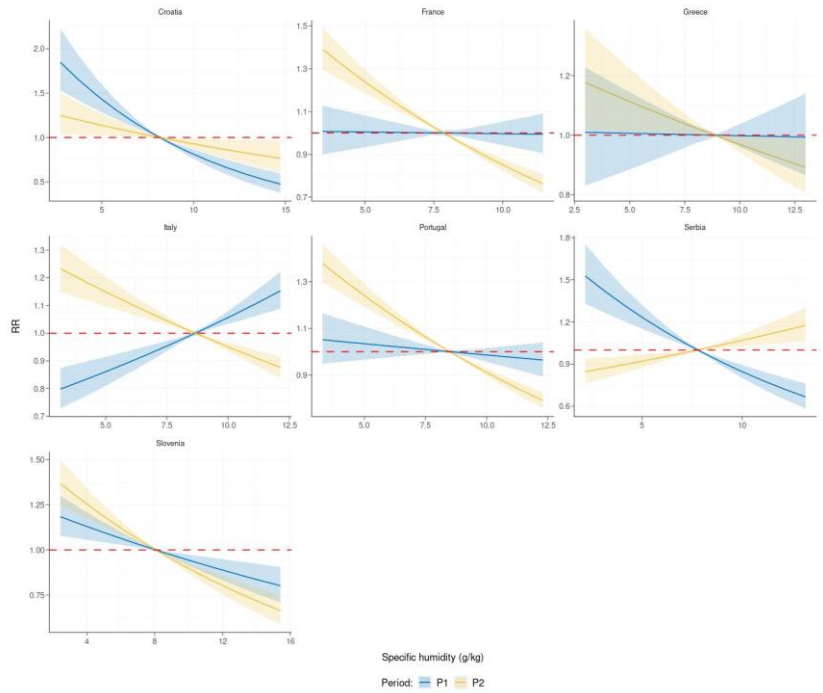
Supplementary Figure S11. **Estimation of the evolution R_e from the influence of SH (g/kg) predicted by the GAM models insolating the rest of variables for European countries**. Graphic representation of the R_e predicted by the GAM models with SH as meteorological variable and considering the rest of variables remain constant for the different European countries. The blue color represents the period without vaccination (P1) and the yellow color represents the period with vaccination (P2).



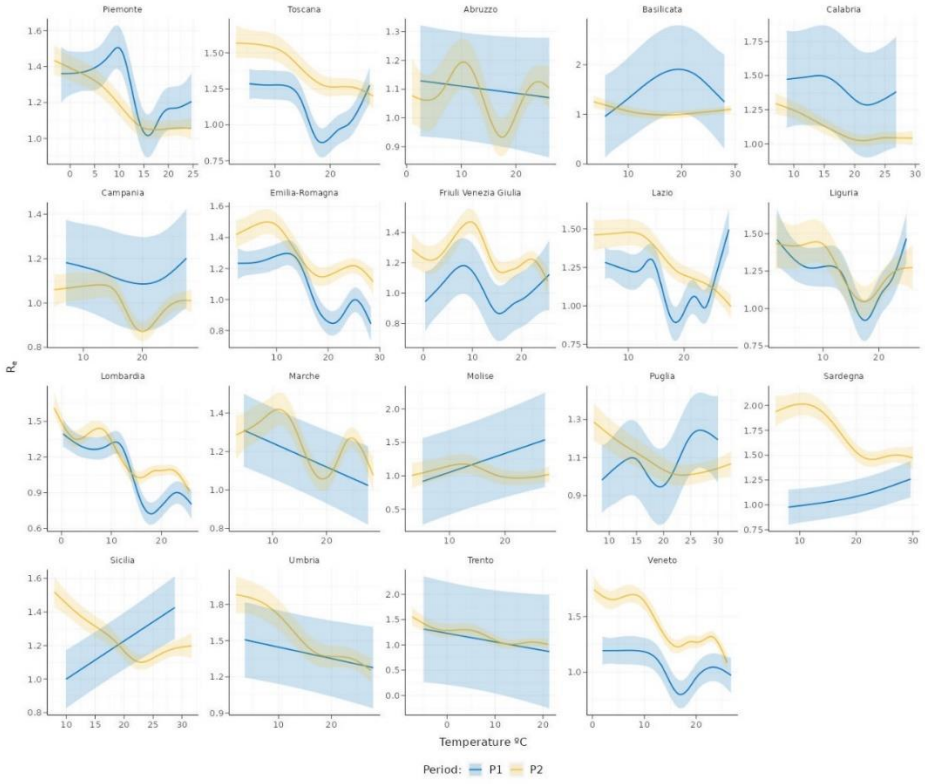
Supplementary Figure S12. **Evolution of the vaccination rate in the different countries of Europe.** Longitudinal plot representing the different vaccination rates (%) for the European countries considered in the study. The grey line represents a percentage of vaccination rate equal to 50%.



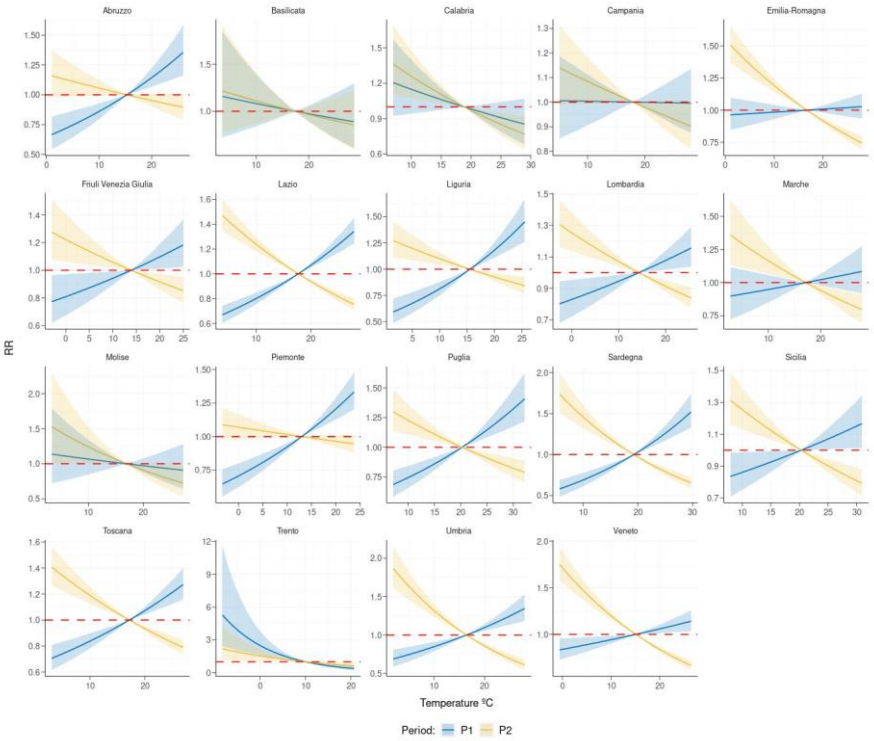
Supplementary Figure S13. Specific association curves for the temperature of the different European countries. Graphic representation of the evolution of Relative Risk (RR) of COVID-19 infection depending on the temperature obtained for each region. The global mean temperature of each region is used as a reference. The blue color represents the period without vaccination (P1) and the yellow color represents the period with vaccination (P2).



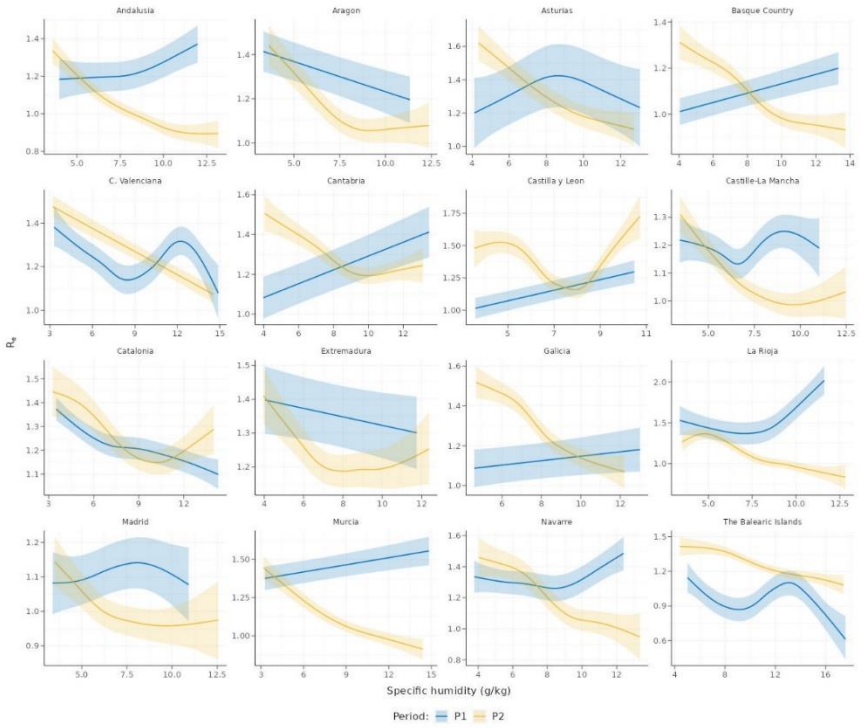
Supplementary Figure S14. Specific association curves for the SH of the different European countries. Graphic representation of the evolution of Relative Risk (RR) of COVID-19 infection depending on the SH obtained for each region. The global mean SH of each region is used as a reference. The blue color represents the period without vaccination (P1) and the yellow color represents the period with vaccination (P2).



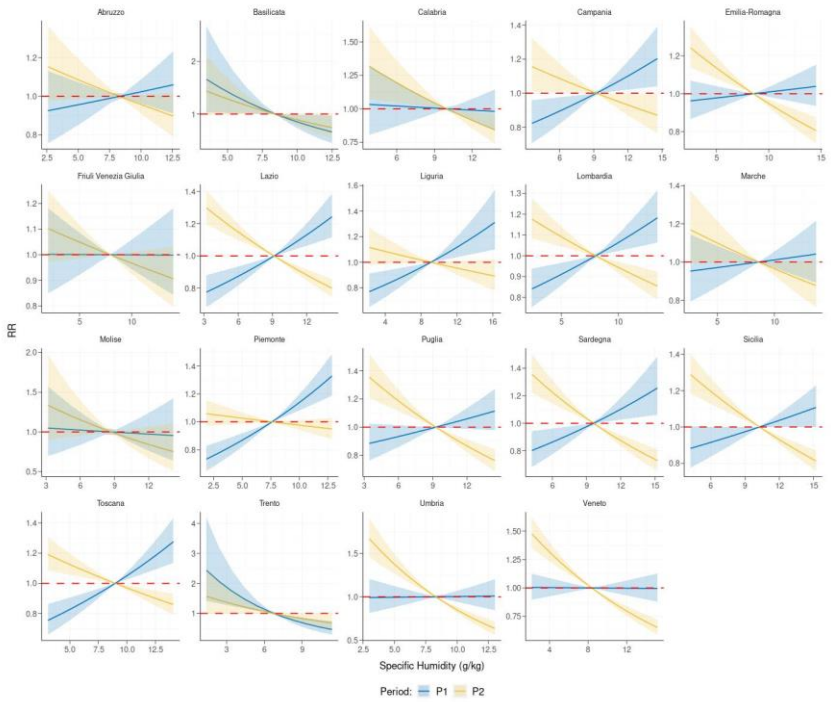
Supplementary Figure S15. **Estimation of the evolution R_e from the influence of the temperature (ºC) predicted by the GAM models insolating the rest of variables for Italian regions**. Graphic representation of the R_e predicted by the GAM models with temperature as meteorological variable and considering the rest of variables remain constant for the different Italian regions. The blue color represents the period without vaccination (P1) and the yellow color represents the period with vaccination (P2).



Supplementary Figure S16. Specific association curves for the temperature of the different Italian regions. Graphic representation of the evolution of Relative Risk (RR) of COVID-19 infection depending on the temperature obtained for each region. The global mean temperature of each region is used as a reference. The blue color represents the period without vaccination (P1) and the yellow color represents the period with vaccination (P2).



Supplementary Figure S17. **Estimation of the evolution R_e from the influence of the SH (g/kg) predicted by the GAM models insolating the rest of variables for Italian regions**. Graphic representation of the R_e predicted by the GAM models with SH as meteorological variable and considering the rest of variables remain constant for the different Italian regions. The blue color represents the period without vaccination (P1) and the yellow color represents the period with vaccination (P2).



Supplementary Figure S18. Specific association curves for the SH of the different Italian regions. Graphic representation of the evolution of Relative Risk (RR) of COVID-19 infection depending on the SH obtained for each region. The global mean SH of each region is used as a reference. The blue color represents the period without vaccination (P1) and the yellow color represents the period with vaccination (P2).

Supplementary Tables

Supplementary Table S1. Results of the GAM models with temperature as meteorological factor for European countries

P1 (1 June 2020 to 31 December 2020)						
Autonomous community	**Deviance explained**	**R^2**	**Corrected P-values**			
			Temperature	**SI**	**Vaccination Rate**	**Variants**
France	50.64	0.49	0.0012	<0.0001	NA	0.9663
Italy	68.33	0.67	<0.0001	<0.0001	NA	0.9663
Greece	30.51	0.28	<0.0001	0.1028	NA	0.9663
Portugal	20.04	0.18	<0.0001	0.0049	NA	0.8610
Slovenia	36.82	0.35	<0.0001	<0.0001	NA	0.9696
Croatia	9.00	0.07	0.0135	0.2590	NA	0.8610
Serbia	55.02	0.53	<0.0001	<0.0001	NA	0.8610
Median	**33.66**	**0.31**	**<0.0001**	**<0.0001**	**NA**	**0.9663**
P2 (1 June 2021 to 31 December 2021)						
Autonomous community	**Deviance explained**	**R^2**	**Corrected p-values**			
			Temperature	**SI**	**Vaccination Rate**	**Variants**
France	63.56	0.62	<0.0001	<0.0001	0.0151	<0.0001
Italy	70.69	0.69	<0.0001	0.6649	<0.0001	<0.0001
Greece	62.45	0.60	<0.0001	<0.0001	0.4115	<0.0001
Portugal	60.22	0.58	<0.0001	0.0013	<0.0001	0.5013
Slovenia	59.73	0.58	0.0008	0.2057	<0.0001	<0.0001
Croatia	56.71	0.55	<0.0001	0.0362	<0.0001	0.0056
Serbia	76.25	0.75	<0.0001	0.0275	0.0483	<0.0001
Median	**62.49**	**0.60**	**<0.0001**	**0.0275**	**<0.0001**	**<0.0001**

Note: Results of the GAM models with temperature as meteorological variable for the different European countries. The first and the second columns represent the deviance explained and the adjusted R^2 of the different models. The rest of columns represent the corrected p-values for the different variables included in the models.

Supplementary Table S2. Results of the GAM models with SH as meteorological factor for European countries

P1 (1 June 2020 to 31 December 2020)						
Conutry	**Deviance explained**	**R^2**	**Corrected P-values**			
			SH	**SI**	**Vaccination Rate**	**Variants**
France	48.96	0.48	0.0220	<0.0001	NA	0.9054
Italy	41.51	0.40	0.0005	<0.0001	NA	0.9248
Greece	15.00	0.13	0.0220	0.0159	NA	0.9054
Portugal	13.99	0.11	0.0194	0.0001	NA	0.4743
Slovenia	32.59	0.30	0.0021	<0.0001	NA	0.9248
Croatia	6.55	0.05	0.0520	0.1470	NA	0.4743
Serbia	45.77	0.44	<0.0001	<0.0001	NA	0.4743
Median	**23.80**	**0.22**	**0.0194**	**<0.0001**	**NA**	**0.9054**
P2 (1 June 2021 to 31 December 2021)						
Country	**Deviance explained**	**R^2**	**Corrected p-values**			
			SH	**SI**	**Vacciantion Rate**	**Variants**
France	59.03	0.57	<0.0001	<0.0001	0.4586	<0.0001
Italy	60.58	0.59	<0.0001	0.7956	<0.0001	<0.0001
Greece	42.23	0.41	0.1241	<0.0001	0.4586	<0.0001
Portugal	53.28	0.52	<0.0001	<0.0001	<0.0001	0.5291
Slovenia	60.27	0.59	0.0001	0.0002	<0.0001	<0.0001
Croatia	46.99	0.45	0.0002	0.7097	0.0344	0.0002
Serbia	70.68	0.69	<0.0001	0.0007	<0.0001	<0.0001
Median	**56.16**	**0.55**	**0.0001**	**0.0002**	**<0.0001**	**<0.0001**

Note: Results of the GAM models with SH as meteorological variable for the different European countries. The first and the second columns represent the deviance explained and the adjusted R^2 of the different models. The rest of columns represent the corrected p-values for the different variables included in the models.

Supplementary Table S3. Results of the GAM models with temperature as meteorological factor for Italian regions

P1 (1 June 2020 to 31 December 2020)						
Region	**Deviance explained**	**R^2**	**Corrected P-values**			
			Temperature	**SI**	**Vaccination Rate**	**Variants**
Abruzzo	15.34	0.14	0.6710	<0.0001	NA	0.9836
Basilicata	5.01	0.03	0.3631	0.6089	NA	0.9836
Calabria	8.24	0.06	0.4525	0.0029	NA	0.9836
Campania	21.81	0.20	0.6710	<0.0001	NA	0.9836
Emilia-Romagna	58.80	0.57	<0.0001	<0.0001	NA	0.9836
Friuli Venezia Giulia	17.86	0.15	0.0179	0.0003	NA	0.9836
Lazio	54.66	0.52	<0.0001	<0.0001	NA	0.9836
Liguria	40.59	0.38	<0.0001	<0.0001	NA	0.9836
Lombardia	69.26	0.68	<0.0001	<0.0001	NA	0.9836
Marche	12.36	0.11	0.0535	<0.0001	NA	0.9904
Molise	4.54	0.03	0.2448	0.3182	NA	0.9836
Piemonte	52.00	0.50	<0.0001	<0.0001	NA	0.9836
Puglia	18.38	0.16	0.0039	0.0584	NA	0.9836
Sardegna	20.98	0.20	0.0699	0.0015	NA	0.9836
Sicilia	15.85	0.15	0.0029	0.2360	NA	0.9836
Toscana	60.26	0.59	<0.0001	<0.0001	NA	0.9836
Umbria	14.35	0.13	0.3631	<0.0001	NA	0.9836
Veneto	29.82	0.27	<0.0001	<0.0001	NA	0.9836
Trento	0.89	-0.01	0.6404	0.2360	NA	0.9836
Median	**18.38**	**0.16**	**0.0179**	**<0.0001**	**NA**	**0.9836**

P2 (1 June 2021 to 31 December 2021)						
Region	**Deviance explained**	**R^2**	**Corrected p-values**			
			Temperature	**SI**	**Vacciantion Rate**	**Variants**
Abruzzo	42.67	0.40	<0.0001	0.0015	0.0305	0.0002
Basilicata	40.48	0.38	0.009	<0.0001	0.0019	<0.0001
Calabria	53.54	0.52	<0.0001	0.0312	<0.0001	<0.0001
Campania	68.11	0.67	<0.0001	0.6829	0.0002	<0.0001
Emilia-Romagna	58.89	0.57	<0.0001	0.01125	<0.0001	<0.0001
Friuli Venezia Giulia	47.23	0.44	<0.0001	0.0009	<0.0001	<0.0001
Lazio	53.10	0.51	<0.0001	0.0246	<0.0001	<0.0001
Liguria	24.61	0.21	<0.0001	0.6829	0.0047	0.0005
Lombardia	77.36	0.76	<0.0001	0.1736	<0.0001	<0.0001
Marche	58.05	0.56	<0.0001	<0.0001	<0.0001	<0.0001
Molise	32.40	0.30	0.0734	<0.0001	0.9052	0.9600
Piemonte	72.52	0.71	<0.0001	0.0066	<0.0001	<0.0001
Puglia	55.63	0.54	0.0004	0.0013	<0.0001	<0.0001
Sardegna	53.38	0.51	<0.0001	0.0013	<0.0001	<0.0001
Sicilia	56.19	0.54	<0.0001	0.5416	<0.0001	<0.0001

Toscana	64.25	0.63	<0.0001	0.007	<0.0001	<0.0001
Umbria	54.27	0.53	<0.0001	0.6829	<0.0001	<0.0001
Veneto	69.10	0.67	<0.0001	0.0013	<0.0001	<0.0001
Trento	41.24	0.39	<0.0001	0.0033	0.0001	<0.0001
Median	**54.27**	**0.53**	**<0.0001**	**0.0066**	**<0.0001**	**<0.0001**

Note: Results of the GAM models with temperature as meteorological variable for the different Italian regions. The first and the second columns represent the deviance explained and the adjusted R^2 of the different models. The rest of columns represent the corrected p-values for the different variables included in the models.

Supplementary Table S4. Results of the GAM models with SH as meteorological factor for Italian regions

P1 (1 June 2020 to 31 December 2020)						
Autonomous community	**Deviance explained**	**R^2**	**Corrected P-values**			
			SH	**SI**	**Vaccination Rate**	**Variants**
Abruzzo	19.91	0.18	0.1461	<0.0001	NA	0.9772
Basilicata	4.68	0.03	0.3844	0.0266	NA	0.9772
Calabria	7.92	0.07	0.0819	0.0001	NA	0.9772
Campania	27.52	0.25	0.0224	<0.0001	NA	0.9772
Emilia-Romagna	27.71	0.26	0.0079	<0.0001	NA	0.9772
Friuli Venezia Giulia	9.24	0.08	0.7774	0.0011	NA	0.9772
Lazio	28.92	0.28	0.1448	<0.0001	NA	0.9772
Liguria	40.23	0.38	<0.0001	<0.0001	NA	0.9772
Lombardia	49.17	0.47	<0.0001	<0.0001	NA	0.9772
Marche	11.53	0.10	0.1607	<0.0001	NA	0.9772
Molise	3.92	0.03	0.5424	0.0696	NA	0.9772
Piemonte	43.72	0.41	<0.0001	<0.0001	NA	0.9772
Puglia	13.35	0.12	0.0446	0.0312	NA	0.9772
Sardegna	20.53	0.19	0.4926	<0.0001	NA	0.9772
Sicilia	18.29	0.16	0.0224	0.0467	NA	0.9772
Toscana	42.57	0.41	0.5027	<0.0001	NA	0.9772
Umbria	21.90	0.20	0.0020	<0.0001	NA	0.9772
Veneto	10.96	0.10	0.9013	0.0003	NA	0.9772
Trento	0.74	-0.01	0.9013	0.3890	NA	0.9772
Median	**19.91**	**0.18**	**0.1448**	**<0.0001**	**NA**	**0.9772**

P2 (1 June 2021 to 31 December 2021)						
Autonomous community	**Deviance explained**	**R^2**	**Corrected p-values**			
			SH	**SI**	**Vacciantion Rate**	**Variants**
Abruzzo	32.70	0.30	0.0221	0.0139	0.0775	<0.0001
Basilicata	37.92	0.36	0.0737	<0.0001	<0.0001	<0.0001
Calabria	52.65	0.51	<0.0001	0.0194	<0.0001	<0.0001
Campania	57.13	0.56	0.0786	0.6907	0.0139	<0.0001
Emilia-Romagna	46.12	0.44	0.0011	0.2038	<0.0001	<0.0001
Friuli Venezia Giulia	31.35	0.29	0.0072	0.0069	<0.0001	<0.0001
Lazio	42.10	0.41	<0.0001	0.0141	<0.0001	<0.0001
Liguria	19.78	0.17	0.0016	0.7851	0.0130	0.0003
Lombardia	66.91	0.66	<0.0001	0.2506	<0.0001	<0.0001
Marche	41.28	0.39	0.0087	0.0021	<0.0001	<0.0001
Molise	29.12	0.27	0.4845	<0.0001	0.2454	0.3936
Piemonte	70.43	0.69	<0.0001	0.0139	<0.0001	<0.0001
Puglia	52.65	0.52	0.0019	0.0023	<0.0001	<0.0001
Sardegna	42.32	0.40	<0.0001	0.0102	<0.0001	<0.0001
Sicilia	47.89	0.47	<0.0001	0.2416	<0.0001	<0.0001

Toscana	58.26	0.57	0.0017	0.0102	<0.0001	<0.0001
Umbria	50.90	0.49	<0.0001	0.6907	<0.0001	<0.0001
Veneto	54.78	0.53	<0.0001	0.0115	<0.0001	<0.0001
Trento	36.52	0.35	0.0008	0.0102	0.0004	<0.0001
Median	**46.12**	**0.44**	**0.0011**	**0.0139**	**<0.0001**	**<0.0001**

Note: Results of the GAM models with SH as meteorological variable for the different Italian regions. The first and the second columns represent the deviance explained and the adjusted R^2 of the different models. The rest of columns represent the corrected p-values for the different variables included in the models.

10. PRODUCCIÓN CIENTÍFICA

En esta sección, proporcionamos un resumen de la producción científica del doctorando durante el desarrollo de la tesis, lo cual abarca tanto trabajos y publicaciones alineados con los objetivos específicos de la investigación como aquellos que trascienden dichos objetivos. Para los artículos publicados en revistas indexadas en el Journal Citation Reports (JCR), se incluye información sobre el factor de impacto (I.F) y su posición relativa en términos de cuartiles (Q) o deciles (D). Además, se indicará con un "*" cuando los autores compartan la primera autoría,

10.1. ARTÍCULOS CON RESULTADOS DE LA TESIS

1. Toro-Domínguez D*, **Villatoro-García JA***, Martorell-Marugán J, Román-Montoya Y, Alarcón-Riquelme ME, Carmona-Sáez P. A survey of gene expression meta-analysis: methods and applications. Briefing Bioinformatics. Published online February 25, 2020. doi:10.1093/bib/bbaa019. **IF:13.994 (D1)**
2. **Villatoro-García JA**, Martorell-Marugán J, Toro-Domínguez D, Román-Montoya Y, Femia P, Carmona-Sáez P. DExMA: An R Package for Performing Gene Expression Meta-Analysis with Missing Genes. *Mathematics*. 2022;10(18):3376. doi:10.3390/math10183376. **IF:2.4 (D1)**
3. **Villatoro-García JA**, López-Domínguez R, Martorell-Marugán J, Luna J de D, Lorente JA, Carmona-Sáez P. Exploring the interplay between climate, population immunity and SARS-CoV-2 transmission dynamics in Mediterranean countries. *Sci Total Environ*. 2023;897:165487. doi:10.1016/j.scitotenv.2023.165487. **IF:8.2 (D1)**
4. Martorell-Marugán J*, **Villatoro-García JA***, García-Moreno A, et al. DatAC: A visual analytics platform to explore climate and air quality indicators associated with the COVID-19 pandemic in Spain. *Sci Total Environ*. 2021;750:141424. doi:10.1016/j.scitotenv.2020.141424. **IF: 10:754 (D1)**
5. **Villatoro-García, JA***, Jurado-Bascón P*, Carmona-Sáez P. A new methodology for conducting pathway enrichment meta-analysis. *En revisión.*

10.2. COLABORACIONES COMO CO-AUTOR

1. Garcia-Moreno A, López-Domínguez R, **Villatoro-García JA**, Ramirez-Mena A, Aparicio-Puerta E, Hackenberg M, Pascual-Montano A, Carmona-Saez P. Functional Enrichment Analysis of Regulatory Elements. Biomedicines. 2022; 10(3):590. https://doi.org/10.3390/biomedicines10030590. IF: 4.7 (Q1).

2. Martorell-Marugán J., López-Domínguez R., García-Moreno A, Toro-Domínguez D **Villatoro-García JA**, Barturen G, Martín-Gómez A, Troule K, Gómez-López Gonzalo, Al-Shahrour F, González-Rumayor V, Peña-Chilet M, Dopazo J, Sáez-Rodruíguez J, Alarcón-Riquelme ME, Carmona-Sáez P. A. comprehensive database for integrated analysis of omics data in autoimmune diseases. BMC Bioinformatics 22, 343 (2021). https://doi.org/10.1186/s12859-021-04268-4. IF: 3.3 (Q2)

3. Martorell-Marugán J., López-Domínguez R, **Villatoro-García JA**, Toro-Domínguez D., Chierici M., Jurman G and Carmona-Sáez P. Explainable deep learning for predicting samples phenotypes from single-cell transcriptomics. En revisión en la revista *Briefings in Bioinformatics*.
4. Caracuel-Peramos R, Rodríguez-Baena FJ, Redondo-García S, **Villatoro-García JA**, Garcia-Munoz A, Peris-Torres C, Plaza-Calonge MC, Lopez-Millan B, Ricciardelli C, Russell D, Carmona-Sáez P, and Rodriguez-Manzaneque JC. Loss of the extracellular protease ADAMTS1 reveals an antitumorigenic program involving the action of NIDOGEN-1 on macrophage polarization. En revisión en la revista *Journal of Biomedical Science.*