Escola Tècnica Superior d'Enginyeria ETSE-UV

MÁSTER UNIVERSITARIO EN BIOINFORMÁTICA

TRABAJO DE FIN DE MÁSTER

EVALUACIÓN DE GENES HOUSEKEEPING EN TEJIDO ADIPOSO, POR SEXO Y ESPECIE, MEDIANTE EL ANÁLISIS MASIVO DE DATOS TRANSCRIPTÓMICOS

AUTORA:
MARIA GUAITA CÉSPEDES

TUTORES:
FRANCISCO GARCÍA GARCÍA
MARTA R. HIDALGO GARCÍA
Mª AMPARO GALÁN ALBIÑANA

SEPTIEMBRE, 2020

MÁSTER UNIVERSITARIO EN BIOINFORMÁTICA

TRABAJO DE FIN DE MÁSTER

EVALUACIÓN DE GENES HOUSEKEEPING EN TEJIDO ADIPOSO, POR SEXO Y ESPECIE, MEDIANTE EL ANÁLISIS MASIVO DE DATOS TRANSCRIPTÓMICOS

AUTORA:
MARIA GUAITA CÉSPEDES

TUTORES:
FRANCISCO GARCÍA GARCÍA
MARTA R. HIDALGO GARCÍA
Mª AMPARO GALÁN ALBIÑANA

SEPTIEMBRE, 2020

TRIBUNAL:

PRESIDENTE/A:

VOCAL 1:

VOCAL 2:

FECHA DE DEFENSA:

CALIFICACIÓN:

Resumen

Los genes *housekeeping* participan en funciones esenciales para el mantenimiento celular y son conocidos por mantener niveles de expresión estables en todos los tipos celulares, motivo por el cual son utilizados como controles internos en estudios de la expresión génica. No obstante, multiples estudios han demostrado que aquellos genes descritos en la literatura como genes *housekeeping* no presentan unos niveles de expresión estables a lo largo de los diferentes tipos de células, tejidos y condiciones, introduciendo un error sistemático en los resultados experimentales. Su adecuada validación y selección en las condiciones estudiadas son clave para la validez de los resultados obtenidos.

En este trabajo realizamos la evaluación de los perfiles de expresión de seis genes *housekeeping* clásicos, cuatro metabólicos (HPRT, GAPDH, PPIA y UBC) y dos ribosomales (18S y RPL19), y de diferentes colecciones de genes abarcados en los datos de expresión génica obtenidos con microarrays disponible en la base de datos *Gene Expression Omnibus* (GEO), con el objetivo de determinar su estabilidad de la expresión en el tejido adiposo de *Homo sapiens* y *Mus musculus*.

Palabras clave: genes *housekeeping*, perfiles de expresión, microarrays, transcriptómica, tejido adiposo

Abstract

Housekeeping genes are involved in the maintenance of basic cell functions and have been known to have constant expression levels in all cell types, hence why are used as internal controls in gene expression studies. Nevertheless, multiple studies have shown that those described in the literature as housekeeping genes do not have stable expression levels across different cells, tissues and conditions, introducing a systematic error in the experimental results. Their proper selection and validation in the specific studied conditions is key for the validity of the obtained results.

In this work, we evaluate the expression profiles of six classical housekeeping genes, four metabolic (HPRT, GAPDH, PPIA and UBC) and two, ribosomal (18S and RPL19), and different collections of genes covered in the expression profiling by array data available at Gene Expression Omnibus database (GEO), in order to determine the stability of their expression in adipose tissue of Homo sapiens and Mus musculus.

Key words: housekeeping genes, gene expression profiling, microarrays, transcriptomics, adipose tissue

Agradecimientos

Gracias a Paco García por darme la oportunidad de participar en este proyecto, guiarme en todo el proceso y mantener la energía positiva en todo momento.

A mis compañeros de la Unidad de Bioinformática y Bioestadística del CIPF por acogerme como una más del equipo y ayudarme en todo lo que he necesitado.

A mis compañeros del máster, a Isma, a Antonio y a Luis, por acompañarme día a día en esta aventura.

A mis amigos de toda la vida, a Juan, a Pablo y a Cris, y a mis biólogas, por animarme en todo momento y sacarme de la rutina.

A Fran, por confiar en mí y estar siempre a mi lado.

A mi familia, por el apoyo incondicional a todas las decisiones que he tomado en la vida.

Sin vosotros todo este trabajo no habría sido posible.

Índice general

Índice de figuras

Índice de tablas

Glosario de Acrónimos

ADN ácido desoxirribonucleico

ARN ácido ribonucleico

PCR Reacción en cadena de la polimerasa

RT-qPCR Reacción en cadena de la polimerasa cuantitativa con transcripción reversa

HK *Housekeeping*

HPRT Hipoxantina-guanina fosforibosiltransferasa

GAPDH Gliceraldehído-3-fosfato deshidrogenasa

PPIA Peptidilprolil isomerasa A

UBC Ubiquitina C

RPL19 Proteina ribosómica 60S

18S ARN ribosómico 18S

GEO *Gene Expression Omnibus*

NCBI *National Center for Biotechnology Information*

NGS *Next-generation sequencing*

CV Coeficiente de Variación

IQR Rango intercuartílico

MAD Desviación mediana absoluta

RP *Rank Product*

HTML Lenguaje de marcas de hipertexto

SQL Lenguaje de consulta estructurada

1. Introducción

1.1. Conceptos básicos de Genética Molecular y tecnologías de alto rendimiento

El ácido desoxirribonucleico (ADN) es la molécula que contiene la información genética que codifica para todas las características, procesos y funcionalidades de una célula, está formada por dos cadenas antiparalelas polinucleotídicas. Estas cadenas de ADN se componen de cuatro nucleótidos: Adenina (A), Citosina (C), Guanina (G), Timina (T) [1].

En el núcleo celular se lleva a cabo el proceso de transcripción génica mediante el cual se transfiere la información contenida en la secuencia del ADN a una molécula de ácido ribonucleico (ARN), sustituyendo la Timina por Uracilo (U), para la síntesis de un producto biológico funcional que puede ser el propio ARN o su traducción en el citoplasma a los aminoácidos que conforman una proteína. Denominamos gen a la unidad de información del ADN que codifica un producto génico. El conjunto completo de genes de un organismo compone su genoma [1].

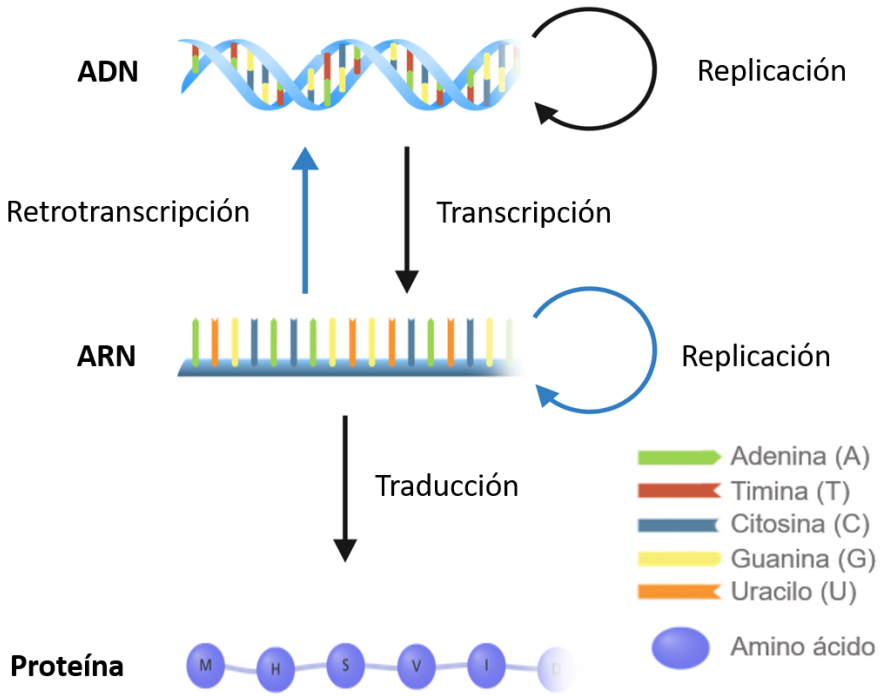

Figura 1.1: Representación del dogma central de la Biología Molecular. En negro el diseño inicial de Francis Crick y en azul las modificaciones posteriores.

El descubrimiento de la estructura en forma de doble hélice del ADN [2] junto a la formulación del dogma central de la biología, que recoge los tres procesos principales de utilización de la información genética (replicación, transcripción y traducción) y la relación entre sus componentes (figura 1.1) [3], sentaron las bases para el entendimiento de la complejidad del entorno genético y el estudio de las dinámicas de la expresión génica.

El objeto de interés de los estudios de la expresión génica es el transcriptoma, el conjunto de todas las moléculas de ARN, también llamadas transcritos, presentes en una célula en un momento determinado [4]. En base al papel al papel que cumplen dentro de la célula se distinguen diferentes tipos de moléculas de ARN: el ARN mensajero (ARNm) codificante de proteínas, y el ARN no codificante, que engloba el ARN ribosomal (ARNr), transferente (ARNt), de interferencia (ARNi), micro ARN (mi-ARN) y ARN largo no codificante (en inglés, *lncRNA*), entre otros [5].

En las últimas décadas el importante progreso de la tecnología ha dado lugar al origen de un conjunto de métodos que permiten estudiar la complejidad biológica a gran escala, las técnicas ómicas [6]. La era de las ómicas comenzó con la Genómica, término propuesto en 1986, con el que se hace referencia al estudio sistemático del genoma completo de un organismo [7]. Más adelante surgió la Transcriptómica, que tiene como objeto de estudio el conjunto de moléculas de ARN de una célula [4]. Entre sus aplicaciones directas destacan catalogar todos los tipos de transcritos y determinar la estructura transcripcional de los genes [8]. Existen otras técnicas ómicas como la Proteómica o la Metabolómica, o las que han emergido más recientemente, la Epigenómica, Interactómica y la Metagenómica.

En su conjunto, todas las técnicas ómicas caracterizan simultáneamente miles de variables en una muestra, generando un volumen importante de información compleja, de gran dimensionalidad, que requiere del uso de herramientas computacionales para su interpretación biológica.

1.1.1. Datos transcriptómicos y técnicas de análisis del transcriptoma

En los estudios de la expresión génica se llevan a cabo diferentes metodologías para analizar el transcriptoma, cuantificar la abundancia de los transcritos y determinar los perfiles de expresión génica bajo unas condiciones concretas.

RT-qPCR

Una de las técnicas clásicas que se realizan para estudiar los niveles de transcripción es la Reacción en Cadena de la Polimerasa cuantitativa con Trancripción Reversa (RT-qPCR). Es una técnica específica y muy sensible que permite la detección y cuantificación de un número limitado de transcritos de interés, donde el ARN se retro-transcribe a ADN complementario (ADNc), se amplifica mediante una PCR con cebadores específicos para las regiones genómicas de interés y se cuantifica la abundancia en función de la acumulación del producto amplificado [4].

Ha sido de las metodologías más utilizadas en los laboratorios hasta la llegada de las tecnologías de alto rendimiento que permiten la cuantificación simultánea de los niveles

de expresión de miles de genes y a estudiar el perfil transcriptómico del genoma completo, entre las que destacan los microarrays y la secuenciación masiva de ARN (RNA-Seq) [9].

Microarrays

Los microarrays están formados por una serie de sondas de ADN complementarias a regiones específicas del genoma fijadas de manera ordenada en un soporte sólido que suele ser de vídrio. Es una técnica basada en la hibridación propia de ácidos nucleicos, donde las sondas, que pueden ser clones de ADN, productos de PCR de longitud variable (ADN bicatenario) u oligonucleótidos sintéticos específicos, se unen al soporte e hibridan con los transcritos diana de secuencia complementaria (figura 1.2) [10].

Figura 1.2: Tecnología de microarray. Se muestran las sondas prefijadas al soporte sólido que hibridan con las muestras de ARN aislado con marcaje fluorescente de secuencia complementaria.

La abundancia de transcritos se infiere de forma indirecta a partir de la intensidad de la hibridación [4]. La implementación de esta metodología requiere del conocimiento previo de la secuencia de los genes que se quieren interrogar para el diseño las sondas, los transcritos que no sean complementarios a las secuencias de las sondas del no serán detectados por lo que no es capaz de identificar nuevos transcritos.

El protocolo técnico (figura 1.3) consiste en la extracción y purificación del ARN de las células, el material genético extraído se retro-transcribe a ADNc y durante el proceso se realiza el marcaje de la muestra con fluorescencia [11]. Se suelen emplear uno o dos fluorocromos diferentes, el uso de dos marcadores (rojo y verde) permite la hibridación de dos muestras simultáneamente en un mismo array y la detección de expresión génica diferencial [12]. La muestra marcada se hibrida con las sondas correspondientes y se realizan lavados para eliminar las hibridaciones inespecíficas [13]. Terminado el proceso de hibridación, se emplea un escáner láser para excitar los fluoróforos y adquirir una imagen digital donde cada píxel se corresponde con la señal emitida por el marcaje. En este paso pueden darse diferentes situaciones que afecten a la calidad de la imagen obtenida, como la saturación del color de los píxeles o puntos de sondas no alineados. La calidad de la imagen es indispensable para poder obtener un valor cuantitativo de la expresión, en

caso de obtener una imagen de mala calidad conviene repetir el proceso de escaneado o incluso la hibridación. Una vez obtenidas las imágenes se procede a identificar los puntos, se realiza la corrección del ruido de fondo local para cada punto, se normalizan los datos para eliminar la componente no biológica de la variación y se cuantifican las intensidades de fluorescencia [14].

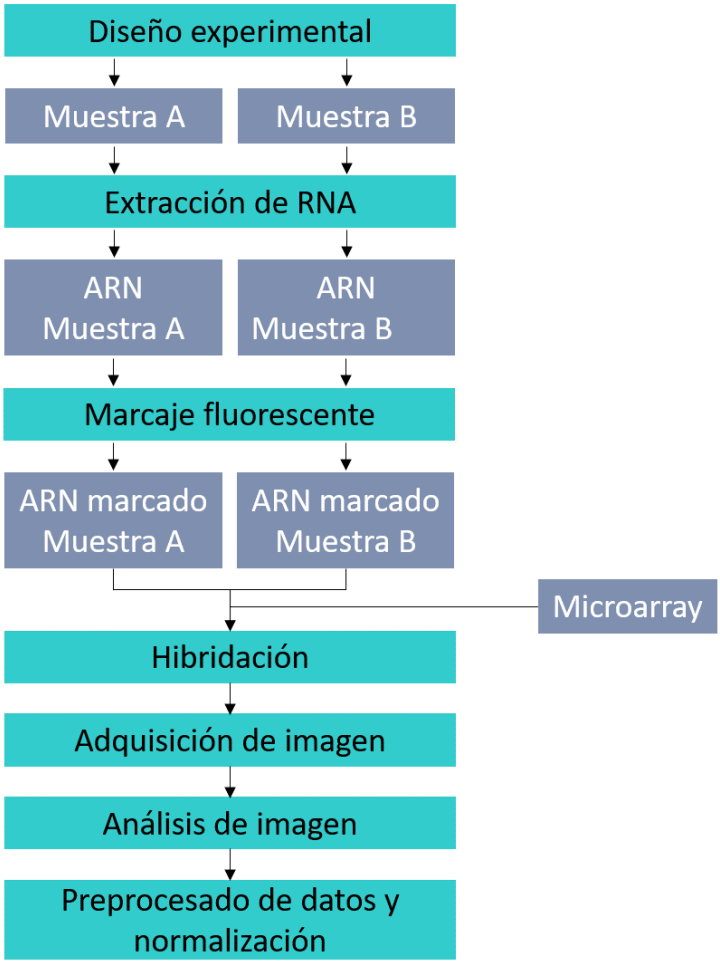

Figura 1.3: Esquema del proceso experimental con microarrays. Una vez procesados los datos se puede proceder a su interpretación biológica.

Al final del proceso obtenemos una matriz con la colección de intensidades de fluorescencia correspondientes a cada sonda del array, la intensidad de la señal detectada es la medida de la expresión génica del transcrito de interés [12, 11, 15]. Esta matriz de expresión génica es la información que se analiza para encontrar el significado biológico.

Así pues, en un experimento con microarrays se generan tres niveles de información relevante, las imágenes escaneadas (los datos crudos), las matrices cuantitativas de las intensidades de fluorescencia medidas de cada microarray y la matriz de expresión génica derivada de las anteriores que recoge los resultados obtenidos con cada una de las muestras analizadas (figura 1.4).

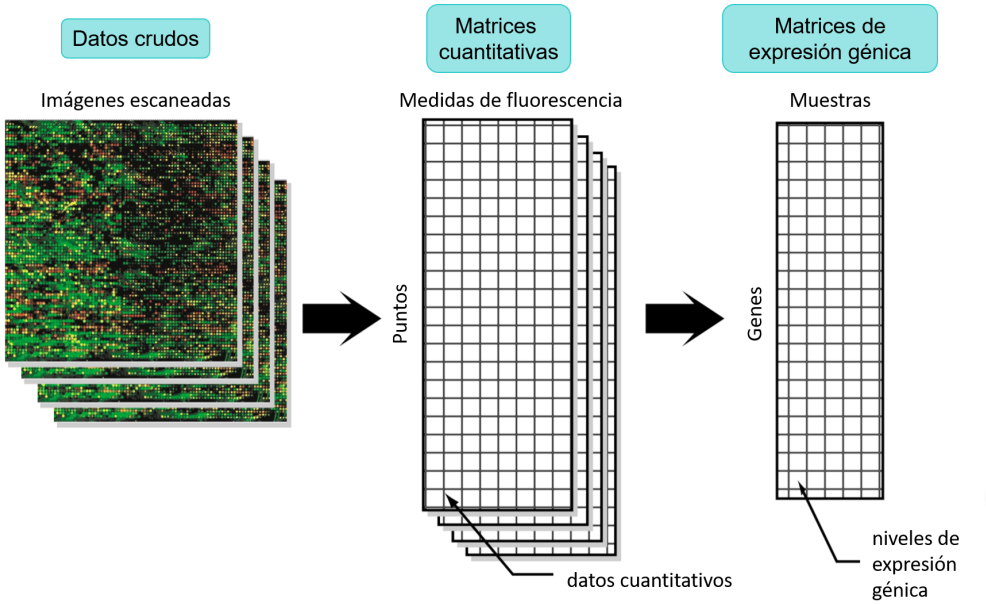

Figura 1.4: Niveles de informácion generada con microarrays. Figura adaptada de Brazma, Alvis, et al. (2001) [45].

RNA-Seq

La técnica de secuenciación masiva del ARN (RNA-Seq) ofrece una visión global del transcriptoma completo. Determina la secuencia nucleotídica de los transcritos mediante el uso de adaptadores de ADN complementarios generados por transcripción reversa del ARN extraído. Al final del proceso se obtienen unas lecturas de 30-400pb que se se mapean contra el genoma de referencia, el número de lecturas mapeadas es la medida del nivel de expresión génica de esa región genómica [15].

El protocolo técnico general está explicado por Wilhelm et al. [16]. Conesa et al. [17] recogen una serie consideraciones para diseño experimental. A diferencia de los microarrays su aplicación no está limitada a tener un conocimiento previo de la región genómica de interés, permitiendo la detección de transcritos no conocidos [8], no obstante, el coste de su aplicación es mayor.

1.1.2. Importancia de estas técnicas en estudios biomédicos

En su conjunto toda la información proporcionada por el estudio del transcriptoma nos permite entender las dinámicas del metabolismo celular y tisular. En el ámbito de la Biomedicina, generan retratos a nivel molecular de la expresión génica de cualquier tipo celular, evidenciando el efecto de las alteraciones asociadas a enfermedades en pacientes [18].

Las técnicas de alto rendimiento, que monitorizan simultáneamente la expresión génica de un gran número de genes, encuentran su aplicación directa en la detección a gran escala de alteraciones en los niveles de expresión relacionados a una condición particular, como el cáncer, la respuesta a un tratamiento o un fármaco; o la identificación de patrones de expresión diferenciales en muestras independientes.

La importancia de la determinación de patrones de expresión génica concretos bajo condiciones de enfermedad reside en el traslado de este conocimiento a la clínica, encontrando genes clave que pueden actuar como biomarcadores de la afectación a nivel molecular, adquiriendo un papel diagnóstico, ayudando a la detección temprana de la afección, o pronóstico, que permita dirigir el tratamiento [19].

1.2. Genes housekeeping

1.2.1. ¿Qué son los genes housekeeping?

Los genes *housekeeping* (HK), también denominados de mantenimiento, llevan a cabo las funciones esenciales para el mantenimiento de la célula. Son genes que realizan un papel indispensable para la supervivencia celular y se expresan a nivel basal en todas las células y tejidos independientemente de la etapa del desarrollo del organismo o las condiciones ambientales [20].

Participan de procesos asociadoss al ciclo celular, la traducción, metabolismo del ARN y proteinas, biogénesis ribosomal, transporte celular y fosforilación oxidativa. [21]. Al realizar actividades básicas de la viabilidad celular, tradicionalmente, se ha considerado que debían expresarse de manera constante en todos los tipos celulares y tejidos independientemente de las condiciones fisiopatológicas [22]. Sus características estructurales y su evolución, diferentes a los genes con expresión específica en un tejido (genes *tissue-specific*), han sido objeto de estudio [23].

De acuerdo con la última definición [24], un gen HK ideal, además de llevar a cabo funciones básicas para el mantenimiento de la vida celular, debe presentar unos patrones de expresión constantes con variaciones mínimas en todas las células y condiciones. Algunos genes descritos clásicamente en la literatura como housekeeping son: Gliceraldehído-3-fosfato deshidrogenasa (GAPDH), beta-actina (ACTB), Hipoxantina-guanina fosforibosiltransferasa (HPRT), Ubiquitina C (UBC), glucosa-6-fosfato deshidrogenasa (G6PD), lactato deshidrogenasa A (LDHA), o los ARN de las subunidades ribosomales 18s y 28s.

1.2.2. Papel de los genes housekeeping en los estudios biomédicos

Las técnicas cuantitativas de la expresión génica, como la RT-qPCR y los microarrays, requieren de la normalización de los valores que se generan para corregir variaciones intrínsecos al desarrollo de la metodología técnica, como errores experimentales debidos a la manipulación de la muestra o errores estocásticos (ruido). El uso de controles internos es la estrategia más común para normalizar las medidas de expresión. En el diseño experimental tenemos genes de los que queremos conocer sus perfiles de expresión bajo unas condiciones concretas y genes que empleamos como referencia contra los que comparamos la expresión de los genes de interés.

Los genes que se emplean como controles internos han de cumplir con dos propiedades: 1) han de ser esenciales para el mantenimiento de la función y viabilidad celular, por tanto, se han de expresar en todos los tejidos y su expresión no debe variar en las entidades biológicas que se están analizando; 2) su transcripción no se puede ver afectada por las condiciones experimentales que se están estudiando [25].

Bajo estas premisas, los genes HK son adecuados para su uso como genes de referencia para normalizar los datos cuantitativos de niveles de ARN [22, 25], ya que sus niveles de expresión relativamente estables permiten corregir estos errores experimentales aumentando la robustez del análisis y la confianza en los resultados obtenidos.

De este modo, adquieren un carácter instrumental de valor añadido como herramienta de calibración en estudios biomédicos al responder a la necesidad de controles que permitan comparar cuantitativamente los perfiles de expresión génica obtenidos de diferentes condiciones experimentales.

Usos y limitaciones de los genes housekeeping como control interno

Durante muchos años los genes HK han sido considerados como buena referencia y su uso ha sido muy extendido como genes control.

Ofrecen una serie de ventajas frente a otras alternativas como normalizar mediante la cantidad total de ARN, la cual requiere de mayores cantidades de muestra, necesita de métodos de cuantificación precisa y fiable, y además, no corrige el posible error experimental cometido durante el proceso [26].

Así mismo, también presentan una serie de limitaciones a su uso:

- Pueden no presentar niveles de expresión constante bajo las condiciones de estudio

Aunque inicialmente se ha considerado que los genes HK debían presentar niveles de expresión constante, ya a finales de siglo XX, diferentes estudios recogen diferentes escenarios experimentales/contextos biológicos donde la expresión de varios genes HK de uso muy común se ve afectada (inducida o reprimida) por las condiciones condiciones fisiopatológicas de la célula [22, 25, 27], se hace evidente que la expresión de los genes HK no era estrictamente estable [28].

En la actualidad numerosos estudios han demostrado que los genes HK presentan variaciones significativas en los niveles de expresión en función de las condiciones celulares y en respuesta a diversos factores [26, 28, 29, 30, 31, 32] ni se expresan de forma constitutiva en todos los tejidos [21, 33].

En consecuencia, el gen HK adecuado como control interno para unas condiciones experimentales puede no ser el apropiado para otras, siendo esta la principal limitación a su uso genérico.

- Requieren de su validación para las condiciones de estudio

Ante la posibilidad de presentar una expresión variable, es necesario realizar un estudio previo para evaluar los niveles de expresión, validar la estabilidad de la expresión y seleccionar los genes más estables bajo las condiciones experimentales en las que se van a emplear como referencia. Este paso es indispensable para la confianza en los resultados obtenidos.

- La selección de genes con niveles de expresión variable bajo nuestras condiciones experimentales puede generar resultados erróneos

La consecuencia directa de emplear un gen de referencia con niveles de expresión fluctuante es la introducción de un error sistemático. Esto implica que podríamos no ser capaces de detectar, si existen, las diferencias más pequeñas que se puedan dar entre los genes de interés, perdiendo esta importante información [26].

- A nivel general, los genes control no se pueden emplear en contextos biológicos en los que la transcripción se ve comprometida

Existen condiciones histopatológicas en las que el proceso de la transcripción celular se ve afectado, se reduce parcialmente o no se lleva a cabo en su totalidad. En estos escenarios, lo más recomendable es escoger un método alternativo para la normalización de nuestros datos.

No obstante, estas limitaciones no los invalidan para su uso, más bien pone de manifiesto la necesidad de precaución y el conocimiento previo de su comportamiento bajo las condiciones de interés.

Diversos estudios han empleado datos cuantitativos obtenidos con varias metodologías de análisis del ARN (microarrays, marcadores de secuencia (EST) y RNA-Seq) y diferentes criterios estadísticos como el coeficiente de variación (CV), el ratio los valores de de expresión mínimos y máximos (*maximum fold-change* (MFC)<2) o que la expresión media ha de ser menos que el máximo valor de expresión obtenido dentro de dos desviaciones estándar, para analizar la variabilidad existente en los niveles de expresión génica de los genes HK a lo largo de los múltiples tejidos e identificar aquellos que presentan patrones de expresión más estables que permitan su proposición como genes candidatos a controles internos [26, 28, 29, 30, 31, 32, 34].

1.2.3. Métodos de validación y herramientas para la selección de controles internos

Ante la necesidad de seleccionar aquellos genes con expresión más estables para nuestras condiciones de interés, han surgido múltiples métodos matemáticos/estadísticos para determinar aquellos más óptimos y calcular un factor de corrección para la normalización. Los más comunes se han implementado en paquetes de R y en herramientas web que facilitan su uso por parte de los investigadores.

Entre los algoritmos que permite la identificación del gen óptimo para la normalización entre el set de genes candidatos, destacan:

- geNorm

Emplea la media geométrica para el cálculo del factor de normalización en base a los niveles de expresión de los mejores genes porque controla mejor los outliers y las diferencias en la abundancia entre diferentes genes [35]. Es una herramienta excel, su implementación en R se encuentra en el paquete NormqPCR.

- Normfinder

Normfinder crea un ranking en base a la estabilidad de la expresión en una muestra y un diseño experimental concretos, del set de posibles genes candidatos, calcula el factor

de normalización como la media geométrica de los genes de referencia considerados. Los genes que ocupen las primeras posiciones del ranking son los que introducirán el menor error sistemático cuando se usen para normalizar. Está implementado como una función en R se descarga de la web de MOMA (Medicina Molecular del Hospital Universitario Aarhus de Dinamarca, `https://moma.dk/normfinder-software`).

- Bestkeeper

Solo toma valores crudos de Ct de RT-qPCR como datos de entrada y calcula un coeficiente de correlación [36]. Valores de correlación (r) próximos a 1 implican mayor estabilidad en la expresión. Implementado en el paquete ctrlGenes como función BestKeeper.

Todas estas estrategias descritas estan incorporadas en la herramienta web RefFinder desarrollada para facilitar al investigador la evaluación de la variabilidad de los genes de referencia (`https://www.heartcure.com.au/for-researchers/`) [37].

Una vez evaluada la variabilidad en los niveles de expresión, la estrategia más común es normalizar a un solo gen con expresión constitutiva bajo nuestras condiciones experimentales, si un gen no es suficiente como referencia, se suelen emplear dos o más genes para la normalización. En su mayoría los estudios experimentales se limitan a diseños muy restringidos (uno o pocos tejidos, caso-control, control-tratamiento) por lo que es posible encontrar genes que expresen de forma estable [38].

1.3. Los repositorios públicos en el abordaje masivo de datos transcriptómicos

El gran volumen de datos generados con el aumento de la popularidad de las técnicas de alto rendimiento, debido al avance del conocimiento y al abaratamiento de los costes, han impulsado la creación de de múltiples bases de datos y repositorios públicos específicos, que responden a las necesidades de almacenamiento, manejo y accesibilidad de toda esta información; ya que sólo tiene sentido si tenemos las herramientas para integrarla y darle un significado biológico [39].

A fecha de enero 2020, según el informe anual sobre la colección de repositorios en la investigación de ácidos nucleicos [40] existen 1637 bases de datos. Los repositorios públicos más conocidos de datos de expresión génica son *Gene Expression Omnibus* (GEO) y *Array Express* [41].

1.3.1. Gene Expression Omnibus (GEO)

Gene Expression Omnibus (GEO, `https://www.ncbi.nlm.nih.gov/geo/`) es un repositorio público internacional que almacena datos de expresión génica obtenidos mediante microarrays y secuenciación masiva de ácidos nucleicos (NGS), y otras tecnologías de alto rendimiento [42]. Fue desarrollada por el NCBI (*National Center for Biotechnology Information*) [43], quién también se encarga de su mantenimiento, con el objetivo de ofrecer un espacio de almacenamiento estructurado de los datos generados con metodologías de alto rendimiento con una arquitectura flexible que permitiera adaptarse al desarrollo de nuevas tendencias. La idea de esta base de datos surge como respuesta a la necesidad de acceso libre a los datos de los estudios publicados para poder contrastar los resultados.

La arquitectura de GEO

La arquitectura de la base de datos GEO se compone de tres tipos de entidades relacionadas: las Plataformas, las Muestras y las Series (figura 1.5).

Una Plataforma es un registro que contiene la descripción de los elementos del array o secuenciador que definen el conjunto de moléculas que van a ser detectadas y cuantificadas en los experimentos que emplean esa plataforma. A cada registro se le asigna un número de accesión único (GPL). Una plataforma referencia a todas las muestras que la han empleado (relación 1:n).

Los registros de las muestras contienen dos partes: 1) una descripción de las características del material biológico, las condiciones de manipulación y el procesamiento experimental particular de cada muestra; y 2) las medidas de las moléculas detectadas y cuantificadas a partir de esta. A cada registro muestral se le asigna un número de accesión único (GSM), que referencia a una única plataforma (relación 1:n) y puede estar incluida en varias series (relación n:m, donde una muestra puede pertenecer a varias series, en una serie recoge varias muestras).

Las series agrupan el conjunto de muestras de un mismo estudio, incluyen la descripción general del estudio, un resumen de la metodología empleada y, opcionalmente, las conclusiones obtenidas. A cada serie se le asigna un número de accesión único (GSE).

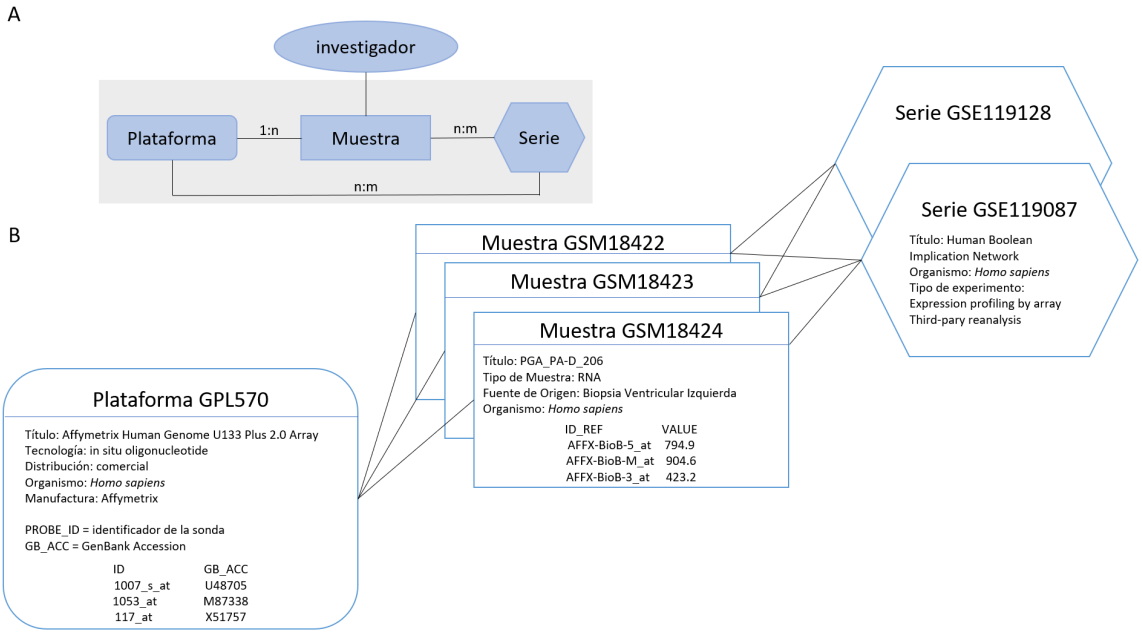

Figura 1.5: Esquema de la arquitectura de GEO. A. Relación entre los tres tipos de entidades que componen la base de datos y la fuente que proporciona los datos, el investigador. B. Ejemplo de registros de las tres entidades y sus atributos, las flechas indican la relación entre los objetos.

Gracias a la organización de la información GEO permite filtrar, localizar y obtener de forma manual o programática datos crudos y procesados, en diferentes formatos, relevantes para los intereses particulares de un estudio. Además, en la propia web se han incorpo-

rado diferentes herramientas para consultar, analizar y visualizar los datos, facilitando el procesamiento de la información a investigadores [44].

Tipos de datos que admite GEO

GEO admite una variedad considerable de datos generados principalmente mediante las siguientes tecnologías:

- *Gene expression profiling by microarray or next-generation sequencing.*
- *Non-coding RNA profiling by microarray or next-generation sequencing.*
- *Chromatin immunoprecipitation (ChIP) profiling by microarray or next-generation sequencing.*
- *Genome methylation profiling by microarray or next-generation sequencing.*
- *High-throughput RT-PCR.*
- *Genome variation profiling by array (arrayCGH).*
- *SNP arrays.*
- *Protein arrays.*

La tabla 1.1 recoge el número de registros que almacena GEO en el momento de redacción de este trabajo.

Tabla 1.1: Historial de GEO en agosto 2020.

	Público	Inédito	Total
Series	134,035	16,331	150,366
Plataformas	21,252	215	21,467
Muestras	3,741,022	621,131	4,362,153

1.3.2. La importancia de los estándares de la información

Muchas revistas y concesiones públicas tienen como requisito para publicar o financiar un estudio que sus datos estén alojados en GEO. Por su parte, GEO establece unos formatos para recepción de los datos experimentales que siguen los criterios mínimos de información MIAME y MINSEQE.

Los criterios MIAME (*Minimum Information About a Microarray Experiment*) [45], describen la información mínima necesaria para asegurar que los datos de expresión génica generados en experimentos con microarrays pueden ser interpretados sin ambiguedades y los resultados pueden ser verificados de manera independiente por terceras partes. El nivel de detalle de la información recogida bajo estos criterios permite interpretar el estudio, contrastar los resultados y reproducir el experimento, además la información está estructurada de modo que permite su filtrado mediante consultas programáticas y el procesamiento automatizado. Previamente a la existencia de este estándar, emular los

resultados de un estudio era complicado porque en la metodología explicada y la información suplementaria de los artículos se omitía información crucial [46].

Las directrices MINSEQE (*Minimum Information About a Next-generation Sequencing Experiment*), emitidas en el 2012, son el equivalente para los experimentos de RNA-Seq. Se encuentran recogidos en `http://fged.org/projects/minseqe`.

Estas directrices no son las únicas guías que recogen principios de transparencia de la información que se han desarrollado con el fin de facilitar la comprensión y la reproducibilidad de los resultados. En 2016 se publicaron los principios FAIR (*Findability, Accessibility, Interoperability and Reusability*) se aplican tanto al conjunto de datos generados como a los algoritmos, las herramientas y los flujos de trabajo que han producido esos datos. Siguiendo estos principios todos los componentes del proceso de investigación deben estar disponibles para asegurar la usabilidad de los datos publicados, así mismo, ponen el foco en la importancia de facilitar la localización y procesamiento automatizado de los datos [47].

1.3.3. Papel de los repositorios públicos

Los repositorios públicos como GEO recogen de forma organizada enormes cantidades de información biológica y conceden el libre acceso a la información almacenada, facilitan su usabilidad y permiten la reproducibilidad de los análisis y el contraste de los resultados. Estos tres aspectos son clave en tanto que los datos generados adquieren valor añadido al permitir su reanálisis y su reutilización por parte de la comunidad científica para desarrollar y probar nuevas hipótesis. Un ejemplo es el caso de los estudios que llevan a cabo un metaanálisis, donde la utilización de conjuntos de datos de diferentes estudios en un mismo contexto clínico incrementa la robustez y la confianza en los resultados al aumentar el número de muestras, o los estudios *in silico* (i.e análisis, simulaciones y modelizaciones computacionales).

Esta corriente de reutilización de la información ha impulsado el desarrollo de una gran variedad de herramientas que emplean registros de repositorios públicos con el fin de aprovechar la enorme cantidad de datos disponibles para generar nuevo conocimiento.

Para GEO concretamente, se encontramos múltiples programas y herramientas web con diferentes funcionalidades:

- Software de consulta de la información disponible en la base de datos como GEOmetadb [48].

- Software de descarga de registros como GEOQuery [49].

- Software de reanálisis de registros individuales, como por ejemplo GEO2R o shinyGEO, o colecciones de registros, como ScanGEO e imaGEO, que llevan a cabo análisis de expresión diferencial de forma, entre otros [50] (ver tabla 1.2).

Tabla 1.2: Herramientas de GEO. Tabla adaptada de Wang et al. 2019 [50].

Herramienta	Análisis	Tipo
GEO2R	Individual	Web
shinyGEO	Individual	Web
GEOquery	Individual	Paquete R
GEO2Enrichr	Individual	Extensión web
BioJupies	Individual	Web
ScanGEO	Multiple	Web
ImaGEO	Multiple	Web
GEOracle	Multiple	Web

1.4. La perspectiva del sexo en Biomedicina

El sexo de un individuo influye en la expresíon génica afectando muchos procesos biológicos [51]. A nivel clínico, estas variaciones entre sexos se traducen en diferencias en el riesgo, el nivel de afectación, la progresión y la eficacia del tratamiento de muchas enfermedades. Se conocen diferencias en la respuesta inmune, en enfermedades cardiovasculares, neurológicas, en la susceptibilidad a infecciones, el metabolismo de fármacos, entre otras [52].

No obstante, a pesar de ser conscientes, existe la preferencia de considerar únicamente uno de los dos sexos en los diseños experimentales, en su mayoría individuos masculinos, para realizar los estudios biomédicos con pacientes o modelos animales. Es por ello que es necesario poner en consideración estas diferencias relacionadas con el sexo.

Incluir la perspectiva del sexo en el diseño experimental es fundamental para evidenciar la existencia de diferencias en el objeto de estudio asociadas al sexo, o por el contrario, determinar que ambos sexos no presentan diferencias en ese rasgo [53]. En caso de encontrar diferencias se abren nuevas vías para entender los mecanismos diferenciales de la enfermedad y adecuar las correspondientes etapas de diagnóstico y tratamiento. Un ejemplo de esto es el uso de la genómica para la detección de biomarcadores específicos.

2. Objetivos

Los resultados de expresión obtenidos tras la normalización varían en función del gen empleado como referencia que utiliza en el proceso de normalización. El análisis de datos de microarrays, mediante diferentes métodos estadísticos, ha demostrado su utilidad para evaluar la expresión de los genes *housekeeping* y proponer nuevos genes candidatos que muestran expresión constante en diferentes tejidos y condiciones [29, 28, 30, 31, 32, 21, 54]. Hasta el momento se han realizado aproximaciones revisitando los datos disponibles de un número reducido de estudios [34].

En este trabajo se propone la evaluación de los niveles de variabilidad de la expresión génica en diferentes tipos celulares y condiciones del tejido adiposo de *Homo sapiens* y *Mus musculus*, mediante el abordaje masivo de los datos de microarrays disponibles en GEO para identificar aquellos genes que muestran mayor estabilidad en su expresión independientemente de las condiciones fisiopatológicas muestra. Empleamos tres estimadores diferentes de la variabilidad relativa para cada gen y en cada estudio, e integramos los resultados para incrementar la cobertura muestral, tratando el efecto "batch", el posible error experimental y los valores extremos. Además, incluimos el sexo como variable en el estudio. Ponemos especial interés en revisar los perfiles de expresión de los genes metabólicos: HPRT, PPIA, UBC, GAPDH y los genes ribosomales: 18S y RPL19, de uso habitual como controles endógenos en análisis cuantitativo de transcritos. Estos genes han sido propuestos por la Dra. Amparo Galán, investigadora especialista en Neuroendocrinología Molecular, que también ha detectado la variación en sus resultados de expresión dependiente del gen empleado como referencia para la normalización. Cuantificada la variación, podemos conocer el error sistemático que se está introduciendo en la normalización de la expresión. Con todo ello:

El objetivo de este trabajo es evaluar la estabilidad de expresión de los genes HK clásicos de uso común en estudios de expresión génica de tejido adiposo e identificar nuevos candidatos. Para lograrlo, se han propuesto los siguientes cuatro objetivos específicos:

1. Obtención la información de todos los estudios con datos transcriptómicos de tejido generados con microarrays disponibles en GEO adiposo en Humano y Ratón.

2. Determinación del nivel de variabilidad en la expresión de los genes metabólicos: HPRT, PPIA, UBC, GAPDH y los genes ribosomales: RNA18S, RPL19; para el conjunto de todos los estudios que analicen el tejido adiposo cuyos datos de microarrays están disponibles en GEO.

3. Detectción de diferencias en los perfiles de expresión asociados con el sexo.

4. Identificación de nuevos genes de referencia candidatos.

3. Material y métodos

Todo el análisis bioinformático descrito a continuación (figura 3.1) se ha realizado mediante scripts propios con los lenguajes de programación Python 3.0 y R [55]. El código y las funciones desarrolladas estan disponibles en github (`https://github.com/mguaita/Eval_HKG`), así como todos ficheros generados para el preprocesamiento de la información de GEO. Todas las librerias y paquetes indicados y los scripts empleados cada paso se detallan respectivamente en las tablas B.1 y B.2 del Anexo B.

Debido al gran volumen de datos con los que trabajamos, se ha necesitado el uso de una infraestructura computacional que presente características de almacenamiento y cómputo superiores a las que presenta un ordenador personal. Por ello en este trabajo se empleó el clúster de cómputo del Centro de Investigación Príncipe Felipe (CIPF) que cuenta con 44 nodos computacionales, 600 unidades centrales de procesamiento (CPUs) y 11 TeraBytes de memoria RAM. Posee un sistema de ficheros distribuido denominado Lustre y un sistema de gestión SLURM.

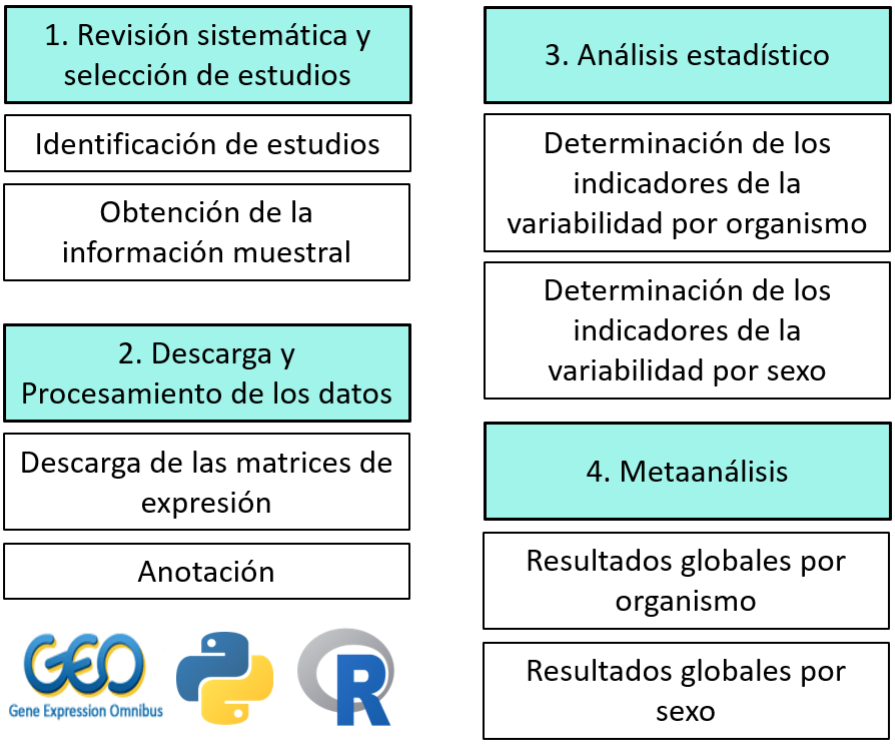

Figura 3.1: Flujo de trabajo seguido.

3.1. Revisión sistemática y selección de los estudios

3.1.1. Identificación de los estudios disponibles en GEO por tejido y organismo

Algunos de los estudios disponibles en GEO con muestras de tejido adiposo emplean varias plataformas e incluyen diferentes tipos de información biológica como polimorfismos de nucleótido único (*SNPs*), variantes del genoma, micro-ARNs o metilación, así como datos de otros tejidos, obtenidos con microarrays u otras tecnologías como RNA-Seq u otros organismos. Esta naturaleza múltiple de los datos incrementa la complejidad de la selección y la descarga de los datos de interés en nuestro estudio.

Hemos desarrollado una metodología de procesamiento y estructuración de la información de los registros de GEO, que permite obtener los datos cuantitativos de expresión génica de muestras de tejido adiposo producidos con microarrays para, en última instancia, analizar estadísticamente la variabilidad de la expresión de cada gen específico. Cabe tener en cuenta que cada chip de microarray contiene una batería de sondas propias empleada para medir la expresión génica, y que comparten muchos genes pero no la totalidad. Por ello, el procesamiento de los estudios se ha de realizar a nivel de plataforma.

Así pues, en primer lugar necesitamos identificar todos los estudios disponibles en GEO con muestras de tejido adiposo que han realizado un análisis cuantitativo de la expresión génica mediante microarrays y las plataformas que han utilizado. Se ha escogido la opción de *web scrapping* frente a la alternativa de emplear el paquete GEOmetadb de R, el cual permite obtener una imagen local de la base de datos de GEO y acceder a los registros mediante consulta SQL, porque la información contenida en el esquema descargado se ha encontrado incompleta respecto de los resultados devueltos directamente desde el repositorio web.

Se ha llevado a cabo una búsqueda avanzada en la web de GEO para obtener todos los estudios que contienen datos cuantitativos de expresión génica de muestras del tejido adiposo de humanos y ratones obtenidos con microarrays (figura 3.2). Introducimos el origen de la muestra, el tipo de estudio y el organismo de interés con las palabras clave "gse", "adipose", "expression profiling by array", "*Homo sapiens*" y "*Mus musculus*" en el buscador.

Las consultas realizadas han sido las siguientes:

- Para *Homo sapiens* (en marzo 2020):
  "*Homo sapiens*"[organism] AND "adipose"[sample source] AND "expression profiling by array"[DataSet Type] AND "gse"[Entry Type]
- Para *Mus musculus* (en junio 2020):
  "*Mus musculus*"[organism] AND "adipose"[sample source] AND "expression profiling by array"[DataSet Type] AND "gse"[Entry Type]

Estas consultas devuelven todos los estudios (GSE) con datos de microarrays ("expression profiling by array") que contienen muestras de tejido adiposo, indicado explícitamente en el atributo "sample source" del registro muestral, de *Homo sapiens* y *Mus musculus*, respectivamente. Consideramos únicamente aquellos estudios con muestras de tejido adiposo

cuyo origen se haya indicado de forma explícita con palabra clave "adipose" en el atributo "sample source". Este es un campo de texto libre donde se anota el origen del material biológico, está redactado por el investigador que facilita los datos, por lo que puede no incluir esta informacón. De esta forma, aplicamos un filtro restrictivo que nos asegura que las muestras incluidas en el estudio pertenecen al tejido adiposo.

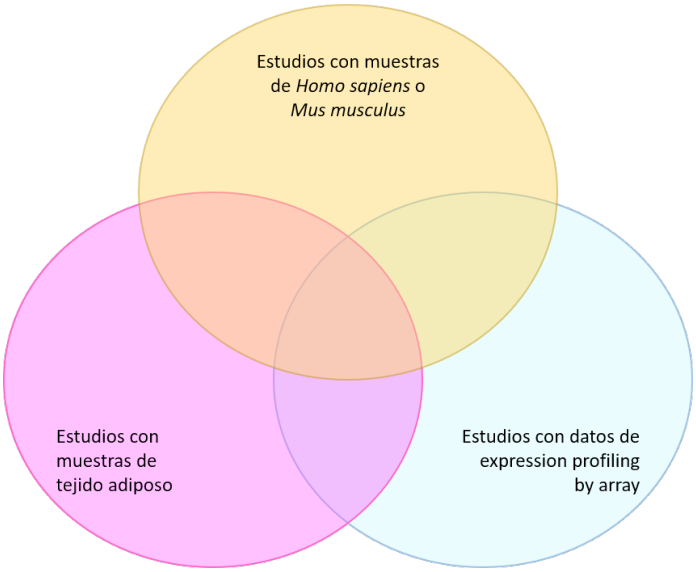

Figura 3.2: Estudios de interés en GEO. Los estudios que nos interesan son aquellos con datos de perfiles de expresión de muestras de tejido adiposo de humanos y ratones obtenidos con microarrays.

El resultado de cada consulta se guarda localmente en forma de fichero HTML que procesamos con la librería BeautifulSoup de Python (script 1), la cual nos permite acceder a las etiquetas HTML y extraer la información clave: los identificadores de los estudios (GSE), los identificadores de las plataformas empleadas (GPL) y el tipo de estudio. La información obtenida se escribe en un fichero csv con la estructura de la tabla 3.1.

Tabla 3.1: Estructura de los resultados de la búsqueda de GEO. Se muestra un fragmento de la tabla generada con el script *webscrapGEO* de Python que estructura la información contenida en el fichero HTML resultado del filtrado web de GEO. Este extracto pertenece al fichero *adipose_GSEs.csv* de *Homo sapiens*.

GSE	GPL	Study_type
GSE62832	GPL6244	Expression profiling by array
GSE25910	GPL6244	Expression profiling by array
GSE25401	GPL6244	Expression profiling by array
GSE41223	GPL6244, GPL8786	Expression profiling by array, Non-coding RNA profiling by array
GSE25402	GPL6244, GPL8786	Methylation profiling by array, Expression profiling by array, Non-coding RNA profiling by array
GSE56635	GPL6246, GPL6244	Expression profiling by array

La revisión del tipo de estudio es relevante porque todos los estudios que pasan el primer filtro contienen datos de microarrays y, además, pueden incluir otro tipo de información biológica obtenida con diferentes tecnologías. Como se ve en la tabla 3.1, tenemos estudios que recogen datos de más de una plataforma y no todas las plataformas se corresponden con microarrays. Por ello, debemos conocer a qué descripción corresponden los identificadores de las plataformas para identificar aquellas plataformas pertenecientes a microarrays y, dado que el procesamiento estadístico se realiza a nivel de plataforma, agrupar los estudios en este nivel para dirigir posteriormente el flujo de trabajo.

Procesamos la información extraída del fichero HTML y generamos una tabla que relacione los identificadores de los estudios y los identificadores de las plataformas, donde la clave primaria es la combinatoria de identificadores estudio-plataforma (script 2). Como podemos ver en el extracto de la tabla 3.2, se trata de una relación n:m donde un estudio tiene tantas entradas como el número de plataformas que emplea, y una plataforma tantos registros como estudios la utilizan, de este modo podemos cuantificar el número de estudios que emplean una misma plataforma con SQL.

Tabla 3.2: Tabla relacional GSE-GPL. Fragmento de la tabla generada con el script *GSE_GPL_relational* de Python, relaciona los identificadores de los estudios con las plataformas que emplean. Su procesamiento nos permite realizar el conteo del número de estudios por plataforma de forma sencilla con una consulta SQL.

GSE	GPL
GSE43346	GPL96
GSE43346	GPL97
GSE43346	GPL570
GSE27121	GPL6947
GSE17170	GPL570
GSE43642	GPL11532
GSE99316	GPL96
GSE99316	GPL97
GSE99316	GPL570
GSE99316	GPL10999
GSE66159	GPL570
GSE101492	GPL20265
GSE72158	GPL10558
GSE19494	GPL4133
GSE43471	GPL6947

Esta estructura de la información facilita los identificadores únicos de las plataformas así podremos obtener las descripciones de las plataformas, así como, agrupar y cuantificar el número de estudios en los que se ha utilizado cada plataforma, para identificar aquellas más comunes.

Empleamos el paquete sqldf de R para procesar la tabla relacional (script 3):

- Extraemos los identificadores de todas las plataformas en forma de vector, los unificamos en una cadena ("string") y construimos la consulta sql que nos devuelve sus descripciones con el paquete GEOmetadb (Anexo A Query 1).
- Cuantificamos los estudios pertenecientes a cada plataforma.
- Escribimos los identificadores de las plataformas con su descripción respectiva y el conteo de los estudios en formato csv.

De esta forma conseguimos a partir del filtro de la web de GEO, identificar los estudios con muestras de tejido adiposo analizadas con microarrays, conocer las plataformas correspondientes a microarrays y cuantificar el número de estudios que emplean cada plataforma.

3.1.2. Selección de las principales plataformas

Ante el gran número de plataformas diferentes empleadas en los estudios se acota el analisis al top de plataformas que incluyen el mayor número de estudios: 4 plataformas para humanos y 5 plataformas para ratones.

3.1.3. Extracción de la información muestral con GEOmetadb

Información muestral de los estudios por organismo

En el conjunto de estudios disponibles de las plataformas seleccionadas, tenemos estudios que analizan la expresión génica exclusivamente en tejido adiposo (monotejido) y estudios que investigan además otros tejidos (multitejido). El pre-procesamiento de los datos para el análisis estadístico es diferente en cada caso, para los estudios que sólo examinan tejido adiposo utilizamos directamente toda la matriz de expresión génica, en cambio, para los estudios que incluyen más tejidos necesitamos identificar qué muestras corresponden a tejido adiposo y extraer sus valores de dichas matrices.

Empleamos el listado de identificadores GSE de los estudios de cada plataforma para construir las consultas de GEOmetadb que nos devuelvan los identificadores de todas las muestras (GSM) de cada estudio, su origen y el organismo al que pertenecen (Anexo A Query 2). La información extraída de cada consulta se incorpora a una tabla que empleamos para clasificar los estudios en "monotejido" o "multitejido", en función de si el origen de todas las muestras está anotado como tejido adiposo con la palabra clave "adipo" en el campo "sample source" o no. Para ello, contamos el total de muestras del estudio de una plataforma y el número de muestras de dicha plataforma cuya descripción contenga la palabra 'adipo', si el número es el mismo se trata de un estudio centrado en tejido adiposo, en caso contrario se trata de un estudio que abarca más tejidos. Se ha creado la variable "indicadorMT" para detallar este dato que posteriormente nos será de utilidad para dirigir el flujo de trabajo. Además, para los estudios que incluyen muestras de varios tejidos, apuntamos los identificadores GSM de las muestras que pertenecen a tejido adiposo (script 4).

Los resultados de cada plataforma se anotan en un fichero csv resumen, que denominamos *Summary*, en el cual se recoge toda la información de los estudios elegibles para el análisis estadístico:

- columna *GSEIDs*: identificadores de los estudios pertenecientes a esa plataforma.
- columna *nSamples_tot*: número de muestras del estudio que se han analizado con esa plataforma.
- columna *nSampples_adi*: número de muestras del estudio que pertenecen a tejido adiposo.
- columna *indicadorMT*: informa de si es un estudio "monotejido" o "multitejido".
- columna *adiGSMs*: identificadores GSM de las muestras de tejido adiposo en caso de ser multitejido.

Esta estructura descrita de la información (ver tabla 3.3) nos ofrece una imagen completa de los datos disponibles en GEO de una misma plataforma con los que podemos trabajar. Repetimos este proceso con cada una de las plataformas que vamos a analizar, y generamos para cada una de ellas su fichero resumen.

Tabla 3.3: Fichero *Summary* de la plataforma GPL570. Fragmento de la tabla summary de la plataforma GPL570 de *Homo sapiens* que recoge todos los estudios de la plataforma con muestras de tejido adiposo en GEO. Podemos ver estudios con muestras exclusivamente de tejido adiposo (monotejido) y estudios que incluyen varios tejidos (multitejido) con los identificadores correspondientes a las muestras de tejido adiposo.

GSEIDs	nSamples_tot	nSamples_adi	indicadorMT	adiGSMs
GSE27916	375	375	MONOTEJIDO	ALL
GSE13070	364	84	MULTITEJIDO	'GSM342612', 'GSM342613', (...)
GSE17170	70	70	MONOTEJIDO	ALL
GSE41168	140	70	MULTITEJIDO	'GSM1009752', 'GSM1009753', (...)
GSE39117	54	54	MONOTEJIDO	ALL
GSE39118	54	54	MONOTEJIDO	ALL
GSE26339	50	50	MONOTEJIDO	ALL
GSE115799	49	49	MONOTEJIDO	ALL
GSE13506	49	49	MONOTEJIDO	ALL
GSE28005	38	38	MONOTEJIDO	ALL

Concretamos el número de muestras del estudio que se han analizado con la plataforma de interés porque las muestras de un estudio son propias de la plataforma con las que se han analizado. Es decir, en un estudio tenemos: el número total de muestras del estudio, el número de muestras total de la plataforma de interés en el estudio y el número de muestras de la plataforma de interés que pertenecen a tejido adiposo en el estudio.

Como ya se ha mencionado, la imagen de GEOmetadb no incluye los registros de varios estudios, la información muestral ausente se ha completado manualmente directamente de la web de GEO.

Identificación del sexo muestral

Queremos diferenciar las muestras en función del sexo con el objetivo de identificar patrones diferenciales en la variabilidad de expresión. Para ello, en primer lugar, se ha de identificar aquellos estudios que incluyen información del sexo en las caracteristicas de las muestras y después, obtener los identificadores de las muestras correspondientes a hombres/mujeres en humanos y machos/hembras en ratones.

Las características de la muestra es otro de los campos de texto libre de los registros de GEO, ante la variabilidad encontrada en la anotación del sexo, se ha optado por especificar como se incluye esta información en cada estudio. Se han revisado todos los estudios disponibles de cada plataforma de humano y ratón para determinar aquellos que incluyen la información sexo del material biológico empleado. Este proceso se ha realizado manualmente a través de la web de GEO. En la tabla GSE_GPL_sexsamples (ver muestra de la plataforma GPL570 en la tabla B.3 del Anexo B) se ha incorporado la información de todos los estudios necesaria para identificar las muestras pertenecientes a cada sexo y dirigir el flujo de procesamiento de la información muestral de forma programática:

- Identificador GSE del estudio.
- Identificador GPL de la plataforma.
- Número de muestras de tejido adiposo de la plataforma.
- indicadorMT.
- InformaciónSexo, sí en algún campo de texto libre del registro del estudio se informa del sexo de las muestras (SI/NO) y podemos identificarlas claramente.
- MuestrasSexo, en caso de informar del sexo muestral, qué sexos se han incluido en el estudio (HOMBRES/MUJERES/AMBOS).
- AnotaciónSexo, si el sexo de la muestra está correctamente anotado en las caracteristicas de la muestra del registro muestral (SI/NO), este dato es relevante porque se puede acceder a la información de este campo con GEOmetadb.
- *Sex_ tags*, en caso de estar anotado el sexo en las caracteristicas de la muestra, qué palabras clave se han empleado en el estudio que nos permiten identificar claramente el sexo de la muestra.

Podemos observar que tenemos un porcentaje considerable de estudios que no incluyen la información del sexo de las muestras. Continuamos el análisis con todos los estudios que sí facilitan este dato, tenemos estudios que han investigado con un solo sexo y estudios que han incluido ambos.

Con la información recogida en esta tabla, filtramos aquellos estudios que incluyan la información del sexo ejecutando la siguiente consulta con el paquete sqldf (script 5):

```
SELECT GSEID, GPL, nSamples_adi, indicadorMT, muestrasSexo, anotacionSexo,
sex_tags
FROM GSE_GPL_sexsamples
WHERE InformacionSexo = "SI"
```

Para cada estudio construimos la consulta específica que nos devuelva toda la información del registro muestral para la plataforma asociada (Anexo A Query 3). De nuevo la información que no está recogida en GEOmetadb se completa manualmente. Una vez obtenida la información de las muestras del estudio, procesamos diferencialmente aquellos estudios que incluyen muestras de ambos sexos de aquellos que solo incluyen uno, empleamos la variable muestrasSexo para dirigir el flujo de procesamiento.

Si el estudio incluye muestras de ambos sexos, empleamos la etiquetas de identificación del sexo (*sex_ tags*) específicas de cada estudio para obtener las muestras de tejido adiposo correspondientes a cada sexo con las consultas:

- SELECT gsm, characteristics
  FROM sampleinformation_df
  WHERE
  sample_source LIKE " %adipo %" AND
  characteristics LIKE " %male_sextag %" AND
  organism LIKE " %*Homo sapiens* %"

- SELECT gsm, characteristics
  FROM sampleinformation_df
  WHERE
  sample_source LIKE " %adipo %" AND
  characteristics LIKE " %female_sextag %" AND
  organism LIKE " %*Homo sapiens* %"

De este modo, obtenemos los identificadores GSM de todas las muestras en cuyo apartado de características de la muestra esten incluidas las palabras clave "adipo" clasificándolas mediante la etiqueta específica con la que está anotada el sexo. La etiqueta de hombre-macho y mujer/hembra se separa previamente formateando la cadena de *sex_ tag* por el caracter "///", diferenciando las palabras clave que identifican el sexo de la muestra en cada uno de los estudios.

Para los estudios que solo incluyen muestras de un solo sexo no es necesario obtener los identificadores GSM, se incorporan directamente todas las muestras del estudio al bloque del sexo correspondiente.

Por último, incorporamos toda la información a la tabla anterior para finalmente realizar el análisis estadístico. La tabla B.4 en el Anexo B detalla la información recogida de todos los estudios de *Homo sapiens* que informan del sexo de las muestras.

3.1.4. Criterios de exclusión de los estudios

La revisión sistemática es una metodología desarrollada para responder a una pregunta concreta contrastando toda la evidencia empírica que cumple con unos criterios de selección definidos previamente. Se caracteriza por tener unos objetivos claros que organizan la búsqueda de los estudios que cumplan con los criterios específicos y la evaluación de su calidad. En el desarrollo de este trabajo se han seguido las directrices y flujo de trabajo propuestos en la declaración PRISMA [56, 57], se han establecido una serie de criterios para la selección de los estudios orientados a la automatización del proceso de análisis de la información (figura 3.3).

Dentro de los estudios disponibles en la base de datos de GEO, con datos de expresión génica del tejido adiposo obtenidos con microarrays, algunos incluyen otros tipos de información biológica como datos sobre polimorfismos de nucleótido único (SNPs), variantes del genoma, micro-ARNs o metilación, obtenidos con arrays u otras tecnologías como RNA-Seq, o datos de otros tejidos u organismos; que tenemos que identificar y desechar.

De los estudios que se han obtenido en el filtrado de la web se han seleccionado aquellos que empleaban las principales plataformas de microarrays de análisis del transcriptoma completo, es decir, aquellas que se han empleado en el mayor número de estudios.

Seguidamente, de los estudios elegibles para el análisis estadístico, se han descartado aquellos estudios que:

- **Criterio 1:** No tienen muestras de tejido adiposo para la plataforma que estamos procesando.

Esta situación ocurre con los estudios que analizan varios tejidos con diferentes plataformas, han pasado la búsqueda exhaustiva de GEO porque poseen datos de microarrays y muestras de tejido adiposo, pero estas muestra se han interrogado experimentalmente con otra plataforma que no es la que estamos procesando.

- **Criterio 2:** Contienen muestras duplicadas de otros estudios.

Las superseries engloban registros muestrales de varios estudios individuales, las identificamos en la revisión de los estudios que se van a procesar por plataforma porque comparten los identificadores GSM de las muestras.

- **Criterio 3:** Presentan menos de 10 muestras de tejido adiposo ($n<10$).

Cuando se consideran diferentes fisiologías y tipos celulares, se necesita un número suficiente de muestras para obtener unos resultados representativos del transcriptoma existente [21], empleamos un tamaño muestral mínimo por estudio de 10.

- **Criterio 4:** No presentan sus resultados de forma estandarizada.

En casos muy concretos, los estudios no presentaban los valores de expresión génica en el formato de matriz indicado por GEO y en consecuencia no se ha podido acceder a ellos de forma programática ni interpretarlos para incorporarlos manualmente.

Los estudios elegibles de cada plataforma se recogen en las tablas summary específicas, los estudios incluidos finalmente en el análisis estadístico se recogen en las tablas *ProcessedGSEs* (con la misma estructura que las tablas *summary*). El resumen de todos los estudios elegibles, incluidos y excluidos del análisis estadístico para humanos y ratones se detallan, respectivamente, en las tablas B.5 y B.6 del Anexo B.

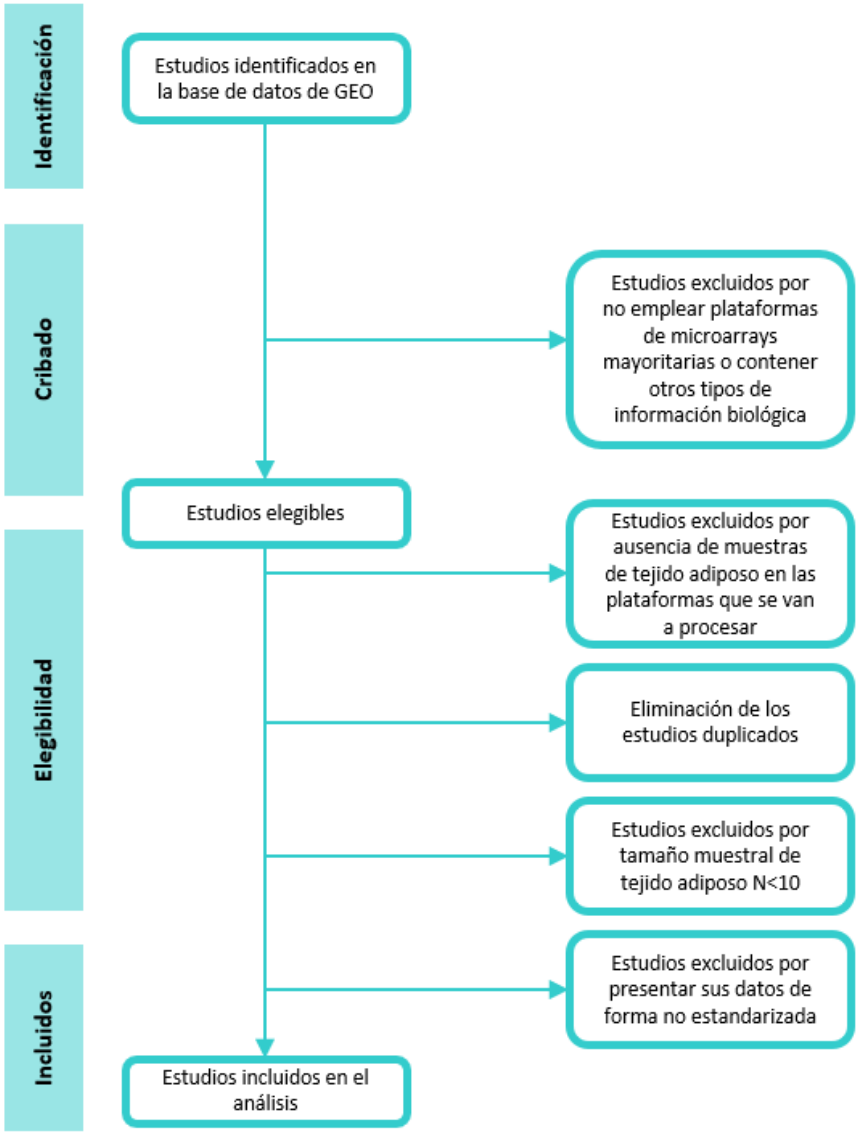

Figura 3.3: Revisión sistemática. Flujo de la información en las diferentes fases de la revisión sistemática. Esquema PRISMA adaptado de Liberati, A et al. 2009 [56].

3.2. Descarga y procesamiento del conjunto de los datos

Los siguientes pasos se repiten para el conjunto de estudios seleccionados de cada plataforma. Todo el procesamiento de datos se ha llevado a cabo con el lenguaje de programación R desde Rstudio empleando como información de entrada el identificador GPL de la plataforma y la información de los estudios elegidos contenida en la tabla *ProcessedGSEs*.

3.2.1. Descarga de los datos de los estudios

Para la descarga de los datos se ha creado la función ***getGEOdata*** (script 6), la cual a partir del listado de los identificadores GSE y el identificador GPL de la plataforma que se quiera procesar, llama a la función *getGEO* del paquete GEOQuery, descarga las matrices de expresión génica normalizadas de los estudios introducidos en forma de objeto ExpressionSet y selecciona aquellas pertenecientes a la plataforma de interés. Esta función nos permite automatizar la obtención de los datos todos los estudios sin distinguir aquellos que emplean más de una plataforma.

3.2.2. Anotación del conjunto de datos

La anotación de las sondas es específica del modelo y la casa comercial del microarray, para cada una de las plataformas se ha programado una función propia de conversión de los códigos de las sondas a un identificador común, el código Gene Symbol, a partir del fichero de anotación disponible en GEO. Se ha comparado con la anotación que ofrecen las librerías de anotación de Bioconductor [58] y se ha decidido continuar con la información de GEO por ser más completa.

La función ***get_geneExpression*** (script 6), engloba el proceso de anotación, toma como argumentos los identificadores del estudio y la plataforma, el objeto ExpressionSet del estudio, el indicadorMT y los identificadores GSM de las muestras de tejido adiposo en caso de ser un estudio multitejido. Esta información se encuentra en los ficheros processedGSEs de la plataforma que se analice. En cada iteración del bucle realiza la anotación de la matriz de expresión de un estudio:

1. Para los estudios que analizan varios tejidos (indicadorMT = "multitejido"), se extraen las muestras pertenecientes a tejido adiposo con los identificadores GSM correspondientes.

2. Se corrigen las medidas de expresión negativas debidas al método de normalización empleado por los investigadores.

3. Llama a la función de anotación expecífica de cada plataforma.

4. Se realiza el tratamiento de registros duplicados.

5. Llama a la función de escritura de la matriz anotada en el directorio de resultados (***writeGSE_cleandata***).

La conversión introduce duplicados debido a la presencia de varias sondas del array que hibridan con diferentes regiones de un mismo gen. Con la función aggregate, colapsamos la matriz de expresión agrupando los valores por el identificador Symbol y calculamos la mediana de los niveles de expresión de todas las sondas que apuntan al mismo gen, para quedarnos con un único valor de expresión de gen por muestra.

Finalizado el proceso, de cada estudio obtenemos la matriz con valores únicos de expresión de todos los genes analizados por la plataforma en cada muestra de tejido adiposo. Además, se informa por pantalla del número máximo de genes que hemos sido capaces de identificar, este dato se utiliza posteriormente para construir la matriz final del procesamiento estadístico.

La función ***getwrite_ cleanGeneExpressionDF***, se ha desarrollado para dirigir el tratamiento diferencial de los estudios con muestras de varios tejidos, formatea los identificadores de las muestras de tejido adiposo de cadena a vector para que la función ***get_ geneExpression*** pueda extraer directamente sus valores de las matrices de expresión.

3.2.3. Revisión de la anotación de los genes housekeeping

Por otro lado, se ha llevado una revisión manual exhaustiva de los diferentes ficheros de anotación de GEO (*soft-table* y *full annotation table*) y los correspondientes paquetes de anotación correspondientes de Bioconductor de las sondas en estas plataformas. Hemos unificado la información de las diferentes fuentes incorporando al FeatureData (fData) del ExpressionSet de un estudio completo de cada plataforma la conversión de los identificadores de las sondas a identificador Symbol realizada con Bioconductor como una columna más. Se ha comprobado el conjunto de la información de las sondas que apuntaban a estos genes, se ha comparado los datos que ofrecían las diferentes y se han encontrado conflictos importantes en la anotación de algunos de los genes *housekeeping* de interés. Ante este escenario, se ha contrastado la descripción de las sondas y sus dianas en el genoma con el NCBI.

Todas las excepciones encontradas se han corregido en las funciones de anotación específicas de cada plataforma. Cabe mencionar que en el código Symbol los genes de humanos están anotados en letras mayusculas, mientas que en ratón presentan la primera letra en mayuscula y el resto en minuscula, estas diferencias entre organismos se han respetado en el proceso de anotación.

3.3. Análisis estadístico individual de cada estudio

El análisis estadístico para cada estudio se llevó a cabo considerando todas las muestras disponibles de los diferentes grupos experimentales.

3.3.1. Determinación de los indicadores de variabilidad de expresión

Empleamos tres estimadores de la variabilidad relativa de cada gen en cada estudio, para identificar qué genes presentan una mayor o menor estabilidad a lo largo de todas las muestras de los diferentes estudios, finalmente resumimos los valores obtenidos en todos los estudios para una misma plataforma.

- **Coeficiente de Variación**

El primer estimador empleado es el coeficiente de variación (C.V.) relaciona la media aritmética y la desviación estándar describiendo la variabilidad de una variable, se calcula con la siguiente fórmula:

$$CV = \frac{\sigma}{\overline{x}} \tag{3.1}$$

donde σ es la desviación estándar y $\overline{x}$ la media aritmética.

Este parámetro es interesante en tanto que es adimensional, nos permite comparar valores a diferentes escalas de magnitud, y no posee unidades, a diferencia de la desviación standard que tiene las unidades de la media. No obstante, al basarse en la media, al igual que esta, se ve afectado por valores extremos. Por este motivo se ha decidido ampliar la perspectiva empleando otros estadísticos de variación basados en la mediana que son más robustos en distribuciones sesgadas [59].

- **IQR/median**

Esta alternativa emplea en el numerador el rango intercuartílico ($IQR = Q3 - Q1$), es una medida de la dispersión de los datos, equivalente a la diferencia entre los percentiles 75 y 25, y la mediana en el denominador como medida de centralidad. Un factor de corrección de 0.75 hace este estadístico comparable al C.V para una distribución normal.

$$RCV_Q = 0{,}75 \times \frac{IQR}{median} \tag{3.2}$$

- **MAD/median**

Emplea la mediana de la desviación absoluta, definida como $MAD = median(|X_i - \widetilde{X}|)$, y la mediana en el denominador. El valor obtenido se multiplica por el factor 1.4826 para generar la equivalencia entre la mediana de la desviación absoluta y la desviación estándar en un modelo normal.

$$RCV_M = 1{,}4826 \times \frac{MAD}{median} \tag{3.3}$$

Para cada estadístico se ha implementado una función independiente. A partir de las matrices anotadas con las medidas de expresión génica de las muestras y conociendo el número de muestras de cada estudio, realizamos el cálculo estadístico para cada gen, de modo que los valores obtenidos nos describen su nivel de variabilidad dentro de un mismo estudio. Se generan tres matrices, una para cada estadístico, que recogen el resultado de cada gen en todos los estudios de una misma plataforma.

Se ha creado la función ***geneExpressionDF_loopprocesssig*** que engloba la lectura de las matrices de expresión anotadas y el procesamiento estadístico de los datos. Toma como argumentos el listado de identificadores GSE de los estudios de la plataforma que se está analizando, el identificador GPL de la plataforma, el nombre del estadístico que se quiere calcular y el número máximo de genes que somos capaces de identificar en dicha plataforma. En cada iteración:

1. Llama a la función de lectura ***geneExpressionDF_reader***, que accede al directorio de la plataforma que se esté analizando e importa al entorno de trabajo la matriz de expresión del estudio.

2. Llama a la función propia del estadístico que se ha pedido y obtiene su valor para todos los genes del estudio.

3. Llama a la función ***add_resPAR_to_finalRes***, que almacena los resultados obtenidos incorporando una columna con el nombre del estudio a la matriz final.

El número máximo de genes que somos capaces de anotar define el número máximo de filas que vamos a tener en la matriz final, conocer este dato nos permite construir esta matriz añadiendo en cada iteración los resultados de cada estudio. Facilita el procesamiento emplear en la primera iteración un estudio completo con datos del máximo número de genes de la plataforma, asimismo, en caso de procesar un estudio que no tenga medidas de expresión de todos los genes de la plataforma, los valores ausentes se incorporan como valores nulos. Finalizado el bucle, tenemos una matriz que recoge los valores obtenidos del estadístico descriptivo de la variabilidad de todos los genes en todos los estudios de una misma plataforma, donde las filas se corresponden con los genes y las columnas a los estudios.

Todas las funciones descritas se encuentran implementadas en el script 6. Para su ejecución es necesario definir la plataforma con la que se va a trabajar y la lectura del fichero ProcessedGSEs específico de la plataforma para extraer la información relevante: los identificadores GSE, el indicadorMT y los identificadores GSM de las muestras de tejido adiposo. Con esta información podemos ejecutar el cuerpo principal del programa que llama a todas las funciones de anotación anidadas en *getwrite_ cleanGeneExpressionDF*, conocer el número máximo de genes que somos capaces de identificar, y realizar el cálculo individual de los tres estadísticos con la función *geneExpressionDF_ loopprocessing*.

3.3.2. Determinación de los indicadores de variabilidad de expresión por sexo

Esta parte del análisis estadistico solo se ha completado en humanos, no ha sido posible realizar el estudio de diferencias en sexo en ratón porque no disponemos de suficientes muestras de hembras. Empleamos las matrices de expresión de las muestras de tejido adiposo ya anotadas, y empleamos las mismas funciones ya implementadas para realizar el cálculo de los tres estadísticos. Repetimos el procesamiento ya descrito, pero esta vez, calculamos los coeficientes de la variabilidad diferenciando entre las muestras pertenecientes a cada sexo, de modo que para los estudios que consideran ambos sexos, un mismo gen tiene dos medidas de variabilidad con cada estadistico, el correspondiente a las muestras de hombres y el correspondiente a las muestras de mujeres. Empleamos los identificadores GSE de los estudios para leer sus matrices de expresión y los identificadores GSM obtenidos en el punto anterior para extraer de la matriz las muestras correspondientes a cada sexo, seguidamente estimamos la variabilidad de cada gen en cada condición con los tres estadísticos. En el caso de estudios que solo examinan un sexo, empleamos todas sus muestras para el cálculo de la variabilidad.

Los valores obtenidos de cada estadístico, sexo y estudio, para todos los genes de una plataforma, se incorporan en una matriz, donde las filas se corresponden a los genes interrogados por la plataforma y las columnas a los estudios. Para construir esta matriz se ha empleado la misma constante, el número máximo de genes que somos capaces de identificar en una plataforma, y en estudios incompletos se han introducido valores nulos para los genes ausentes. Generamos seis matrices por plataforma que recogen los valores calculados de los tres estimadores de la variabilidad para todos los genes en las muestras de hombres y mujeres de forma independiente.

3.3.3. Ranking de variabilidad de expresión de los genes por plataforma

Una vez obtenidos los valores descriptivos del nivel de variabilidad de la expresión génica de todos los genes de las plataformas consideradas, generamos un ranking en base a los valores medianos de variabilidad de cada estadístico y plataforma. Ordenamos de menor a mayor los niveles de variabilidad de la expresión génica, donde la posición de los genes indica su estabilidad de forma relativa (script 7). Los genes que ocupan los primeros puestos en el ranking son aquellos con mayor estabilidad de expresión.

Este ranking nos permite inferir la estabilidad relativa de los seis genes HK y compararla con la variabilidad del resto de genes de una misma plataforma, conociendo su posición global podemos responder a la pregunta de cuál de los genes HK interrogados presentan una expresión más estable o son más variables en la plataforma considerada. Por otro lado, podemos identificar otros genes con expresión estable para proponerlos como genes candidatos a genes de referencia.

3.3.4. Ranking "naive"

Dado que las plataformas presentan muchos genes comunes, para cada especie, se ha desarrollado una aproximación "naive" que nos permita integrar la información obtenida con las plataformas individualmente y generar un ranking de variabilidad de expresión génica.

Cargamos en el entorno de trabajo las matrices que recogen los valores de los estadísticos de variabilidad de expresión de todos los genes en todos los estudios de cada plataforma, extraemos los identificadores Symbol y con la función unique eliminamos los duplicados. Para cada gen de la lista, recorremos las matrices, recogemos su valor del estimador de la variabilidad en los estudios en los que aparezca y resumimos la variación de su expresión mediante la mediana. Consideramos todos los estudios disponibles para cada gen, aunque haya una descompensación del número de estudios. Por último, ordenamos los valores medianos de menor a mayor y generamos un ranking "naive" (script 8). Esto se ha realizado tanto para humanos como para ratones.

De forma esquemática, para cada estadístico: CV, IQR/median y MAD/median, interrogamos cada gen de la lista de identificadores symbol para extraer sus valores del estadístico en cada plataforma (script 7):

1. Extraemos sus valores de los estudios de cada plataforma.
2. Unimos los resultados en un único vector de todos los estudios en los que aparece.
3. Obtenemos el valor correspondiente a la mediana y anotamos el número de estudios que lo contienen.
4. Almacenamos el valor mediano en un vector de medianas.
5. Construimos el resultado, una tabla con 3 columnas: los identificadores Symbol, los valores medianos y el número de estudios de cada gen.
6. Generamos un ranking ordenado por este valor mediano.

Para las diferencias en sexo se han seguido los pasos ya descritos, integramos los valores de variabilidad obtenidos para todos los genes en las muestras de hombres y mujeres, de manera independiente. Se ha generado el ranking de variabilidad por plataforma y también se ha integrado la información de las plataformas en un único ranking "naive".

3.4. Metaanálisis de los resultados

Para integrar todos los resultados generados de las diferentes plataformas hemos realizado un metanálisis con el método de Rank Product. El Rank Product (RP) es un estadístico no-paramétrico que se emplea principalmente para identificar elementos que ocupan posiciones altas en listas de rankings de forma consistente. Encuentra su aplicación directa en la detección de genes con expresión diferencial en estudios transcriptómicos de microarrays con réplicas biológicas [60], no obstante puede utilizarse para otros fines, lo encontramos implementado en R en el paquete RankProd [61, 62] de Bioconductor. El cálculo del RP es equivalente a calcular la media geométrica de las posiciones que ocupa un elemento en los diferentes rankings.

$$RP_i = \left(\prod_{j=1}^{K} rank_{i,j} \right)^{1/K} \tag{3.4}$$

La explicación detallada del algoritmo se encuentra en Mitchell, L 2001 [63].

En nuestro caso, lo aplicamos para obtener un único ranking por estadística y especie, empleamos los valores medianos de variabilidad obtenidos de cada gen donde cada plataforma es una réplica. Adaptamos nuestros datos al formato de entrada requerido por la función RankProd (tabla 3.4).

Tabla 3.4: Formato de los datos de entrada de la función RP. Extracto del formato de los datos de entrada de la función RankProd para el CV de *Homo sapiens*. Recoge el listado completo de los genes que somos capaces de identificar y sus valores medianos del CV en todas las plataformas. Las filas se corresponden a los identificadores de los genes y las columnas a los valores medianos obtenidos con cada plataforma.

gene	medianGPL570	medianGPL6244	medianGPL10558	medianGPL6947
BRCA1	0.0932	0.0884	0.0509	0.1203
BRCA2	0.1338	0.0787	0.0382	0.2135
BRCC3	0.0629	0.0536	0.0458	0.1460
BRD1	0.0541	0.0267	0.0286	0.1560
BRD2	0.0744	0.0232	0.0408	0.1478

Se ha ejecutado la función RP con los valores resultantes del cómputo con los tres estadísticos descriptivos de la variabilidad para todas las condiciones: todas las muestras de *Homo sapiens*, las muestras de hombres, las muestras de mujeres y todas las muestras de *Mus musculus*. Finalmente, se ha generado un meta-ranking ordenando de menor a mayor los valores RP obtenidos (script 9).

3.5. Identificación de nuevos genes candidatos a genes de referencia

Como ya se ha mencionado, la estabilidad de expresión es la principal característica de los genes empleados como referencia para normalizar los valores de los genes de interés en las técnicas cuantitativas de la abundancia de transcritos. Por tanto, cualquier gen con expresión estable puede ser potencialmente utilizado como gen de referencia para la normalización, siempre y cuando esta no se vea afectada por las condiciones particulares del estudio.

El flujo de trabajo que se ha llevado a cabo nos permite identificar aquellos genes que presentan menor varibilidad en sus niveles de expresión en cada condición estudiada. Por ello, en último lugar, a partir de los resultados obtenidos en el metaanálisis, se han extraido los 10 genes con expresión más estable en humanos y ratones para proponerlos como nuevos genes candidatos para la normalización de datos cuantitativos del transcriptoma. Se han excluido las localizaciones genómicas y los identificadores correspondientes a pseudogenes y micro-ARNs, según el NCBI y la base de datos GeneCards [64].

4. Resultados

En este apartado se muestran los principales resultados obtenidos en este trabajo, siguiendo el orden establecido en el apartado de Material y Métodos. Dado el importante volumen de información procesada, todos los ficheros generados y los resultados completos estan accesibles en los correspondientes directorios del repositorio de github `https://github.com/mguaita/Eval_HKG`. La información y las tablas complementarias de cada apartado se muestran en los Anexos indicados.

4.1. Revisión sistemática y selección de los estudios

4.1.1. Identificación de los estudios disponibles en GEO por tejido y organismo

Definido el origen de la muestra, el tipo de estudio y el organismo de interés con las palabras clave en el buscador web de GEO se identificaron un total de 187 estudios candidatos en *Homo sapiens* y 214 estudios candidatos de *Mus musculus*.

Hasta 109 estudios con humanos y 111 con ratones fueron excluidos por no emplear las plataformas de microarrays de uso más común o contener otros tipos de información biológica analizadas con microarrays.

De los 78 estudios elegibles en humanos y 103 en ratones, se descartaron 29 y 90, respectivamente, por no superar alguno de los criterios de inclusión establecidos: no presentar muestras de tejido adiposo analizadas con las plataformas seleccionadas, contener registros muestrales duplicados (superseries), tener menos de diez muestras de tejido adiposo o presentar sus resultados de forma no estandarizada.

Finalmente, 49 estudios de *Homo sapiens* y 43 estudios de *Mus musculus* superaron el proceso de selección y han sido incluidos en el análisis.

En la figura 4.1 recoge el resumen de los resultados obtenidos en este apartado.

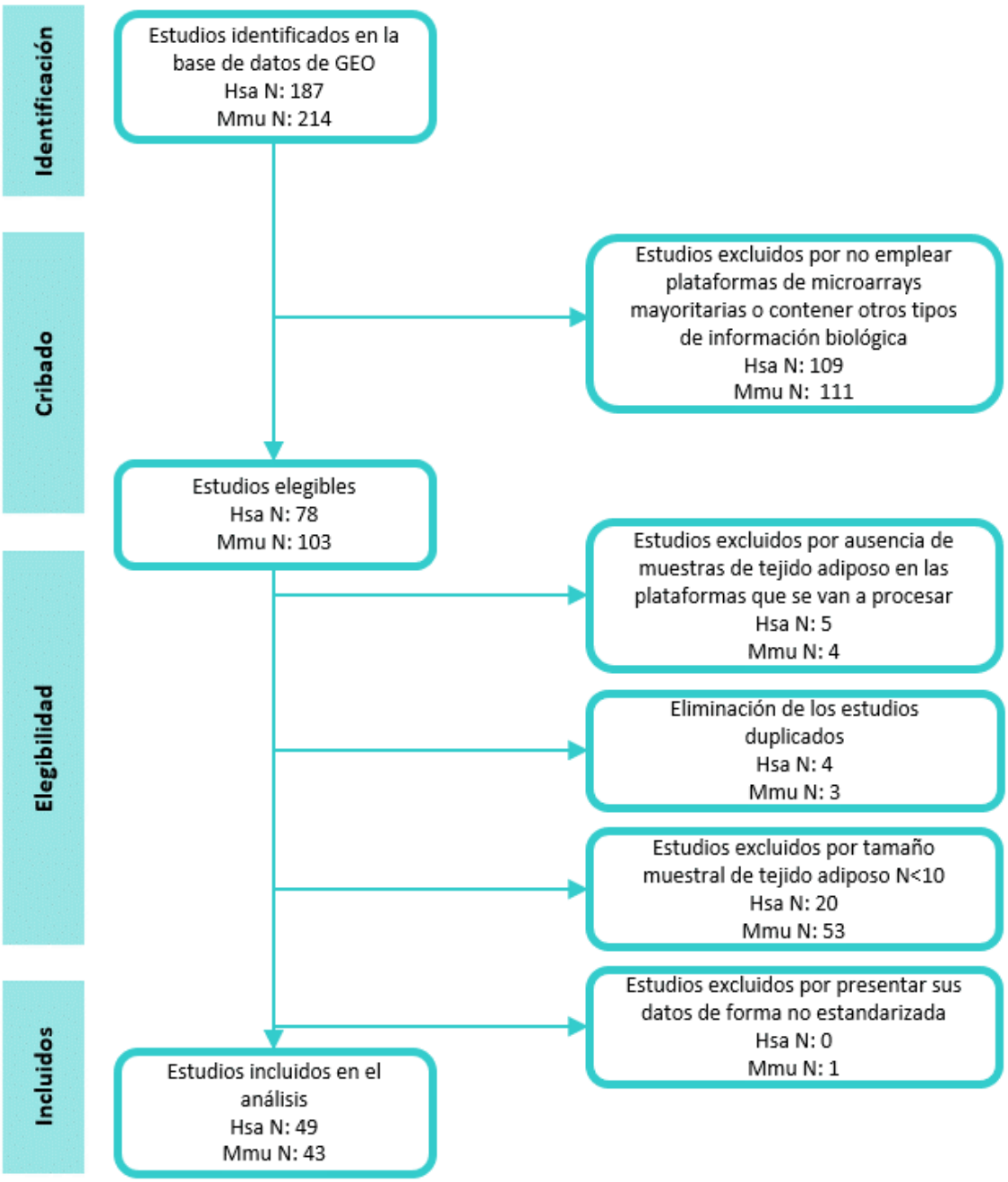

Figura 4.1: Resultado de la revisión sistemática. Desglose de la revisión sistemática y la selección de estudios realizada, siguiendo la declaración PRISMA.

4.1.2. Selección y descripción de principales plataformas

Estructurada la información contenida en el fichero HTML de los registros obtenidos de la búsqueda de GEO, se identificaron un total de 101 plataformas de *Homo sapiens* y 80 de *Mus musculus*, que incluían tanto plataformas de microarrays como de otro tipo de tecnologías. Se descartaron todas aquellas que no pertenecieran a microarrays y se seleccionaron las que contenían el mayor número de estudios, 4 para humanos y 5 para ratones. La información de las plataformas mayoritarias de *Homo sapiens* y *Mus musculus* se recogen en las tablas 4.1 y 4.2, respectivamente. Para cada plataforma se incluye el número de estudios elegibles, el número de estudios incluidos finalmente en el análisis, el número de sondas que contienen y el número máximo de genes que somos capaces de anotar.

Tabla 4.1: Principales plataformas en *Homo sapiens*. Resumen de las plataformas seleccionadas para el análisis en *Homo sapiens*.

Plataforma	Descripción	Estudios elegibles	Estudios incluidos	Sondas	Genes identificados
GPL570	Affymetrix Human Genome U133 Plus 2.0 Array	37	20	54675	22881
GPL6244	Affymetrix Human Gene 1.0 ST Array transcript (gene) version	15	13	33297	23307
GPL10558	Illumina HumanHT-12 V4.0 expression BeadChip	14	7	47323	31426
GPL6947	Illumina HumanHT-12 V3.0 expression BeadChip	12	9	48803	25159

Tabla 4.2: Principales plataformas en *Mus musculus*. Resumen de las plataformas seleccionadas para el análisis en *Mus musculus*.

Plataforma	Descripción	Estudios elegibles	Estudios incluidos	Sondas	Genes identificados
GPL1261	Affymetrix Mouse Genome 430 2.0 Array	34	16	45101	21496
GPL6246	Affymetrix Mouse Gene 1.0 ST Array transcript (gene) version	24	6	35557	24213
GPL6887	Illumina MouseWG-6 v2.0 expression BeadChip	20	8	45281	30886
GPL6885	Illumina MouseRef-8 v2.0 expression BeadChip	15	8	25697	18120
GPL16570	Affymetrix Mouse Gene 2.0 ST Array transcript (gene) version	10	5	41801	24647

4.1.3. Información muestral de los estudios por organismo

A partir de los identificadores GSE se pudo acceder con GEOmetadb a las características de los estudios candidatos, identificar aquellos estudios que contenían muestras de varios tejidos y obtener los identificadores GSM de las muestras correspondientes a tejido adiposo de las plataformas de uso más común. Los resultados obtenidos de todos los estudios elegibles de cada plataforma se pueden ver en los ficheros con nombre *summary_GPLID* en el repositorio de github.

Con esta información se ha determinado qué estudios se incluyen finalmente en el análisis. En el Anexo se desglosa la información (identificadores, plataforma, nº de muestras de tejido adiposo, indicadorMT y identificadores de las muestras de tejido adiposo) de los estudios seleccionados para ambas especies. La estructura de la información es relevante para llevar a cabo el análisis estadístico de forma automatizada.

4.1.4. Anotación del sexo de las muestras

Homo sapiens

En la revisión sistemática de los estudios *Homo sapiens*, de los 49 estudios seleccionados inicialmente, se han identificado 24 estudios donde se informa del sexo de sus muestras (tabla 4.3). De estos 24 estudios (49 % del total), 10 incluyen ambos sexos en el análisis (19 %), 11 incluyen únicamente muestras de mujeres (25 %) y solo 3 incorporan únicamente muestras de hombres (6 %) (figura 4.2). En total se han identificado 681 muestras de hombres y 875 muestras de mujeres, detallando esta información por plataforma en la tabla 4.3.

Tabla 4.3: Estudios de *Homo sapiens* que incluyen la información del sexo.

GSEID	GPL	Muestras	Hombres	Mujeres
GSE27657	GPL570	18	4	14
GSE27916	GPL570	375	113	262
GSE41168	GPL570	70	0	70
GSE61302	GPL570	15	0	15
GSE66159	GPL570	38	0	38
GSE71416	GPL570	20	5	15
GSE88837	GPL570	30	0	30
GSE9624	GPL570	11	10	1
GSE25401	GPL6244	56	0	56
GSE25910	GPL6244	36	0	36
GSE33070	GPL6244	26	10	16
GSE73655	GPL6244	20	6	14
GSE41223	GPL6244	20	4	16
GSE54280	GPL6244	12	6	6
GSE73108	GPL10558	12	0	12
GSE65221	GPL10558	136	63	73
GSE119717	GPL10558	60	60	0
GSE115645	GPL10558	24	21	3
GSE43471	GPL6947	96	0	96
GSE32512	GPL6947	204	204	0
GSE29231	GPL6947	24	0	24
GSE29226	GPL6947	24	0	24
GSE27666	GPL6947	175	175	0
GSE112307	GPL6947	54	0	54

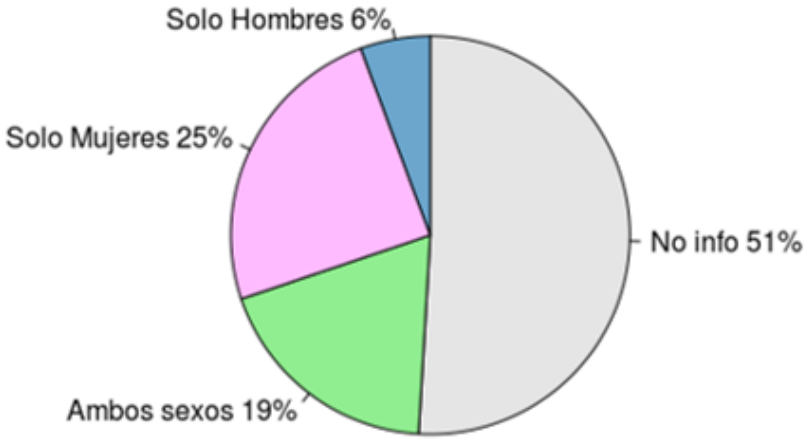

Figura 4.2: Proporción de la información del sexo en los estudios seleccionados de *Homo sapiens*.

Mus musculus

En la revisión sistemática de los estudios en *Mus musculus*, de 43 estudios iniciales, se ha encontrado un total de 22 que informan del sexo de las muestras (tabla 4.4). De estos 22 estudios (51 % del total), 1 incluyen ambos sexos en el análisis (2 %), 2 incluyen únicamente muestras de mujeres (5 %) y 19 sólo muestras de hombres (44 %) (figura 4.3). Se han identificado 559 muestras de machos y 34 muestras de hembras, en consecuencia, con un único estudio que incluya información de ambos sexos en sus muestras, no es posible realizar un análisis estadístico que permita la comparación de la variabilidad en los niveles de expresión por sexo.

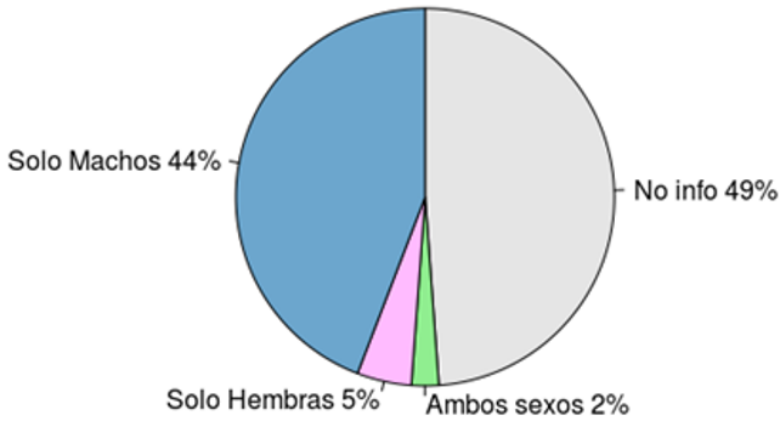

Figura 4.3: Propoción de la información del sexo en los estudios seleccionados de *Mus musculus*.

Tabla 4.4: Estudios de *Mus musculus* que incluyen la información del sexo.

GSEID	GPL	Muestras	Machos	Hembras
GSE117352	GPL1261	10	10	0
GSE140953	GPL1261	12	6	6
GSE110531	GPL1261	23	23	0
GSE66132	GPL1261	16	16	0
GSE77943	GPL1261	21	21	0
GSE97240	GPL1261	34	34	0
GSE71367	GPL1261	27	27	0
GSE67389	GPL1261	16	16	0
GSE51080	GPL1261	18	0	18
GSE13432	GPL1261	12	12	0
GSE38321	GPL6246	10	0	10
GSE79434	GPL6246	24	24	0
GSE55272	GPL6246	12	12	0
GSE37514	GPL6246	15	15	0
GSE113808	GPL6885	16	16	0
GSE70857	GPL6885	48	48	0
GSE57659	GPL6885	179	179	0
GSE97145	GPL6887	11	11	0
GSE62612	GPL6887	11	11	0
GSE50647	GPL6887	42	42	0
GSE87661	GPL16570	24	24	0
GSE79711	GPL16570	12	12	0

4.1.5. Conflictos de anotación de los genes housekeeping

Inicialmente, el gen 18S solo se localizaba en dos de las cuatro plataformas seleccionadas en humanos y en dos de las cinco consideradas para ratón. Después de la revisión se encontró que:

- En las plataformas Human Genome U133 Plus 2.0 Array y Mouse Genome 430 2.0 Array de Affymetrix, la expresión de este gen estaba siendo interrogada por sondas control positivas cuya información no estaba incluida en el campo de los identificadores Symbol de los ficheros de anotación de GEO ni en las librerías de Bioconductor, y como consecuencia estaba siendo eliminado en el tratamiento de valores nulos.

- En las plataformas HumanHT-12 V4.0 expression BeadChip y HumanHT-12 V3.0 expression BeadChip de Illumina, la primera presenta tres sondas ILMN_1703337, ILMN_3239610 y ILMN_3243593, y la segunda, solo la sonda ILMN_1703337, que están anotadas, tanto en GEO como en Bioconductor, como complementarias al gen 18S. Se han contrastado las localizaciones genómicas de las anotaciones con el NCBI para confirmar si son complementarias al gen 18S. Las localizaciones genómicas retiradas del NCBI se han retirado del análisis.

El resto de genes HK estaban anotados correctamente con el mismo identificador Symbol en las plataformas en humano. Sin embargo, en ratón, Hprt y Gapdh se han encontrado anotados con sinónimos y alias, Ppia, estaba anotado en segunda posición los identificadores Symbol, y Ubc no tiene sonda complementaria en la plataforma MouseWG-6 v2.0expression BeadChip de Illumina.

En el repositorio github se recogen los featureData de las plataformas combinando ambas anotaciones de GEO y Bioconductor de las plataformas problemáticas.

4.2. Análisis estadístico individual de cada estudio

4.2.1. Determinación de indicadores de variabilidad de expresión y ranking de variabilidad de expresión de los genes

Tras la descarga de los datos normalizados y la anotación de todos los estudios se realizó el cálculo de los tres estimadores de la variabilidad basados en la media y la mediana. Se han obtenido para todos los genes de las plataformas consideradas, su nivel de variabilidad en cada estudio de la plataforma, en base a los valores de expresión génica de las muestras de tejido adiposo. Se han generado un total de 12 matrices para humanos y 15 para ratón (una por cada combinación de estadístico, plataforma y organismo), que recogen el nivel de variabilidad estimado con los tres estadísticos de todos los genes identificados de cada plataforma, en base a los valores de expresión génica de las muestras de tejido adiposo.

Los niveles de variabilidad obtenidos con cada estadístico se han explorado para generar un ranking de estabilidad para cada plataforma plataforma, a partir del valor mediano de cada gen, ordenando los genes de menor a mayor variabilidad. Los rankings se pueden observar en el repositorio de github. Aquellos genes que ocupan los primeros puestos presentan los mayores niveles de estabilidad de expresión en tejido adiposo.

A continuación, se muestra el resumen de la variabilidad mediana obtenida con cada estadístico y la posición en los rankings generales de la primera plataforma considerada de cada especie, para los seis genes *housekeeping* que queremos evaluar (ver tabla 4.5) para *Homo sapiens* y tabla 4.6 para *Mus musculus*). Para mayor claridad se ha incluido la columna "rank HKG" que ordena los genes HK de mayor a menor estabilidad. Los resultados del resto de plataformas se muestran en las tablas B.9 y B.12 del Anexo B.

Tabla 4.5: Resultados obtenidos de los tres estadísticos con la plataforma GPL570 de *Homo sapiens*. Esta plataforma recoge la información de 23 estudios y podemos idetificar un total de 22881 genes. En "rank HKG" se indica el orden de estabilidad relativo entre los 6 genes HK de interés, en "general ranking" se muestra la posición que ocupa cada gen HK en el ranking global considerando todos los genes de la plataforma, la última columna muestra el valor mediano de la variabilidad estimada del gen HK, en el conjunto de estudios de la plataforma, con cada estadístico.

rank HKG	gene	general ranking	median CV
1	PPIA	47	0,0257
2	RPL19	57	0,0272
3	UBC	92	0,0302
4	GAPDH	1033	0,0457
5	HPRT1	6837	0,0778
6	RNA18S	7336	0,0801

rank HKG	gene	general ranking	median IQR/m
1	UBC	77	0,0222
2	RPL19	81	0,0224
3	PPIA	106	0,0234
4	GAPDH	1187	0,0369
5	HPRT1	9411	0,0719
6	RNA18S	10591	0,0762

rank HKG	gene	general ranking	median MAD/m
1	PPIA	36	0,0270
2	UBC	75	0,0323
3	RPL19	82	0,0329
4	GAPDH	1069	0,0533
5	RNA18S	10750	0,1134
6	HPRT1	12586	0,1233

Tabla 4.6: Resultados obtenidos con la plataforma GPL1261 de *Mus musculus*. Esta plataforma recoge la información de 16 estudios y podemos idetificar un total de 21496 genes. En "rank HKG" se indica el orden de estabilidad relativo entre los 6 genes HK de interés, en "general ranking" se muestra la posición que ocupa cada gen HK en el ranking global considerando todos los genes de la plataforma, la última columna muestra el valor mediano de la variabilidad estimada del gen HK, en el conjunto de estudios de la plataforma, con cada estadístico.

rank HKG	gene	general ranking	median CV
1	Ppia	8	0,0269
2	Ubc	65	0,0354
3	Rpl19	1462	0,0519
4	Hprt	4842	0,0643
5	Gapdh	15569	0,0992
6	Rn18s	21399	0,2711

rank HKG	gene	general ranking	median IQR/m
1	Ppia	16	0,0235
2	Ubc	703	0,0387
3	Hprt	3672	0,0510
4	Rpl19	4313	0,0528
5	Gapdh	9573	0,0670
6	Rn18s	21077	0,1841

rank HKG	gene	general ranking	median MAD/m
1	Ppia	5	0,0264
2	Ubc	83	0,0435
3	Rpl19	2756	0,0720
4	Hprt	7977	0,0935
5	Gapdh	10502	0,1040
6	Rn18s	19862	0,1901

4.2.2. Determinación de indicadores de variabilidad de expresión y ranking de variabilidad de expresión de los genes por sexo

Este apartado se ha completado únicamente con *Homo sapiens*. Para cada estudio seleccionado se han extraído las muestras correspondientes a hombres y mujeres de las matrices de expresión normalizadas previamente anotadas y se han calculado los niveles de variabilidad de todos los genes con los tres estimadores para ambos sexos de forma independiente.

De nuevo, se han explorado los resultados generales para extraer de todos los genes sus niveles de variabilidad mediana estimada y generar un ranking de estabilidad, ordenando de menor a mayor las medidas de variabilidad, para cada plataforma haciendo distinción

por sexos. Los rankings de las cuatro plataformas se pueden explorar en el repositorio de github.

Ponemos especial interés en identificar, si existen, posibles diferencias en los niveles de variabilidad que presentan nuestros seis genes *housekeeping* asociadas con el sexo, recogemos los valores medianos de variabilidad de estos genes y la posición que ocupan en el ranking general para cada plataforma y estadístico. A continuación mostramos los resultados obtenidos con la plataforma GPL570 en Hombres (tabla 4.7) y Mujeres (tabla 4.8), recordamos que hemos podido identificar 22881 genes en esta plataforma. El conjunto de los resultados se detallan en las tablas B.11 y B.10 del Anexo B.

Tabla 4.7: Resultados obtenidos con la plataforma GPL570 para los indicadores de la variabilidad en muestras de hombres. En "rank HKG" se indica el orden de estabilidad relativo entre los 6 genes HK de interés, en "general ranking" se muestra la posición que ocupa cada gen HK en el ranking global considerando todos los genes de la plataforma, la última columna muestra el valor mediano de la variabilidad estimada del gen HK, en el conjunto de estudios de la plataforma, con cada estadístico.

rank HKG	gene	general ranking	median CV
1	UBC	39	0,0184
2	RPL19	368	0,0288
3	PPIA	502	0,0310
4	GAPDH	3188	0,0540
5	RNA18S	9819	0,0897
6	HPRT1	15211	0,1209

rank HKG	gene	general ranking	median IQR/m
1	UBC	116	0,0156
2	RPL19	777	0,0246
3	PPIA	802	0,0249
4	GAPDH	2099	0,0336
5	HPRT1	15922	0,0994
6	RNA18S	16740	0,1050

rank HKG	gene	general ranking	median MAD/m
1	UBC	213	0,0238
2	RPL19	853	0,0354
3	PPIA	1314	0,0407
4	GAPDH	2408	0,0511
5	HPRT1	11911	0,1172
6	RNA18S	15895	0,1524

Tabla 4.8: Resultados obtenidos con la plataforma GPL570 para los indicadores de la variabilidad en mujeres. En "rank HKG" se indica el orden de estabilidad relativo entre los 6 genes HK de interés, en "general ranking" se muestra la posición que ocupa cada gen HK en el ranking global considerando todos los genes de la plataforma, la última columna muestra el valor mediano de la variabilidad estimada del gen HK, en el conjunto de estudios de la plataforma, con cada estadístico.

rank HKG	gene	general ranking	median CV
1	RPL19	15	0,0315
2	PPIA	137	0,0502
3	UBC	615	0,0646
4	GAPDH	1877	0,0822
5	HPRT1	7730	0,1234
6	RNA18S	22756	0,4176

rank HKG	gene	general ranking	median IQR/m
1	UBC	60	0,0206
2	PPIA	69	0,0217
3	RPL19	94	0,0238
4	GAPDH	1257	0,0479
5	HPRT1	7805	0,0987
6	RNA18S	22869	0,6204

rank HKG	gene	general ranking	median MAD/m
1	UBC	51	0,0302
2	RPL19	72	0,0332
3	PPIA	79	0,0338
4	GAPDH	1101	0,0688
5	HPRT1	8404	0,1533
6	RNA18S	22858	0,6382

4.2.3. Ranking "naive"

La aproximación "naive" del metaanálisis propuesta permite agregar los resultados obtenidos de las plataformas de cada especie. Con cada estadístico, se han recogido todas las estimaciones de la variabilidad disponibles de todos los genes identificados.

Para *Homo sapiens* se han explorado tres condiciones: todas las muestras, solo muestras de hombres y solo muestras de mujeres. Se han identificado 41975 genes únicos (41973 en hombres).

Para *Mus musculus*, ante la imposibilidad de interrogar diferencias de expresión por sexos, se han explorado todas las muestras sin hacer distinción entre machos y hembras. Se han identificado un total de 47203 genes únicos.

Para cada condición, se ha generado un ranking global de la variabilidad mediana estimada con cada estadístico. Asimismo, para cada gen se ha apuntado el número de estudios (GSEs) que lo interrogan. Estos rankings se pueden observar en el repositorio de github. A continuación, se muestra el resumen de la variabilidad mediana obtenida de cada estadístico con todos los valores disponibles para los seis genes HK, así como su posición global y el número de estudios que interrogan su expresión (ver tablas 4.9, 4.10, 4.11 y 4.12)

Tabla 4.9: Resultados "naive" ranking para el conjunto de muestras *Homo sapiens*. Para cada estadístico, en "ranking" se muestra la posición que ocupa el gen HK en el ranking global considerando todos los genes de todas las plataformas, la tercera columna muestra el valor mediano de la variabilidad estimada del gen HK, en el conjunto de estudios que lo contienen, y la última columna informa del número de estudios que contienen ese gen.

gene	ranking	median CV	Nº GSEs
RPL19	699	0,0312	49
UBC	1024	0,034	48
PPIA	2554	0,0394	48
GAPDH	6973	0,0535	48
HPRT1	15234	0,0741	48
RNA18S5	17771	0,0794	35

gene	ranking	median IQR/m	Nº GSEs
RPL19	289	0,0256	49
UBC	311	0,0259	48
PPIA	790	0,0298	48
GAPDH	3467	0,0387	48
RNA18S5	20887	0,0727	35
HPRT1	22907	0,0759	48

gene	ranking	median MAD/m	Nº GSEs
RPL19	274	0,0376	49
UBC	376	0,0391	48
PPIA	778	0,0438	48
GAPDH	2387	0,0534	48
RNA18S5	20902	0,1074	35
HPRT1	27513	0,1231	48

Tabla 4.10: Resultados "naive" ranking para las muestras de hombres. Para cada estadístico, en "ranking" se muestra la posición que ocupa el gen HK en el ranking global considerando todos los genes de todas las plataformas, la tercera columna muestra el valor mediano de la variabilidad estimada del gen HK, en el conjunto de estudios que lo contienen, y la última columna informa del número de estudios que contienen ese gen.

gene	ranking	median CV	Nº GSEs
UBC	118	0,0184	13
RPL19	1444	0,0297	13
GAPDH	1949	0,0317	13
PPIA	4349	0,0393	13
HPRT1	9661	0,0544	13
RNA18S5	18445	0,0786	10

gene	ranking	median IQR/m	Nº GSEs
UBC	560	0,0188	13
RPL19	1694	0,025	13
GAPDH	3822	0,0315	13
PPIA	3985	0,0319	13
HPRT1	11005	0,0484	13
RNA18S5	15176	0,058	10

gene	ranking	median MAD/m	Nº GSEs
UBC	519	0,0272	13
RPL19	1492	0,0355	13
PPIA	3466	0,0448	13
GAPDH	4884	0,0506	13
HPRT1	12110	0,0752	13
RNA18S5	15346	0,0865	10

Tabla 4.11: Resultados "naive" ranking para las muestras de mujeres. Para cada estadístico, en "ranking" se muestra la posición que ocupa el gen HK en el ranking global considerando todos los genes de todas las plataformas, la tercera columna muestra el valor mediano de la variabilidad estimada del gen HK, en el conjunto de estudios que lo contienen, y la última columna informa del número de estudios que contienen ese gen.

gene	ranking	median CV	Nº GSEs
RPL19	5520	0,0849	20
RNA18S5	8077	0,0986	13
UBC	12048	0,1153	20
HPRT1	19760	0,1366	20
PPIA	20215	0,1378	20
GAPDH	33453	0,1771	19

gene	ranking	median IQR/m	Nº GSEs
RPL19	3935	0,0607	20
UBC	14303	0,1002	20
HPRT1	15918	0,1044	20
RNA18S5	16980	0,1071	13
GAPDH	23057	0,1218	19
PPIA	26812	0,1312	20

gene	ranking	median MAD/m	Nº GSEs
RPL19	3831	0,0819	20
UBC	15557	0,1501	20
HPRT1	15817	0,1512	20
RNA18S5	19334	0,1652	13
PPIA	24378	0,1842	20
GAPDH	24400	0,1843	19

Tabla 4.12: Resultados "naive" ranking para el conjunto de muestras de *Mus musculus*. Para cada estadístico, en "ranking" se muestra la posición que ocupa el gen HK en el ranking global considerando todos los genes de todas las plataformas, la tercera columna muestra el valor mediano de la variabilidad estimada del gen HK, en el conjunto de estudios que lo contienen, y la última columna informa del número de estudios que contienen ese gen.

gene	ranking	median CV	nGSEs
Ppia	2710	0,0137	41
Ubc	16356	0,0290	34
Rpl19	23416	0,0353	41
Gapdh	26153	0,0379	42
Rn18s	29041	0,0410	27
Hprt	35624	0,0505	43

gene	ranking	median IQR/m	nGSEs
Ppia	475	0,0103	41
Ubc	12150	0,0226	34
Rpl19	22638	0,0311	41
Rn18s	23516	0,0317	27
Gapdh	30739	0,0383	42
Hprt	37488	0,0485	43

gene	ranking	median MAD/m	nGSEs
Ppia	205	0,0141	41
Ubc	14128	0,0364	34
Rpl19	21336	0,0448	41
Rn18s	22988	0,0467	27
Gapdh	30278	0,0563	42
Hprt	38164	0,0737	43

4.3. Metaanálisis de resultados

Tras la ejecución de la función RP se ha obtenido un MetaRanking único para cada condición de estudio. Los metaRankings completos de cada condición se pueden explorar en el repositorio de github.

4.3.1. Resultados globales del conjunto de todos los estudios por organismo

A continuación mostramos el extracto de las posiciones que ocupan los genes HK en el metaRanking y el cómputo del estadístico Rank Product según el estimador de la variabilidad empleado (ver tabla 3.4).

Tabla 4.13: Resultados metaanálisis. A. Valores obtenidos con *Homo sapiens* (41975 genes). B. Valores obtenidos con *Mus musculus* (47203 genes).

A) Homo sapiens

CV		
Gene	**metaRanking**	**RP**
UBC	460	1048
PPIA	555	1231
RPL19	737	1633
GAPDH	5055	6770
RNA18S	19506	17687
HPRT1	25528	21849

IQR/median		
Gene	**metaRanking**	**RP**
UBC	233	648.5
PPIA	600	1432
RPL19	1301	2630
GAPDH	5529	7312
RNA18S	18121	16716
HPRT1	26215	22168

MAD/median		
Gene	**metaRanking**	**RP**
UBC	271	712.4
PPIA	494	1225
RPL19	943	2064
GAPDH	5771	7537
RNA18S	18782	17238
HPRT1	21574	19123

B) Mus musculus

CV		
Gene	**metaRanking**	**RP**
Ppia	643	1013
Rn18s	5044	6214
Rpl19	6962	8319
Ubc	10243	11752
Gapdh	13359	14545
Hprt	26657	23664

IQR/median		
Gene	**metaRanking**	**RP**
Ppia	551	891
Rn18s	6600	8075
Rpl19	10597	12008
Ubc	11816	13111
Gapdh	21618	20260
Hprt	24711	22280

MAD/median		
Gene	**metaRanking**	**RP**
Ppia	689	1058
Rn18s	4446	5658
Rpl19	11718	13110
Ubc	13263	14416
Gapdh	21154	20020
Hprt	25565	22876

4.3.2. Resultados globales del conjunto de estudios por sexo

Mostramos el extracto de las posiciones que ocupan los genes HK en el metaRanking y el cómputo del estadístico Rank Product según el estimador de la variabilidad empleado, con las muestras de Hombres y Mujeres (ver tabla 4.14).

Tabla 4.14: Resultados metaanálisis basado en sexos. A. Valores obtenidos con hombres (41973 genes). B. Valores obtenidos con Mujeres (41975 genes).

A) Hombres

CV		
Gene	**metaRanking**	**RP**
UBC	667	1237
PPIA	1458	2473
GAPDH	2495	3848
RPL19	3785	5291
RNA18S	11158	11499
HPRT1	13278	13067

IQR/median		
Gene	**metaRanking**	**RP**
UBC	1394	2384
GAPDH	1607	2709
PPIA	3057	4610
RPL19	3938	5597
HPRT1	8838	9863
RNA18S	10051	10780

MAD/median		
Gene	**metaRanking**	**RP**
UBC	1845	3062
GAPDH	1883	3108
RPL19	3213	4809
PPIA	4270	5897
HPRT1	7236	8598
RNA18S	11133	11588

B) Mujeres

CV		
Gene	**metaRanking**	**RP**
RPL19	4159	5188
UBC	8634	10700
HPRT1	9681	11648
PPIA	10703	12501
GAPDH	17827	17458
RNA18S	22209	20320

IQR/median		
Gene	**metaRanking**	**RP**
RPL19	3524	4679
UBC	3758	5004
PPIA	5657	7430
GAPDH	9816	11515
HPRT1	14020	14670
RNA18S	20029	18707

MAD/median		
Gene	**metaRanking**	**RP**
RPL19	991	1348
UBC	4174	5527
PPIA	7869	9585
GAPDH	11916	13002
HPRT1	18590	17667
RNA18S	19217	18093

4.4. Identificación de genes candidatos a genes de referencia

En la tabla 4.15 se recogen los 10 genes con mayor estabilidad de expresión en *Homo sapiens* y *Mus musculus* en base a los resultados del metaanálisis. De cada gen se detalla la posición ocupada en el metaRanking y el valor estimado del Rank Product.

La tabla 4.16 muestra los 10 genes con mayor estabilidad de expresión en hombres y mujeres.

Tabla 4.15: Top 10 de genes con expresión más estables en *Homo sapiens* y *Mus musculus*.

A) Homo sapiens

CV		
Gene	**metaRanking**	**RP**
CYTB	6	12
COX2	22	45.85
ND3	27	53.23
ND5	33	74.48
KLRAQ1	37	89.76
ND4	42	100.4
DDX39B	45	104.7
COX1	51	116.9
PPIAL4A	69	171.1
NELFA	86	226.3

IQR/median		
Gene	**metaRanking**	**RP**
RNA28S5	1	1.316
CYTB	4	4.865
COX1	6	10.25
ND3	7	12.34
COX2	8	13.5
ND5	21	34.07
ND4	23	37.58
MYL12B	26	41.71
KLRAQ1	45	91.34
EEF1A1	48	105.8

MAD/median		
Gene	**metaRanking**	**RP**
RNA28S5	1	1.189
CYTB	4	5.03
COX1	10	17.66
ND3	11	18.59
COX2	15	23.17
ND4	23	40.59
ND5	26	45.77
MYL12B	34	62.84
KLRAQ1	35	66.26
POTEF	41	78.08

B) Mus musculus

CV		
Gene	**metaRanking**	**RP**
Rn28s1	1	2.178
mt-Rnr1	2	4.282
CYTB	3	5.615
Rn7sk	4	7.34
mt-Rnr2	5	10.93
ND1	6	14.92
ND2	7	15.23
ATP6	10	21.26
ND4	11	23.4
COX1	12	26.38

IQR/median		
Gene	**metaRanking**	**RP**
Rn28s1	1	1.149
Rn7sk	2	5.498
mt-Rnr1	3	8.445
CYTB	4	10.89
G0B-alpha	9	21.44
ND1	11	25.36
mt-Rnr2	14	32.92
COX1	18	41.4
ATP6	19	42.97
GPR106	20	44.66

MAD/median		
Gene	**metaRanking**	**RP**
Rn28s1	1	1.149
Rn7sk	2	5.55
CYTB	4	10.19
mt-Rnr1	5	11.82
COX1	8	18.16
mt-Rnr2	9	19.1
Rdh20	11	21.99
ND1	16	35.42
G0B-alpha	22	48.39
Murr2	24	51.51

Tabla 4.16: Top 10 de genes con expresión más estables en hombres y mujeres.

A) Hombres

CV		
Gene	**metaRanking**	**RP**
CYTB	1	1.189
ND5	3	4.243
ND4	29	45.34
TMPPE	30	46.84
ND3	31	48.24
COX1	34	53.87
ND1	35	54.31
NELFA	36	56.16
COX2	54	94.7
PRR27	61	110.1

IQR/median		
Gene	**metaRanking**	**RP**
SRSF2	7	10.14
CYTB	10	14.42
ND5	26	38.32
ND3	27	39.98
TMPPE	38	62.65
RNA28S5	44	74.69
COX1	46	78.22
PRR23B	51	86.1
IGLV9-49	55	92.07
ND1	57	95.74

MAD/median		
Gene	**metaRanking**	**RP**
CYTB	17	27.09
SRSF2	20	31.69
COX1	22	33.04
TMPPE	23	34.49
ND5	41	63.33
ADTRP	62	98.02
ND3	67	110.2
OR14K1	92	158.5
RNA28S5	114	201.5
PRR27	120	211.6

B) Mujeres

CV		
Gene	**metaRanking**	**RP**
ND5	10	14.56
CYTB	20	27.12
COX2	279	327.6
ZNRF3-AS1	280	328.6
ND4	282	330
COX1	313	364.7
FUSSEL18	326	379.2
ANKRD61	364	422
IGLV9-49	732	832.9
METTL11A	971	1098

IQR/median		
Gene	**metaRanking**	**RP**
CYTB	5	7.158
BECN1L1	6	8.819
COX1	12	16.94
ND4	13	17.2
COX2	60	78.94
ND5	88	114.7
FUSSEL18	214	272.3
SNORD112	439	555.9
METTL11A	728	901.5
IGLV9-49	770	954.6

MAD/median		
Gene	**metaRanking**	**RP**
TTC34	1	1,414
CYTB	8	11.47
SNORD112	39	57.39
COX1	44	63.28
ND4	64	92.94
SNORD126	78	113.2
COX2	157	220.4
GAGE2E	188	264.6
BECN1L1	297	422.8
ND5	319	455.4

5. Discusión

La veracidad de los análisis de la expresión génica depende de la correcta normalización de los valores con genes de referencia para las condiciones de estudio. Para que un gen sea adecuado como control interno debe presentar una expresión estable y constante en todas las condiciones y tipos celulares del estudio, así como ser expresado a unos niveles suficientes para ser detectado fácilmente por la tecnología empleada.

Tradicionalmente, los genes *housekeeping* se consideraban apropiados como genes de referencia y han sido ampliamente utilizados en las diferentes técnicas de cuantificación, relativa y absoluta, del ARN. No obstante, la estabilidad de estos genes ha sido ampliamente cuestionada e interrogada. Múltiples estudios han demostrado como los niveles de expresión fluctuan en función del tipo celular, el tejido o las condiciones fisiopatológicas de las muestras, y cómo esta variación se traduce en una alteración de los resultados dependiendo del gen o conjunto de genes seleccionados para normalizar [22, 25, 27]. Las primeras aproximaciones analizan los perfiles de expresión de un número reducido de genes obtenidos con RT-PCR cuantitativa [27, 26, 28], con el avance de la tecnología y la aparición de las técnicas ómicas se comenzó a interrogar la expresión génica a gran escala con microarrays [21, 29, 30, 31, 32, 54, 65], con secuenciación masiva de transcritos (RNA-Seq) [24] y combinando datos de diferentes fuentes en un metaanálisis [34]. En conjunto, todos evidencian que la validación de los genes empleados como controles es indispensable para la fidelidad de la cuantificación de la abundancia de transcritos realizada y de los perfiles de expresión generados.

En este trabajo hemos evaluado la variabilidad de la expresión génica de seis genes *housekeeping* clásicos, cuatro metabólicos, HPRT, PPIA, UBC, GAPDH, y dos ribosomales, 18S y RPL19, en el tejido adiposo de *Homo sapiens* y *Mus musculus*, mediante el abordaje masivo de los datos de microarrays disponibles en GEO, para identificar aquellos genes que muestran mayor estabilidad en su expresión independientemente de las condiciones de la muestra. Asimismo, se ha evaluado el conjunto de todos los genes incluidos en los estudios, para la detección de nuevos y potenciales genes de referencia. Empleamos tres estimadores diferentes de variabilidad relativa e integramos los resultados en un metanálisis final. Se ha desarrollado una metodología para el filtrado de la información disponible en GEO, la identificación de forma programática de los estudios de interés y la evaluación estadística de los patrones de expresión génica de las muestras de tejido adiposo. Hemos considerado todas las muestras disponibles de cada organismo con la finalidad de determinar la estabilidad de la expresión en el conjunto del tejido adiposo. Adicionalmente, se ha realizado el análisis de la estabilidad desde la perspectiva diferencial por sexo.

5.1. Indicadores de variabilidad de expresión

Ante la gran variedad de plataformas de microarrays empleadas en los estudios con muestras de tejido adiposo, se decidió seleccionar para el análisis aquellas mayoritarias, es decir, que contuviesen el mayor número de estudios. Para cada plataforma, se ha determinado el nivel de variabilidad de cada gen, en las muestras de tejido adiposo de cada uno de los estudios, con los tres estimadores: CV, IQR/median y MAD/median. El CV ha demostrado ser de utilidad para reflejar los niveles de variabilidad de los genes HK a partir de de datos cuantitativos de la expresión génica obtenidos con microarrays [29, 31, 32]. Se trata de un estimador basado en la media, es adimensional y no posee unidades, de modo que permite comparar valores de diferentes magnitudes, sin embargo se ve considerablemente afectado por valores extremos. El IQR/median y el MAD/median, son dos estimadores alternativos más robustos, basados en la mediana, equivalentes al CV [59]. Con cada estadístico, hemos resumido el nivel de variabilidad de cada gen como la mediana de las estimaciones de todos los estudios, y hemos ordenado estos valores medianos para generar un ranking de estabilidad por plataforma.

De cada plataforma, hemos extraído los valores medianos y las posiciones que ocupan en los rankings los genes HK de interés para su evluación. Como hemos visto con los conflictos de anotación no todas las plataformas interrogan todos los genes HK. Dentro de una misma plataforma, podemos observar que los valores medianos obtenidos con cada indicador difieren, aunque muchos son muy próximos, consecuentemente, también lo hacen las posiciones que ocupan estos genes en sus respectivos rankings. Estas diferencias se deben a que cada estadístico sigue un criterio distinto para cuantificar la variabilidad relativa de los datos. Por otro lado, en la comparativa entre plataformas los resultados obtenidos tampoco coinciden, esto nos indica que el mejor gen para normalizar con una plataforma no tiene por qué ser el mismo con otra, es explicable en tanto que cada plataforma y casa comercial emplea sus métodos de medida y normalización. Estudios previos ya mencionan la existencia de esta variabilidad dependiente de la tecnología, aunque concluyen que no es tan determinante como las diferencias existentes en los niveles de expresión de los ARN individuales asocidas a diferencias en las condiciones celulares [28].

De las cuatro plataformas analizadas de *Homo sapiens*, en las plataformas GPL570 y GPL6244 de Affymetrix, los genes RPL19, UBC y PPIA se alternan como los tres más estables con niveles de variabilidad en torno al 0.03 o inferior en los tres estadísticos. Además, en el ranking global de 22881 genes de GPL570 se encuentran en las primeras 100 posiciones. En GPL6244, el gen UBC ocupa la posición 5 del ranking general de 23307 genes con los dos estimadores basados en la mediana y la 22 con el CV. Como se puede apreciar son posiciones muy altas, que se corresponden con unos niveles importantes de estabilidad de expresión. Por otro lado, GAPDH es el cuarto gen HK más estable en ambas plataformas. Por su parte, HPRT1 y RNA18S, presentan los niveles más altos de variabilidad. La plataforma GPL10558 de Illumina, también coincide en situar RPL19, UBC, PPIA y GAPDH como los más estables respecto de HPRT y RNA18S. Por último, la plataforma GPL6947 de Illumina es la más discrepante de las cuatro, aunque los genes RPL19 y UBC muestran la menor variabilidad, HPRT se posiciona con mayor estabilidad que PPIA y GAPDH.

Aunque podemos apreciar algunas tendencias, los valores medianos y las posiciones que

ocupan estos genes en los rankings por plataforma de cada estadístico no son coincidentes en valores absolutos. En conjunto, en humanos, los genes RPL19, UBC, PPIA y GAPDH muestran menores niveles de variabilidad y ocupar posiciones más altas en los rankings, en comparación con HPRT y RNA18S, que muestran valores de variabilidad considerablemente superiores y ocupan posiciones más altas en los rankings de sus respectivas plataformas.

En las cinco plataformas analizadas de *Mus musculus* obtenemos aún mayores diferencias dentro de una misma plataforma y entre plataformas. Podemos apreciar que el gen Ppia es el gen que más veces presenta la menor variabilidad relativa de todos los genes HK, ocupa las posiciones 8 (CV), 16 (IQR/median) y 5 (MAD/median) de los rankings globales de la plataforma GPL1261 de Affymetrix, y en la plataforma GPL6246, también de Affymetrix, muestra niveles de variabilidad mínimos junto al gen 18S.

5.2. Indicadores de variabilidad de expresión por sexo

En el primer análisis global se evalúan todas las muestras sin diferenciar por sexo. en este segundo hemos incorporado este dato como variable, pudiendo comparar los resultados obtenidos con el primer abordaje descrito. Hemos empleado 24 de los 49 estudios disponibles de *Homo sapiens*, los estudios restantes (51 %) no informaban del sexo de las muestras, aún así, el tamaño muestral de hombres y mujeres ha sido suficiente para realizar el análisis diferencial por sexos. No ha sido el caso de *Mus musculus*, donde si bien 21 estudios aportaban el dato del sexo de las muestras, la importante mayoría de estudios empleaban únicamente ratones macho, tan solo 3 estudios incluian muestras de hembras, ascendiendo a un total de 34 muestras de hembras respecto de las más de 500 de machos, en consecuencia no hemos podido realizar este analisis con ratones. Este escenario podía ser esperable ya que cómo se ha apuntado, existe una preferencia a escoger individuos macho frente a hembras o ambos sexos en los estudios biomédicos [53].

El análisis de la variabilidad por sexos de los genes HK con cada plataforma y estadístico genera resultados similares a los obtenidos con el conjunto de todas las muestras, observamos diferencias dentro de una misma plataforma según el estadístico empleado y diferencias entre plataformas asociadas a la tecnologías.

La plataforma más discrepante es la GPL10558, donde observamos una inversión de los resultados con el gen HPRT1, 18s y PPIA.

- Al considerar todas las muestras, HPRT1 muestra los mayores niveles de variabilidad en el análisis conjunto de todas las muestras de *Homo sapiens*, junto al 18S, ocupando puestos en torno a la posición 27000 del ranking general de 31426 genes de la plataforma en los tres estadísticos. Por el contrario, en el análisis diferencial por sexos, en hombres, muestra la menor variabilidad para los tres estadísticos de todos los genes HK ocupando posiciones altas, 792 (CV), 981 (IQR/median) y 1237 (MAD/median), en el ranking global. En mujeres también presenta menores niveles de variación respecto del resto de genes HK pero ocupa posiciones considerablemente peores, 3647 (CV), 7055 (IQR/median) y 13510 (MAD/median) en comparación hombres donde HPRT1 entra dentro de los 1000 primeros genes, aunque los valores de estabilidad calculados son similares para ambos sexos. Podemos inferir que en mujeres existen otros genes con mejor estabilidad que HPRT1.

- Por su parte, el gen 18S ocupa la segunda posición para los tres estadísticos en hombres, en cambio, en el análisis conjunto de todas las muestras ocupa posiciones cercanas a 30000 en el ranking general. Por otro lado, si observamos los valores medianos de cada estadístico en hombres vemos que son prácticamente coincidentes con los de las mujeres, sin embargo en mujeres el gen 18S ocupa posiciones más centrales del ranking, es decir, presenta mayor variabilidad relativa respecto del conjunto de genes de la plataforma, apuntando de nuevo que hay otros genes con mejor estabilidad de expresión en las mujeres. En el resto de plataformas los genes HPRT y 18S presentan la mayor variabilidad en machos.
- El gen PPIA es el gen HK que muestra la mejor estabilidad con los dos estimadores basados en la mediana en los rankings generados con todas las muestras, ocupando las posiciones 5299 en el ranking del IQR/median y 5941 en el ranking del MAD/median. Al contrario, en hombres es el gen que muestra la mayor variabilidad con los tres estadísticos ocupando las posiciones 26783 (CV), 25535 (IQR/median) y 26674 (MAD/median).

El orden de los genes HK en los rankings de la plataforma GPL570 en hombres es coincidente para los tres estadísticos aunque no compartan los valores absolutos.

En este punto, tras la revisión y la evaluación realizada, confirmamos la necesidad de incluir individuos de ambos sexos en los diseños experimentales, e informar del sexo de las muestras de forma sistemática de modo que se puedan identificar inequívocamente. Sería interesante, por ejemplo, la inclusión de este dato como un atributo más en los registros muestrales de los repositorios públicos como GEO, ya que posibilitaría la realización de estudios *in silico* como el actual, más completos e incrementando la potencia estadística. En última instancia, concede la capacidad de hacer extensible el conocimiento generado al conjunto de la población y no solo a un grupo particular.

5.3. Ranking "naive"

Como encontramos una variabilidad considerable a nivel de plataforma, realizamos una aproximación simple para integrar los resultados individuales de cada una. Para ello, determinamos el total de genes únicos que podemos identificar con las plataformas de cada especie y consideramos todos los estudios disponibles para cada gen, aunque haya una descompensación del número de estudios. Hemos generado de nuevo los rankings resumiendo el indicador de la variabilidad de expresión génica de cada estadístico a la mediana de los valores obtenidos con todos los estudios de cada gen, ordenamos estos valores de menor a mayor.

En *Homo sapiens* identificamos un total de 41975 genes únicos. En hombres identificamos 41973 porque ninguno de los tres estudios con muestras de hombres de la plataforma GPL10558 está completo, en esta plataforma somos capaces de anotar hasta 31426 genes, pero el estudio con muestras de hombres que más datos contiene recoge medidas de expresión génica de 31422 genes.

Podemos apreciar como las posiciones relativas de los rankings de los genes HK se repiten con los tres estadísticos en las tres condiciones: todas las muestras, muestras de hombres y muestras de mujeres.

Para el conjunto de todas las muestras el top 3 de genes con menor variabilidad es: 1. RPL19, 2. UBC y 3. PPIA. Si observamos sus posiciones en el ranking generado podemos ver cómo ocupan posiciones dentro de los 1000 primeros puestos para los dos estadísticos más robustos, el IQR/median y el MAD/median con niveles de variabilidad inferiores al 5 %. GAPDH ocupa posiciones posteriores aunque también muy altas respecto de los 41975 genes del ranking. Sin embargo, los genes HPRT y 18S presentan mayores niveles de variabilidad y ocupan posiciones más centrales, el salto en los niveles de variabilidad de estos genes es considerable.

En el análisis basado en el sexo, se observa que:

- El gen UBC aparece como el gen HK más estable en hombres, con una variación alrededor del 2 % en los 3 estadísticos y ocupando los puestos 118 (CV), 560 (IQR/median) y 519 (MAD/median), en cambio, en mujeres ocupan posiciones entre los 8000-10000, con una variación del 10-15 %.
- El gen RPL19 es el segundo más estable en hombres con una variación en torno al 3 %, ocupando posiciones aproximadas a 1500, mientras que en mujeres ocupa el primer lugar con un nivel de variación superior al 6 % y sus posiciones en el ranking están entre los puestos 4000 y 7000.
- El gen 18S, presenta los niveles de variabilidad más altos respecto del resto de genes HK en hombres, en torno al 7 %, ocupando posiciones posteriores a 15000, mientras que en mujeres presenta mayores niveles de variación y esta variación fluctúa considerablemente dependiendo del estimador.

En general los genes HK clásicos analizados presentan valores más altos de variabilidad en mujeres que en hombres. En hombres los genes HK se posicionan dentro de los primeros 18000 puestos (en la parte superior de los rankings), con un UBC muy alto en los rankings, y en mujeres vemos cómo ocupan posiciones más centrales con un mayor rango de variabilidad de la expresión génica, a excepción de RPL19 que se posiciona dentro de los 5000 primeros puestos en los tres estadísticos. Podemos detectar la presencia de genes con niveles de expresión más estable en mujeres. Todo ello apunta a que probablemente sería adecuado utilizar genes HK específicos por sexo.

Por su parte, en *Mus musculus* identificamos 47203 genes, apreciamos una diferencia importante entre Ppia y el resto de genes HK (Tabla de resultados *Mus musculus*). La ordenación relativa de los genes HK según los tres estadísticos vuelve a situar a Ppia, Ubc y Rpl19 como los genes más estables. Ppia presenta un nivel de variabilidad del 1 % en los tres estadísticos ocupando las oposiciones 205 (MAD/median), 475 (IQR/median) y 2710 (CV). En cambio, en el resto de genes HK, aunque presentan niveles de variación entre el 2 % y el 7 %, ocupan posiciones centrales en los rankings como es el caso del 18s y Gapdh comparado con Rpl19, e incluso finales, en el caso de Hprt.

5.4. Metaanálisis de resultados

El Rank Product ya ha sido utilizado previamente para seleccionar genes con estabilidad de expresión estable para su uso como controles internos de RT-qPCR unificando los resultados de microarrays. Este método es interesante porque consigue resultados independientes de las plataformas al combinar listas ordenadas [54]. En nuestro caso lo

utilizamos para combinar de forma robusta los resultados de los diferentes rankings de las plataformas, en un único ranking global por estadísticos. Una menor puntuación de RP se corresponde con un menor nivel de variabilidad.

5.4.1. Resultados globales del conjunto de todos los estudios por organismo

En el análisis con todas las muestras de *Homo sapiens* el gen HK con mejor puntuación es el UBC, seguido de PPIA y RPL19. Esto difiere de lo obtenido con la aproximación "naive" la cual colocaba al gen RPL19 como el menos variable, seguido de UBC y PPIA. Nos quedamos con los resultados del Rank Product ya que esta aproximación presenta una mayor potencia en la integración de resultados. Asimismo, el gen GAPDH continúa siendo el cuarto más estable, y los genes 18S y HPRT han obtenido las peores puntuaciones, siendo HPRT el que presenta el mayor nivel de variabilidad.

El resultado obtenido con *Mus musculus* continúa la línea de los resultados anteriores, Ppia es el gen con menor nivel de variabilidad en tejido adiposo de ratón, con una diferencia considerable del resto de los genes HK. El segundo gen más estable es el 18s, seguido de Rpl19, Ubc y Gapdh. De nuevo, Hprt es el gen que muestra los peores niveles de variación con la puntuación más alta de RP.

5.4.2. Resultados globales del conjunto de estudios por sexo

En hombres obtenemos a UBC como el gen HK más estable con la puntuación de RP más alta de los seis genes, seguido de PPIA y GAPDH, el orden de estos dos últimos depende del estimador de la variabilidad. RPL19 baja al cuarto puesto, cuando en el ranking "naive" ocupaba la segunda posición de estabilidad, y el 18S y HPRT muestran los mayores niveles de variabilidad en su expresión génica, obtienen las peores puntuaciones.

En las mujeres el gen RPL19 es el que obtiene la mejor puntuación RP, coincidiendo con el ranking "naive" es el gen que presenta menor nivel de variabilidad, seguido de UBC, PPIA, HPRT1 y GAPDH muestran diferencias en la posición que ocupan según el estimador de la variabilidad. Sin embargo, a diferencia del ranking "naive", los tres MetaRankings de las muestras mujeres coinciden en que el gen 18S es el más variable, obteniendo la mayor puntuación de RP y ocupando posiciones centrales.

Por otra parte, la revisión de los resultados obtenidos en el conjunto total de genes en su mayoría los primeros puestos de los rankings de estabilidad estan ocupados por pseudogenes, micro-ARNs, ARNs nucleares y regiones genómicas anotadas con el prefijo "LOC-", muchas de ellas sin entrada en las bases de datos de genes más populares como GeneCards. Llama especialmente la atención en mujeres, donde las primeras 1000 posiciones más estables en mujeres pertenecen en su gran mayoría a este tipo de elementos. En el resto de condiciones estudiadas también se observa este fenómeno pero no es necesarío bajar tantas poscionesen el metaRanking para localizar genes propiamente dichos con variaciones mínimas en su expresión.

5.4.3. Identificación de nuevos genes candidatos a genes de referencia

La validación de la estabilidad de la expresión génica en las condiciones de estudio es determinante para la confianza en los resultados obtenidos en un análisis cuantitativo del transcriptoma. La metodología planteada en este trabajo nos permite realizar este paso de evaluación a gran escala interrogando de forma masiva la estabilidad de la expresión génica. Desde las diferentes aproximaciones estadísticas somos capaces de identificar aquellos genes con expresión más estable de toda la batería de genes únicos analizados por las diferentes plataformas. Para cada condición estudiada hemos extraido los 10 genes con menor variabilidad en los niveles de expresión a lo largo de todos los estudios seleccionados en base a los resultados finales de los metaRankings.

Entre los genes identificados se encuentran otros genes *housekeeping*, que también participan de funciones vitales. Los dos genes más estables son el Citocromo b (CYTB) y la subunidad ribosomal 28S, tanto en humanos como en ratones, ambos genes se situan dentro de las primeras 5 posiciones de genes más estables en el metaRanking de estabilidad. En ratones, también aparecen, en los tres estadísticos, ambas subunidades ribosomales mitocondriales la 12s (mt-Rnr1) y 16s (mt-Rnr2). Identificamos más genes mitocondriales, como las subunidades 1 y 2 de la Citocromo oxidasa (COX1 y COX2), entre otros genes nucleares, detallados en los resultados del top 10 de genes más estables de cada condición.

Una parte considerable de estos genes más estables se repiten en humanos, hombres, mujeres y ratones, sobre todo aquellos codificados en el genoma mitocondrial. En este aspecto coinciden donde los tres estimadores de la variabilidad. Especialmente, el Citocromo b (CYTB) muestra unos niveles de variación mínimos en todas las condiciones y apenas aprecian diferencias en el nivel de variabilidad estimado entre hombres y mujeres.

Como ya hemos mencionado en el apartado anterior, confirmamos la diferencia en las posiciones relativas ocupadas en el metaRanking de los genes con mayor estabilidad de expresión más en mujeres. En hombres vemos que localizamos el top de genes más estable dentro de las primeras 150 posiciones, mientras que en mujeres, tenemos que bajar casi a la posición 1000 en los resultados obtenidos con el coeficiente de variación. Esto tampoco ocurre en el análisis de todas las muestras de *Homo sapiens*, ni en las de *Mus musculus*.

Todos los genes recogidos en los resultados de este trabajo presentan niveles mínimos de variabilidad en su expresión génica y potencialmente pueden ser utilizados como genes candidatos para controles internos en estudios de expresión génica del tejido adiposo.

5.5. Estructura de la información

5.5.1. Anotación de los genes housekeeping

Debido a los conflictos entontrados durante la fase de anotación de las matrices de expresión ha sido necesario corregir las funciones propias de cada plataforma y reanalizar la información. Tras la revisión de las principales fuentes de anotación, la mayor dificultad se ha encontrado con la anotación del gen 18S.

Las plataformas seleccionadas de la casa comercial Affymetrix, tanto de humanos como de ratones, emplean el gen 18S como control interno, es decir, las sondas que miden los niveles de expresión de este gen son sondas control. Según su página web, anteriormente al 2003, Affymetrix no incorporaba la anotación de las dianas genómicas de las sondas control, por tanto, no tenemos un identificador Symbol asociado a estas sondas en el *FeatureData* del objeto ExpressionSet descargado de cada estudio ni en las librerias de anotación correspondientes de Bioconductor. En consecuencia, al no tener ningún identificador Symbol asociado, al realizar el tratamiento de nulos, los valores de estas sondas eran eliminados. Ha sido necesario revisar la descripción de las sondas en los ficheros pesados de anotación de GEO (*full annotation tables*) y contrastar esta información con Affymetrix. Finalmente, los valores medidos con las tres sondas control positivas de la plataforma GPL570 de *Homo sapiens* y de la plataforma GPL1261 de *Mus musculus* que apuntaban a este gen se han considerado en el procesamiento estadístico, en cambio, las sondas control negativas de la plataforma GPL6244 se han excluido.

Por otro lado, en los microarrays de Illumina también se han encontrado contradicciones entre la información facilitada por GEO, Bioconductor y la casa comercial respecto al gen 18S. Concretamente, la plataforma GPL10558 presenta tres sondas normales ILMN_1703337, ILMN_3239610 y ILMN_3243593, que GEO y Bioconductor anotan con el identificador Symbol del gen 18S. Sin embargo, al comparar el resto de la información del *fData* de estas sondas, estas también apuntan a localizaciones genómicas: LOC441763, LOC100133565 y LOC100008588, respectivamente. Al contrastar estas localizaciones con el NCBI hemos encontrado que únicamente una de ellas, la sonda ILMN_3243593, con la anotación LOC100008588, se corresponde realmente con la secuencia de este gen, las otras dos sondas restantes corresponden a predicciones hipotéticas que se han desestimado, han sido retiradas de modelos y anotaciones posteriores en el NCBI, por lo tanto sus valores se han retirado del analisis. La plataforma GPL6947, únicamente contiene la sonda ILMN_1703337, por lo que sus medidas también han sido excluidas.

Adicionalmente, ante la presencia de sinónimos y alias puntuales para los genes Hprt, Gapdh y Ppia en *Mus musculus* se ha optado por unificar la nomenclatura de cada gen a un único identificador Symbol.

5.5.2. Procesamiento de metadatos

Uno de los mayores retos de este trabajo ha sido el filtrado y la estructuración de la información de GEO para lograr identificar los estudios con datos transcriptómicos de muestras pertenecientes a tejido adiposo y, posteriormente, clasificar las muestras en función del sexo para los estudios que aportaban esta información. El principal factor limitante ha sido la ausencia de vocabulario estandarizado en los metadatos de los registros de los estudios. Si bien el propio dato de expresión génica está recogido en un formato definido, como es una matriz de expresión, los metadatos asociados a los estudios almacenados en GEO no se adhieren a un vocabulario definido para describir las entidades biológicas.

Concretamente, los metadatos que recogen nuestra información clave, el tipo de tejido y el sexo de la muestra, se encuentran en campos de texto libre redactados por la persona que suministra los datos al repositorio. En nuestro caso, para identificar las muestras de tejido adiposo se decidió considerar únicamente aquellos registros que indicaran explici-

tamente que la muestra pertenecia a tejido adiposo con la palabra clave "adipose" en el campo de origen de la muestra (*sample_source*). La aplicación de este filtro restrictivo se tradujo en una reducción de los estudios identificados en GEO pero, a su vez, nos aseguraba que los registros obtenidos mediante herramientas programáticas pertenecen a muestras de tejido adiposo. Esta aproximación no fue posible con la información del sexo, ya que cada uno de los estudios que facilitaban este dato, lo anotaban de forma diferente o en diferentes atributos del registro muestral, por lo que fue necesario la revisión individual de cada uno de los estudios para conocer si aportaban o no información del sexo de sus muestras y como estaba descrito.

Es importante reflejar las limitaciones de la mineria de datos a gran escala en la ausencia de un vocabulario común y una estructura estandarizada de los metadatos de los estudios que facilite la automatización del procesamiento de la información almacenada de los repositorios públicos [50].

5.6. Automatización del proceso y escalabilidad del proyecto

En último lugar, este trabajo sienta las bases para la continuacón de la línea de investigación. El ya código desarrollado recoge la posibilidad de ampliación para la evaluación de la estabilidad de la expresión génica a otras plataformas o tejidos empleando. En caso de querer incorporar la información de más plataformas únicamente es necesario añadir su función propia de anotación al script 6 (ver tabla B.2). A partir de la información ya procesada que disponemos de GEO podemos extraer los identificadores GSE de los estudios que utilizan otras plataformas que se quieren analizar, identificar las muestras correspondientes al tejido adiposo en los estudios (script 4) y emplear el script principal (script 6) para descargar las matrices de expresión y realizar la anotación de las sondas y el procesamiento estadístico en un solo bucle. También se puede trasladar la metodología a la evaluación de los perfiles de expresión en otros tejidos modificando el origen de la muestra en el filtro inicial de GEO y en las consultas realizadas posteriormente con GEOmetadb para obtener la información muestral. Asimismo, nosotros hemos acotado el analisis a explorar la estabilidad de expresión de seis genes HK, pero como hemos visto podemos extraer los niveles de variabilidad de otros genes, si estan cubiertos por las plataformas que hemos seleccionado para en el analisis.

6. Conclusiones

1. Se ha observado una importante variación de la expresión génica dependiente de la tecnología. El gen más estable obtenido con una plataforma no tiene por qué serlo en otra.

2. Los genes RPL19, UBC, PPIA y GAPDH muestran los menores niveles de variabilidad relativa, en comparación con HPRT y 18S.

3. En *Mus musculus* apreciamos una diferencia importante entre los niveles de estabilidad de Ppia y el resto de genes *housekeeping*. Ppia ha presentado niveles de variabilidad del 1 % en las tres aproximaciones estadísticas realizadas.

4. Cuando se considera el sexo como variable, observamos diferencias en los niveles relativos de variabilidad obtenidos entre hombres y mujeres. También observamos diferencias cuando consideramos el sexo como una variable respecto del análisis con todas las muestras. Por tanto, puede ser apropiado considerar genes de referencia específicos por sexo para evitar introducir errores sistemáticos.

5. Existe la necesidad de incluir muestras de ambos sexos en los diseños experimentales de los estudios biomédicos. Ello posibilitaría la realización de estudios *in silico* como el que presentamos, incrementaría el poder estadístico de los trabajos y lo más importante, el conocimiento generado sería extensible para toda la población, y no solo a una parte de ella.

6. Se han identificado los genes más estables en los diversos escenarios descritos, para proponerlos como candidatos a genes de referencia que mejoren la normalización de los datos cuantitativos de expresión génica.

A. Anexo I. Consultas GEOmetadb

Query 1. Devuelve el identificador y la descripción de las plataformas recogidas en la cadena *GPL_string*:

SELECT gpl.gpl, gpl.title FROM gpl WHERE gpl.gpl IN (GPL_string)

Query 2. Devuelve para el estudio de interés (GSE), el identificador de la muestra (GSM), la plataforma (GPL) que se ha empleado, el origen de la muestra (sample_source) y el organismo al que pertenece:

SELECT DISTINCT gse.gse, gsm.gsm, gsm.gpl,gsm.source_name_ch1 as sample_source, gsm.organism_ch1 as organism
FROM
gsm JOIN gse_gsm ON gsm.gsm=gse_gsm.gsm
JOIN gse ON gse_gsm.gse=gse.gse
WHERE
gse.gse =GSE

Query 3. Devuelve para el estudio de interés (GSE), la información correspondiente a las muestras del tejido adiposo: el identificador de la muestra (GSM), la plataforma (GPL) que se ha empleado, el origen de la muestra (*sample_source*) y el organismo al que pertenece:

SELECT DISTINCT gse.gse, gsm.gsm, gsm.gpl, gsm.source_name_ch1 as sample_source, gsm.characteristics_ch1 as characteristics ,gsm.organism_ch1 as organism
FROM
gsm JOIN gse_gsm ON gsm.gsm=gse_gsm.gsm
JOIN gse ON gse_gsm.gse=gse.gse
WHERE
gsm.source_name_ch1 LIKE '%adipo%' AND
gsm.gpl = GPL AND
gse.gse = GSE

B. Anexo II. Tablas

Tabla B.1: Librerias y paquetes utilizados

Software	Versión
R	3.5.0

Software	Versión
Python	3.0

Paquete	Versión
sqldf	0.4-11
Bioconductor	3.11.1
GEOmetadb	1.44.0
GEOQuery	2.56.0
RankProd	3.14.0

Paquete	Versión
BeautifulSoup	4

Tabla B.2: Detalle del código desarrollado. Resumen de los scripts propios desarrollados, se indica el número referenciado en el texto, el lenguaje de programación y su funcionalidad.

Script	Nombre	Lenguaje	Función
1	WebscrapGEO	Python	Procesa el fichero HTML con los resultados de la consulta de GEO para extraer la información relevante
2	GSE_GPL_relational	Python	A partir de los datos extraidos del fichero HTML crea una tabla relacional de los identificadores GSE-GPL
3	createTableConteos	R	Extrae los identificadores GPL de las plataformas, obtiene sus descripciones y cuenta el número de estudios por plataforma
4	identifySamples	R	A partir de los identificadores GSE de cada plataforma obtiene la información de las muestras de cada uno de los estudios e identifica aquellas que pertenecen a tejido adiposo
5	identifySexSamples	R	A partir de la información recogida de la anotación del sexo de los estudios, obtiene los identificadores GSM de las muestras de cada sexo
6	masterScript	R	Realiza el procesamiento automatizado del conjunto de estudios de cada plataforma, contiene las funciones de descarga de los datos, anotación de las plataformas y el cálculo estadístico
7	EDA_RESULTS_automatizado	R	Exploración de resultados, extrae los valores medianos de cada estadístico para todos los genes y ordena los valores generando un ranking por condición
8	naive_Rank	R	Recoge todos los resultados de un mismo gen para cada estadístico y genera un ranking ordenando los valores medianos de menor a mayor
9	RankProduct	R	Formatea los resultados a la estructura de entrada de la función Rank Product, ejecuta la función y genera los metarankings para cada condición

Tabla B.3: Anotación del sexo en los estudios de GPL570 Extracto de la tabla GSE_GPL_sexsamples que recoge la información relevante en el procesamiento diferencial por sexos. Se muestra únicamente los resultados para la primera plataforma. La tabla completa se puede ver en el repositorio de github.

GSEID	GPL	nSamples_adi	indicadorMT	adiGSMs	InformacionSexo	muestrasSexo	anotacionSexo	sex_tags
GSE115799	GPL570	49	MONOTEJIDO	ALL	NO	NA	NO	NA
GSE120196	GPL570	14	MONOTEJIDO	ALL	NO	NA	NO	NA
GSE13070	GPL570	84	MULTITEJIDO	'GSM342612', (...)	NO	NA	NO	NA
GSE133346	GPL570	24	MONOTEJIDO	ALL	NO	NA	NO	NA
GSE13506	GPL570	49	MONOTEJIDO	ALL	NO	NA	NO	NA
GSE17090	GPL570	10	MONOTEJIDO	ALL	NO	NA	NO	NA
GSE17170	GPL570	70	MONOTEJIDO	ALL	NO	NA	NO	NA
GSE26339	GPL570	50	MONOTEJIDO	ALL	NO	NA	NO	NA
GSE27657	GPL570	18	MONOTEJIDO	ALL	SI	AMBOS	SI	Sex: Male///Sex: Female
GSE27916	GPL570	375	MONOTEJIDO	ALL	SI	AMBOS	SI	Sex: male///Sex: female
GSE27949	GPL570	33	MONOTEJIDO	ALL	NO	NA	NO	NA
GSE28005	GPL570	38	MONOTEJIDO	ALL	NO	NA	NO	NA
GSE3526	GPL570	10	MULTITEJIDO	'GSM80561', (...)	NO	NA	NO	NA
GSE39117	GPL570	54	MONOTEJIDO	ALL	NO	NA	NO	NA
GSE41168	GPL570	70	MULTITEJIDO	'GSM1009752', (...)	SI	MUJERES	SI	gender: Female
GSE61302	GPL570	15	MONOTEJIDO	ALL	SI	MUJERES	SI	gender: female
GSE66159	GPL570	38	MULTITEJIDO	'GSM1615577', (...)	SI	MUJERES	NO	NA
GSE71416	GPL570	20	MONOTEJIDO	ALL	SI	AMBOS	SI	Sex: Male///Sex: Female
GSE88837	GPL570	30	MONOTEJIDO	ALL	SI	MUJERES	SI	gender: Female
GSE9624	GPL570	11	MONOTEJIDO	ALL	SI	AMBOS	SI	gender: M///gender: F

Tabla B.4: Tabla GSE_sexsamples de *Homo sapiens*. Recoge la información muestral necesaria para realizar el analisis diferencial por sexos de todos los estudios de *Homo sapiens*. En el repositorio github se puede acceder a la tabla equivalente para *Mus musculus*.

GSEID	GPL	indicadorMT	muestrasSexo	anotacionSexo	sex_tags	nSamples_adi	nMales	nFemales	MaleGSMs	FemaleGSMs
GSE27657	GPL570	MONOTEJIDO	AMBOS	SI	Sex: Male///Sex: Female	18	4	14	GSM685065', (...)	GSM685066',(...)
GSE27916	GPL570	MONOTEJIDO	AMBOS	SI	Sex: male///Sex: female	375	113	262	GSM689465', (...)	GSM689463',(...)
GSE41168	GPL570	MULTITEJIDO	MUJERES	SI	gender: Female	70	0	70	NULL	GSM1009752',(...)
GSE61302	GPL570	MONOTEJIDO	MUJERES	SI	gender: female	15	0	15	NULL	GSM1501795',(...)
GSE66159	GPL570	MULTITEJIDO	MUJERES	NO	NA	38	0	38	NULL	GSM1615577',(...)
GSE71416	GPL570	MONOTEJIDO	AMBOS	SI	Sex: Male///Sex: Female	20	5	15	GSM1833926', (...)	GSM1833927',(...)
GSE88837	GPL570	MONOTEJIDO	MUJERES	SI	gender: Female	30	0	30	NULL	GSM2349936',(...)
GSE9624	GPL570	MONOTEJIDO	AMBOS	SI	gender: M///gender: F	11	10	1	GSM243215', (...)	NULL
GSE25401	GPL6244	MONOTEJIDO	MUJERES	SI	genotype: female	56	0	56	NULL	GSM623761',(...)
GSE25910	GPL6244	MONOTEJIDO	MUJERES	SI	gender: female	36	0	36	NULL	GSM636662',(...)
GSE33070	GPL6244	MONOTEJIDO	AMBOS	SI	gender: Male///gender: Female	26	10	16	GSM819098', (...)	GSM819096',(...)
GSE73655	GPL6244	MONOTEJIDO	AMBOS	SI	gender: Male///gender: Female	20	6	14	GSM1900070', (...)	GSM1900073',(...)
GSE41223	GPL6244	MONOTEJIDO	AMBOS	SI	gender: Male///gender: Female	20	4	16	GSM1011037', (...)	GSM1011041',(...)
GSE54280	GPL6244	MONOTEJIDO	AMBOS	SI	gender: male///gender: female	12	6	6	GSM1311783', (...)	GSM1311785',(...)
GSE73108	GPL10558	MONOTEJIDO	MUJERES	SI	gender: female	12	0	12	NULL	GSM1886833',(...)
GSE65221	GPL10558	MONOTEJIDO	AMBOS	SI	gender: Male///gender: Female	136	63	73	GSM1590152', (...)	GSM1590151',(...)
GSE119717	GPL10558	MULTITEJIDO	HOMBRES	SI	Sex: Male	60	60	0	GSM3381381', (...)	NULL
GSE115645	GPL10558	MONOTEJIDO	AMBOS	SI	Sex: male///Sex: female	24	21	3	GSM3186618', (...)	GSM3186636',(...)
GSE43471	GPL6947	MONOTEJIDO	MUJERES	SI	gender: female	96	0	96	NULL	GSM1063183',(...)
GSE32512	GPL6947	MONOTEJIDO	HOMBRES	NO	NA	204	204	0	GSM804886', (...)	NULL
GSE29231	GPL6947	MONOTEJIDO	MUJERES	SI	gender: Female	24	0	24	NULL	GSM722948',(...)
GSE29226	GPL6947	MONOTEJIDO	MUJERES	SI	gender: Female	24	0	24	NULL	GSM722735',(...)
GSE27666	GPL6947	MONOTEJIDO	HOMBRES	SI	Sex: male	175	175	0	GSM685160', (...)	NULL
GSE112307	GPL6947	MONOTEJIDO	MUJERES	NO	NA	54	0	54	NULL	GSM3067430',(...)

Tabla B.5: Resumen de los estudios elegibles incluidos y excluidos en el análisis de *Homo sapiens*.

GSE ID	GPL ID	Muestras GPL	Muestras	indicadorMT	Estado	Criterio exclusión
GSE27916	GPL570	375	375	MONOTEJIDO	INCLUIDO	-
GSE13070	GPL570	364	84	MULTITEJIDO	INCLUIDO	-
GSE17170	GPL570	70	70	MONOTEJIDO	INCLUIDO	-
GSE41168	GPL570	140	70	MULTITEJIDO	INCLUIDO	-
GSE39117	GPL570	54	54	MONOTEJIDO	INCLUIDO	-
GSE39118	GPL570	54	54	MONOTEJIDO	EXCLUIDO	2
GSE26339	GPL570	50	50	MONOTEJIDO	INCLUIDO	-
GSE115799	GPL570	49	49	MONOTEJIDO	INCLUIDO	-
GSE13506	GPL570	49	49	MONOTEJIDO	INCLUIDO	-
GSE28005	GPL570	38	38	MONOTEJIDO	INCLUIDO	-
GSE66159	GPL570	76	38	MULTITEJIDO	EXCLUIDO	2
GSE66162	GPL570	150	38	MULTITEJIDO	INCLUIDO	-
GSE27949	GPL570	33	33	MONOTEJIDO	INCLUIDO	-
GSE27951	GPL570	33	33	MONOTEJIDO	EXCLUIDO	2
GSE88837	GPL570	30	30	MONOTEJIDO	INCLUIDO	-
GSE133346	GPL570	24	24	MONOTEJIDO	INCLUIDO	-
GSE71416	GPL570	20	20	MONOTEJIDO	INCLUIDO	-
GSE27657	GPL570	18	18	MONOTEJIDO	INCLUIDO	-
GSE61302	GPL570	15	15	MONOTEJIDO	INCLUIDO	-
GSE9624	GPL570	11	11	MONOTEJIDO	INCLUIDO	-
GSE17090	GPL570	10	10	MONOTEJIDO	INCLUIDO	-
GSE120196	GPL570	10	10	MONOTEJIDO	INCLUIDO	-
GSE3526	GPL570	353	10	MULTITEJIDO	INCLUIDO	-
GSE124226	GPL570	8	8	MONOTEJIDO	EXCLUIDO	3
GSE98421	GPL570	8	8	MONOTEJIDO	EXCLUIDO	3
GSE26626	GPL570	4	4	MONOTEJIDO	EXCLUIDO	3
GSE66084	GPL570	12	3	MULTITEJIDO	EXCLUIDO	3
GSE20033	GPL570	19	3	MULTITEJIDO	EXCLUIDO	3
GSE18391	GPL570	2	2	MONOTEJIDO	EXCLUIDO	3
GSE19238	GPL570	2	2	MONOTEJIDO	EXCLUIDO	3
GSE12843	GPL570	4	2	MULTITEJIDO	EXCLUIDO	3
GSE37896	GPL570	8	2	MULTITEJIDO	EXCLUIDO	3
GSE43346	GPL570	68	1	MULTITEJIDO	EXCLUIDO	3
GSE99316	GPL570	68	1	MULTITEJIDO	EXCLUIDO	3
GSE2109	GPL570	2158	1	MULTITEJIDO	EXCLUIDO	3
GSE34200	GPL570	12	0	MULTITEJIDO	EXCLUIDO	1
GSE17312	GPL570	17	0	MULTITEJIDO	EXCLUIDO	1
GSE25402	GPL6244	92	92	MONOTEJIDO	EXCLUIDO	2
GSE25401	GPL6244	56	56	MONOTEJIDO	INCLUIDO	-
GSE70529	GPL6244	36	36	MONOTEJIDO	INCLUIDO	-
GSE62832	GPL6244	36	36	MONOTEJIDO	INCLUIDO	-
GSE25910	GPL6244	36	36	MONOTEJIDO	INCLUIDO	-
GSE20571	GPL6244	27	27	MONOTEJIDO	INCLUIDO	-
GSE37324	GPL6244	26	26	MONOTEJIDO	INCLUIDO	-
GSE33070	GPL6244	26	26	MONOTEJIDO	INCLUIDO	-

GSE73655	GPL6244	20	20	MONOTEJIDO	INCLUIDO	-
GSE34302	GPL6244	20	20	MONOTEJIDO	INCLUIDO	-
GSE41223	GPL6244	20	20	MONOTEJIDO	INCLUIDO	-
GSE38792	GPL6244	18	18	MONOTEJIDO	INCLUIDO	-
GSE54280	GPL6244	12	12	MONOTEJIDO	INCLUIDO	-
GSE59325	GPL6244	10	10	MONOTEJIDO	INCLUIDO	-
GSE56635	GPL6244	6	0	MULTITEJIDO	EXCLUIDO	1
GSE58559	GPL10558	170	170	MONOTEJIDO	INCLUIDO	-
GSE65221	GPL10558	136	136	MONOTEJIDO	INCLUIDO	-
GSE72158	GPL10558	84	84	MONOTEJIDO	INCLUIDO	-
GSE119717	GPL10558	119	60	MULTITEJIDO	INCLUIDO	-
GSE115645	GPL10558	24	24	MONOTEJIDO	INCLUIDO	-
GSE73108	GPL10558	12	12	MONOTEJIDO	INCLUIDO	-
GSE42809	GPL10558	12	12	MONOTEJIDO	INCLUIDO	-
GSE133803	GPL10558	6	6	MONOTEJIDO	EXCLUIDO	3
GSE42523	GPL10558	6	6	MONOTEJIDO	EXCLUIDO	3
GSE48774	GPL10558	9	6	MULTITEJIDO	EXCLUIDO	3
GSE42560	GPL10558	6	6	MONOTEJIDO	EXCLUIDO	3
GSE55695	GPL10558	12	3	MULTITEJIDO	EXCLUIDO	3
GSE130393	GPL10558	8	0	MONOTEJIDO	EXCLUIDO	1
GSE53081	GPL10558	24	0	MULTITEJIDO	EXCLUIDO	1
GSE32512	GPL6947	204	204	MONOTEJIDO	INCLUIDO	-
GSE27666	GPL6947	175	175	MONOTEJIDO	INCLUIDO	-
GSE22070	GPL6947	142	142	MONOTEJIDO	INCLUIDO	-
GSE43471	GPL6947	96	96	MONOTEJIDO	INCLUIDO	-
GSE27121	GPL6947	70	70	MONOTEJIDO	INCLUIDO	-
GSE112307	GPL6947	54	54	MONOTEJIDO	INCLUIDO	-
GSE23506	GPL6947	36	36	MONOTEJIDO	INCLUIDO	-
GSE29231	GPL6947	24	24	MONOTEJIDO	INCLUIDO	-
GSE29226	GPL6947	24	24	MONOTEJIDO	INCLUIDO	-
GSE30652	GPL6947	239	2	MULTITEJIDO	EXCLUIDO	3
GSE30654	GPL6947	239	2	MULTITEJIDO	EXCLUIDO	3

Tabla B.6: Resumen de los estudios elegibles incluidos y excluidos en el análisis de *Mus musculus*.

GSE ID	GPL ID	Muestras GPL	Muestras	indicadorMT	Estado	Criterio exclusión
GSE140954	GPL1261	22	12	MULTITEJIDO	EXCLUIDO	2
GSE140953	GPL1261	12	12	MONOTEJIDO	INCLUIDO	-
GSE117353	GPL1261	10	10	MONOTEJIDO	EXCLUIDO	2
GSE117352	GPL1261	10	10	MONOTEJIDO	INCLUIDO	-
GSE110531	GPL1261	23	23	MONOTEJIDO	INCLUIDO	-
GSE66132	GPL1261	16	16	MONOTEJIDO	INCLUIDO	-
GSE66131	GPL1261	8	8	MONOTEJIDO	EXCLUIDO	3
GSE66130	GPL1261	8	8	MONOTEJIDO	EXCLUIDO	3
GSE106271	GPL1261	8	8	MONOTEJIDO	EXCLUIDO	3
GSE106270	GPL1261	3	3	MONOTEJIDO	EXCLUIDO	3
GSE87854	GPL1261	12	12	MONOTEJIDO	INCLUIDO	-
GSE87853	GPL1261	11	11	MONOTEJIDO	INCLUIDO	-
GSE77943	GPL1261	62	21	MULTITEJIDO	INCLUIDO	-
GSE97240	GPL1261	74	34	MULTITEJIDO	INCLUIDO	-
GSE71367	GPL1261	27	27	MONOTEJIDO	INCLUIDO	-
GSE67389	GPL1261	16	16	MONOTEJIDO	INCLUIDO	-
GSE51080	GPL1261	18	18	MONOTEJIDO	INCLUIDO	-
GSE53403	GPL1261	16	16	MONOTEJIDO	INCLUIDO	-
GSE46209	GPL1261	21	6	MULTITEJIDO	EXCLUIDO	3
GSE36492	GPL1261	2	2	MONOTEJIDO	EXCLUIDO	3
GSE30247	GPL1261	16	16	MONOTEJIDO	INCLUIDO	-
GSE39562	GPL1261	26	26	MONOTEJIDO	INCLUIDO	-
GSE31940	GPL1261	8	8	MONOTEJIDO	EXCLUIDO	3
GSE27309	GPL1261	10	10	MONOTEJIDO	INCLUIDO	-
GSE24207	GPL1261	73	3	MULTITEJIDO	EXCLUIDO	3
GSE21754	GPL1261	2	2	MONOTEJIDO	EXCLUIDO	3
GSE17923	GPL1261	6	6	MONOTEJIDO	EXCLUIDO	3
GSE7852	GPL1261	18	6	MULTITEJIDO	EXCLUIDO	3
GSE13432	GPL1261	12	12	MONOTEJIDO	INCLUIDO	-
GSE13585	GPL1261	12	6	MULTITEJIDO	EXCLUIDO	3
GSE13582	GPL1261	6	6	MONOTEJIDO	EXCLUIDO	3
GSE10246	GPL1261	182	4	MULTITEJIDO	EXCLUIDO	3
GSE9954	GPL1261	70	3	MULTITEJIDO	EXCLUIDO	3
GSE8044	GPL1261	6	6	MONOTEJIDO	EXCLUIDO	3
GSE122374	GPL6246	6	6	MONOTEJIDO	EXCLUIDO	3
GSE118575	GPL6246	2	0	MULTITEJIDO	EXCLUIDO	1
GSE80148	GPL6246	14	6	MULTITEJIDO	EXCLUIDO	3
GSE80146	GPL6246	6	6	MONOTEJIDO	EXCLUIDO	3
GSE57513	GPL6246	6	6	MONOTEJIDO	EXCLUIDO	3
GSE79166	GPL6246	8	8	MONOTEJIDO	EXCLUIDO	3
GSE79164	GPL6246	8	8	MONOTEJIDO	EXCLUIDO	3
GSE79434	GPL6246	73	24	MULTITEJIDO	INCLUIDO	-
GSE61121	GPL6246	11	5	MULTITEJIDO	EXCLUIDO	3
GSE56635	GPL6246	3	1	MULTITEJIDO	EXCLUIDO	3
GSE56634	GPL6246	3	1	MULTITEJIDO	EXCLUIDO	3

GSE56852	GPL6246	4	4	MONOTEJIDO	EXCLUIDO	3
GSE63358	GPL6246	2	1	MULTITEJIDO	EXCLUIDO	3
GSE54652	GPL6246	288	48	MULTITEJIDO	EXCLUIDO	2
GSE54650	GPL6246	288	48	MULTITEJIDO	INCLUIDO	-
GSE55272	GPL6246	36	12	MULTITEJIDO	INCLUIDO	-
GSE51931	GPL6246	18	6	MULTITEJIDO	EXCLUIDO	3
GSE35026	GPL6246	24	24	MONOTEJIDO	INCLUIDO	-
GSE38321	GPL6246	10	10	MONOTEJIDO	INCLUIDO	-
GSE37514	GPL6246	15	15	MONOTEJIDO	INCLUIDO	-
GSE27525	GPL6246	2	2	MONOTEJIDO	EXCLUIDO	3
GSE26442	GPL6246	9	9	MONOTEJIDO	EXCLUIDO	3
GSE24940	GPL6246	40	3	MULTITEJIDO	EXCLUIDO	3
GSE20121	GPL6246	24	0	MULTITEJIDO	EXCLUIDO	1
GSE125900	GPL6887	4	4	MONOTEJIDO	EXCLUIDO	3
GSE123990	GPL6887	6	6	MONOTEJIDO	EXCLUIDO	3
GSE120243	GPL6887	27	9	MULTITEJIDO	EXCLUIDO	3
GSE97145	GPL6887	11	11	MONOTEJIDO	INCLUIDO	-
GSE82328	GPL6887	4	4	MONOTEJIDO	EXCLUIDO	3
GSE63198	GPL6887	6	6	MONOTEJIDO	EXCLUIDO	3
GSE70300	GPL6887	24	10	MULTITEJIDO	INCLUIDO	-
GSE64718	GPL6887	33	33	MONOTEJIDO	INCLUIDO	-
GSE64060	GPL6887	6	6	MONOTEJIDO	EXCLUIDO	3
GSE62612	GPL6887	41	11	MULTITEJIDO	INCLUIDO	-
GSE62937	GPL6887	24	24	MONOTEJIDO	INCLUIDO	-
GSE54189	GPL6887	18	9	MULTITEJIDO	EXCLUIDO	3
GSE55057	GPL6887	12	12	MONOTEJIDO	INCLUIDO	-
GSE56711	GPL6887	12	0	MULTITEJIDO	EXCLUIDO	1
GSE39549	GPL6887	91	40	MULTITEJIDO	INCLUIDO	-
GSE53980	GPL6887	3	3	MONOTEJIDO	EXCLUIDO	3
GSE50647	GPL6887	42	42	MONOTEJIDO	INCLUIDO	-
GSE37238	GPL6887	4	4	MONOTEJIDO	EXCLUIDO	3
GSE35431	GPL6887	12	6	MULTITEJIDO	EXCLUIDO	3
GSE30116	GPL6887	6	6	MONOTEJIDO	EXCLUIDO	3
GSE33684	GPL6885	16	16	MONOTEJIDO	INCLUIDO	-
GSE123990	GPL6885	12	12	MONOTEJIDO	EXCLUIDO	4
GSE113808	GPL6885	16	16	MONOTEJIDO	INCLUIDO	-
GSE111735	GPL6885	8	8	MONOTEJIDO	EXCLUIDO	3
GSE118034	GPL6885	24	6	MULTITEJIDO	EXCLUIDO	3
GSE104128	GPL6885	31	16	MULTITEJIDO	INCLUIDO	-
GSE97910	GPL6885	72	72	MONOTEJIDO	INCLUIDO	-
GSE63761	GPL6885	8	8	MONOTEJIDO	EXCLUIDO	3
GSE70857	GPL6885	96	48	MULTITEJIDO	INCLUIDO	-
GSE64770	GPL6885	112	0	MONOTEJIDO	EXCLUIDO	1
GSE62324	GPL6885	12	12	MONOTEJIDO	INCLUIDO	-
GSE40428	GPL6885	16	8	MULTITEJIDO	EXCLUIDO	3
GSE57659	GPL6885	179	179	MONOTEJIDO	INCLUIDO	-
GSE39313	GPL6885	48	16	MULTITEJIDO	INCLUIDO	-
GSE29502	GPL6885	8	8	MONOTEJIDO	EXCLUIDO	3
GSE118876	GPL16570	12	12	MONOTEJIDO	INCLUIDO	-

GSE113507	GPL16570	6	6	MONOTEJIDO	EXCLUIDO	3
GSE111809	GPL16570	14	4	MULTITEJIDO	EXCLUIDO	3
GSE111744	GPL16570	4	4	MONOTEJIDO	EXCLUIDO	3
GSE104955	GPL16570	16	16	MONOTEJIDO	INCLUIDO	-
GSE99064	GPL16570	8	8	MONOTEJIDO	EXCLUIDO	3
GSE87661	GPL16570	24	24	MONOTEJIDO	INCLUIDO	-
GSE84809	GPL16570	12	8	MULTITEJIDO	EXCLUIDO	3
GSE79711	GPL16570	12	12	MONOTEJIDO	INCLUIDO	-
GSE60150	GPL16570	37	37	MONOTEJIDO	INCLUIDO	-

Tabla B.7: Detalle de los estudios de *Homo sapiens* seleccionados.

Estudio	Plataforma	Muestras	indicadorMT	GSM IDs
GSE120196	GPL570	10	MONOTEJIDO	ALL
GSE115799	GPL570	49	MONOTEJIDO	ALL
GSE13070	GPL570	84	MULTITEJIDO	'GSM342612', 'GSM342613', 'GSM342618', 'GSM342619', 'GSM342624', 'GSM342625', 'GSM342630', 'GSM342631', 'GSM342636', 'GSM342637', 'GSM342640', 'GSM342641', 'GSM342646', 'GSM342647', 'GSM342652', 'GSM342657', 'GSM342658', 'GSM342663', 'GSM342664', 'GSM342669', 'GSM342670', 'GSM342675', 'GSM342676', 'GSM342681', 'GSM342682', 'GSM342687', 'GSM342688', 'GSM342693', 'GSM342694', 'GSM342699', 'GSM342700', 'GSM342705', 'GSM342710', 'GSM342711', 'GSM342716', 'GSM342721', 'GSM342722', 'GSM342727', 'GSM342728', 'GSM342733', 'GSM342734', 'GSM342739', 'GSM342740', 'GSM342745', 'GSM342746', 'GSM342751', 'GSM342752', 'GSM342757', 'GSM342758', 'GSM342763', 'GSM342764', 'GSM342769', 'GSM342770', 'GSM342775', 'GSM342776', 'GSM342781', 'GSM342782', 'GSM342787', 'GSM342788', 'GSM342793', 'GSM342794', 'GSM342799', 'GSM342800', 'GSM342805', 'GSM342806', 'GSM342811', 'GSM342812', 'GSM342817', 'GSM342818', 'GSM342823', 'GSM342824', 'GSM342829', 'GSM342830', 'GSM342835', 'GSM342840', 'GSM342841', 'GSM342846', 'GSM342847', 'GSM342852', 'GSM342853', 'GSM342858', 'GSM342859', 'GSM342864', 'GSM342865'
GSE133346	GPL570	24	MONOTEJIDO	ALL
GSE13506	GPL570	49	MONOTEJIDO	ALL
GSE17090	GPL570	10	MONOTEJIDO	ALL
GSE17170	GPL570	70	MONOTEJIDO	ALL

GSE26339	GPL570	50	MONOTEJIDO	ALL
GSE27657	GPL570	18	MONOTEJIDO	ALL
GSE27916	GPL570	375	MONOTEJIDO	ALL
GSE27949	GPL570	33	MONOTEJIDO	ALL
GSE28005	GPL570	38	MONOTEJIDO	ALL
GSE3526	GPL570	10	MULTITEJIDO	'GSM80561', 'GSM80562', 'GSM80563', 'GSM80564', 'GSM80580', 'GSM80583', 'GSM80584', 'GSM80588', 'GSM80589', 'GSM80590'
GSE39117	GPL570	54	MONOTEJIDO	ALL
GSE41168	GPL570	70	MULTITEJIDO	'GSM1009752', 'GSM1009753', 'GSM1009756', 'GSM1009757', 'GSM1009760', 'GSM1009761', 'GSM1009764', 'GSM1009765', 'GSM1009768', 'GSM1009769', 'GSM1009772', 'GSM1009773', 'GSM1009776', 'GSM1009777', 'GSM1009778', 'GSM1009779', 'GSM1009782', 'GSM1009783', 'GSM1009784', 'GSM1009785', 'GSM1009788', 'GSM1009789', 'GSM1009796', 'GSM1009797', 'GSM1009800', 'GSM1009801', 'GSM1009804', 'GSM1009805', 'GSM1009806', 'GSM1009807', 'GSM1009810', 'GSM1009811', 'GSM1009812', 'GSM1009813', 'GSM1009816', 'GSM1009817', 'GSM1009820', 'GSM1009821', 'GSM1009824', 'GSM1009825', 'GSM1009828', 'GSM1009829', 'GSM1009832', 'GSM1009833', 'GSM1009836', 'GSM1009837', 'GSM1009840', 'GSM1009841', 'GSM1009844', 'GSM1009845', 'GSM1009848', 'GSM1009849', 'GSM1009854', 'GSM1009855', 'GSM1009858', 'GSM1009859', 'GSM1009862', 'GSM1009863', 'GSM1009866', 'GSM1009867', 'GSM1009872', 'GSM1009873', 'GSM1009874', 'GSM1009875', 'GSM1009878', 'GSM1009879', 'GSM1009882', 'GSM1009883', 'GSM1009886', 'GSM1009887'
GSE61302	GPL570	15	MONOTEJIDO	ALL

GSE66159	GPL570	38	MULTITEJIDO	'GSM1615577', 'GSM1615578', 'GSM1615579', 'GSM1615580', 'GSM1615581', 'GSM1615582', 'GSM1615583', 'GSM1615584', 'GSM1615585', 'GSM1615586', 'GSM1615587', 'GSM1615588', 'GSM1615589', 'GSM1615590', 'GSM1615591', 'GSM1615592', 'GSM1615593', 'GSM1615594', 'GSM1615595', 'GSM1615596', 'GSM1615597', 'GSM1615598', 'GSM1615599', 'GSM1615600', 'GSM1615601', 'GSM1615602', 'GSM1615603', 'GSM1615604', 'GSM1615605', 'GSM1615606', 'GSM1615607', 'GSM1615608', 'GSM1615609', 'GSM1615610', 'GSM1615611', 'GSM1615612', 'GSM1615613', 'GSM1615614'
GSE71416	GPL570	20	MONOTEJIDO	ALL
GSE88837	GPL570	30	MONOTEJIDO	ALL
GSE9624	GPL570	11	MONOTEJIDO	ALL
GSE25401	GPL6244	56	MONOTEJIDO	ALL
GSE70529	GPL6244	36	MONOTEJIDO	ALL
GSE62832	GPL6244	36	MONOTEJIDO	ALL
GSE25910	GPL6244	36	MONOTEJIDO	ALL
GSE20571	GPL6244	27	MONOTEJIDO	ALL
GSE37324	GPL6244	26	MONOTEJIDO	ALL
GSE33070	GPL6244	26	MONOTEJIDO	ALL
GSE73655	GPL6244	20	MONOTEJIDO	ALL
GSE34302	GPL6244	20	MONOTEJIDO	ALL
GSE41223	GPL6244	20	MONOTEJIDO	ALL
GSE38792	GPL6244	18	MONOTEJIDO	ALL
GSE54280	GPL6244	12	MONOTEJIDO	ALL
GSE59325	GPL6244	10	MONOTEJIDO	ALL
GSE73108	GPL10558	12	MONOTEJIDO	ALL
GSE72158	GPL10558	84	MONOTEJIDO	ALL
GSE65221	GPL10558	136	MONOTEJIDO	ALL
GSE58559	GPL10558	170	MONOTEJIDO	ALL
GSE42809	GPL10558	12	MONOTEJIDO	ALL

GSE119717	GPL10558	60	MULTITEJIDO	'GSM3381381', 'GSM3381382', 'GSM3381383', 'GSM3381384', 'GSM3381389', 'GSM3381391', 'GSM3381397', 'GSM3381398', 'GSM3381401', 'GSM3381402', 'GSM3381404', 'GSM3381405', 'GSM3381408', 'GSM3381409', 'GSM3381410', 'GSM3381412', 'GSM3381413', 'GSM3381416', 'GSM3381419', 'GSM3381420', 'GSM3381421', 'GSM3381423', 'GSM3381425', 'GSM3381426', 'GSM3381427', 'GSM3381428', 'GSM3381429', 'GSM3381430', 'GSM3381432', 'GSM3381435', 'GSM3381438', 'GSM3381440', 'GSM3381441', 'GSM3381444', 'GSM3381447', 'GSM3381449', 'GSM3381453', 'GSM3381454', 'GSM3381460', 'GSM3381462', 'GSM3381464', 'GSM3381465', 'GSM3381469', 'GSM3381470', 'GSM3381474', 'GSM3381477', 'GSM3381479', 'GSM3381480', 'GSM3381482', 'GSM3381483', 'GSM3381487', 'GSM3381489', 'GSM3381490', 'GSM3381493', 'GSM3381495', 'GSM3381496', 'GSM3381497', 'GSM3381499', 'GSM3381504', 'GSM3381505'
GSE115645	GPL10558	24	MONOTEJIDO	ALL
GSE43471	GPL6947	96	MONOTEJIDO	ALL
GSE32512	GPL6947	204	MONOTEJIDO	ALL
GSE29231	GPL6947	24	MONOTEJIDO	ALL
GSE29226	GPL6947	24	MONOTEJIDO	ALL
GSE27666	GPL6947	175	MONOTEJIDO	ALL
GSE27121	GPL6947	70	MONOTEJIDO	ALL
GSE23506	GPL6947	36	MONOTEJIDO	ALL
GSE22070	GPL6947	142	MONOTEJIDO	ALL
GSE112307	GPL6947	54	MONOTEJIDO	ALL

Tabla B.8: Detalle de los estudios de *Homo sapiens* seleccionados.

Estudio	Plataforma	Muestras	indicadorMT	GSM IDs
GSE117352	GPL1261	10	MONOTEJIDO	ALL
GSE140953	GPL1261	12	MONOTEJIDO	ALL
GSE110531	GPL1261	23	MONOTEJIDO	ALL
GSE66132	GPL1261	16	MONOTEJIDO	ALL
GSE87854	GPL1261	12	MONOTEJIDO	ALL
GSE87853	GPL1261	11	MONOTEJIDO	ALL
GSE77943	GPL1261	21	MULTITEJIDO	'GSM2061955', 'GSM2061962', 'GSM2061967', 'GSM2061971', 'GSM2061976', 'GSM2061981', 'GSM2061986', 'GSM2061991', 'GSM2061996', 'GSM2062001', 'GSM2062009', 'GSM2062011', 'GSM2062016', 'GSM2062021', 'GSM2062027', 'GSM2062032', 'GSM2062037', 'GSM2062042', 'GSM2062047', 'GSM2062052', 'GSM2062058'
GSE97240	GPL1261	34	MULTITEJIDO	'GSM2560037', 'GSM2560038', 'GSM2560039', 'GSM2560040', 'GSM2560041', 'GSM2560042', 'GSM2560043', 'GSM2560044', 'GSM2560045', 'GSM2560046', 'GSM2560047', 'GSM2560048', 'GSM2560049', 'GSM2560050', 'GSM2560051', 'GSM2560052', 'GSM2560054', 'GSM2560056', 'GSM2560058', 'GSM2560061', 'GSM2560062', 'GSM2560064', 'GSM2560066', 'GSM2560068', 'GSM2560071', 'GSM2560072', 'GSM2560074', 'GSM2560075', 'GSM2560077', 'GSM2560081', 'GSM2560082', 'GSM2560084', 'GSM2560086', 'GSM2560088'
GSE71367	GPL1261	27	MONOTEJIDO	ALL
GSE67389	GPL1261	16	MONOTEJIDO	ALL
GSE51080	GPL1261	18	MONOTEJIDO	ALL
GSE53403	GPL1261	16	MONOTEJIDO	ALL
GSE30247	GPL1261	16	MONOTEJIDO	ALL
GSE39562	GPL1261	26	MONOTEJIDO	ALL
GSE27309	GPL1261	10	MONOTEJIDO	ALL
GSE13432	GPL1261	12	MONOTEJIDO	ALL
GSE38321	GPL6246	10	MONOTEJIDO	ALL

GSE79434	GPL6246	24	MULTITEJIDO	'GSM2095226', 'GSM2095227', 'GSM2095228', 'GSM2095229', 'GSM2095230', 'GSM2095231', 'GSM2095232', 'GSM2095233', 'GSM2095234', 'GSM2095235', 'GSM2095236', 'GSM2095237', 'GSM2095287', 'GSM2095288', 'GSM2095289', 'GSM2095290', 'GSM2095291', 'GSM2095292', 'GSM2095293', 'GSM2095294', 'GSM2095295', 'GSM2095296', 'GSM2095297', 'GSM2095298'
GSE54650	GPL6246	48	MULTITEJIDO	'GSM1321062', 'GSM1321063', 'GSM1321064', 'GSM1321065', 'GSM1321066', 'GSM1321067', 'GSM1321068', 'GSM1321069', 'GSM1321070', 'GSM1321071', 'GSM1321072', 'GSM1321073', 'GSM1321074', 'GSM1321075', 'GSM1321076', 'GSM1321077', 'GSM1321078', 'GSM1321079', 'GSM1321080', 'GSM1321081', 'GSM1321082', 'GSM1321083', 'GSM1321084', 'GSM1321085', 'GSM1321254', 'GSM1321255', 'GSM1321256', 'GSM1321257', 'GSM1321258', 'GSM1321259', 'GSM1321260', 'GSM1321261', 'GSM1321262', 'GSM1321263', 'GSM1321264', 'GSM1321265', 'GSM1321266', 'GSM1321267', 'GSM1321268', 'GSM1321269', 'GSM1321270', 'GSM1321271', 'GSM1321272', 'GSM1321273', 'GSM1321274', 'GSM1321275', 'GSM1321276', 'GSM1321277'
GSE55272	GPL6246	12	MULTITEJIDO	'GSM1333082', 'GSM1333083', 'GSM1333084', 'GSM1333085', 'GSM1333086', 'GSM1333087', 'GSM1333100', 'GSM1333101', 'GSM1333102', 'GSM1333103', 'GSM1333104', 'GSM1333105'
GSE35026	GPL6246	24	MONOTEJIDO	ALL
GSE37514	GPL6246	15	MONOTEJIDO	ALL
GSE70300	GPL6887	10	MULTITEJIDO	'GSM1723369', 'GSM1723371', 'GSM1723376', 'GSM1723377', 'GSM1723379', 'GSM1723380', 'GSM1723382', 'GSM1723383', 'GSM1723385', 'GSM1723389'
GSE97145	GPL6887	11	MONOTEJIDO	ALL

GSE64718	GPL6887	33	MONOTEJIDO	ALL
GSE62612	GPL6887	11	MULTITEJIDO	'GSM1530015', 'GSM1530016', 'GSM1530017', 'GSM1530018', 'GSM1530019', 'GSM1530020', 'GSM1530021', 'GSM1530022', 'GSM1530023', 'GSM1530024', 'GSM1530025'
GSE62937	GPL6887	24	MONOTEJIDO	ALL
GSE55057	GPL6887	12	MONOTEJIDO	ALL
GSE39549	GPL6887	40	MULTITEJIDO	'GSM971546', 'GSM971547', 'GSM971548', 'GSM971549', 'GSM971550', 'GSM971551', 'GSM971552', 'GSM971553', 'GSM971554', 'GSM971555', 'GSM971556', 'GSM971557', 'GSM971558', 'GSM971559', 'GSM971560', 'GSM971561', 'GSM971562', 'GSM971563', 'GSM971564', 'GSM971565', 'GSM971566', 'GSM971567', 'GSM971568', 'GSM971569', 'GSM971570', 'GSM971571', 'GSM971572', 'GSM971573', 'GSM971574', 'GSM971575', 'GSM971576', 'GSM971577', 'GSM971578', 'GSM971579', 'GSM971580', 'GSM971581', 'GSM971582', 'GSM971583', 'GSM971584', 'GSM971585'
GSE50647	GPL6887	42	MONOTEJIDO	ALL
GSE33684	GPL6885	16	MONOTEJIDO	ALL
GSE113808	GPL6885	16	MONOTEJIDO	ALL
GSE104128	GPL6885	16	MULTITEJIDO	'GSM2790383', 'GSM2790384', 'GSM2790385', 'GSM2790386', 'GSM2790387', 'GSM2790388', 'GSM2790389', 'GSM2790390', 'GSM2790391', 'GSM2790392', 'GSM2790393', 'GSM2790394', 'GSM2790395', 'GSM2790396', 'GSM2790397', 'GSM2790398'
GSE97910	GPL6885	72	MONOTEJIDO	ALL

GSE70857	GPL6885	48	MULTITEJIDO	'GSM1820563', 'GSM1820564', 'GSM1820565', 'GSM1820566', 'GSM1820567', 'GSM1820568', 'GSM1820569', 'GSM1820570', 'GSM1820571', 'GSM1820572', 'GSM1820573', 'GSM1820574', 'GSM1820587', 'GSM1820588', 'GSM1820589', 'GSM1820590', 'GSM1820591', 'GSM1820592', 'GSM1820593', 'GSM1820594', 'GSM1820595', 'GSM1820596', 'GSM1820597', 'GSM1820598', 'GSM1820611', 'GSM1820612', 'GSM1820613', 'GSM1820614', 'GSM1820615', 'GSM1820616', 'GSM1820617', 'GSM1820618', 'GSM1820619', 'GSM1820620', 'GSM1820621', 'GSM1820622', 'GSM1820635', 'GSM1820636', 'GSM1820637', 'GSM1820638', 'GSM1820639', 'GSM1820640', 'GSM1820641', 'GSM1820642', 'GSM1820643', 'GSM1820644', 'GSM1820645', 'GSM1820646'
GSE62324	GPL6885	12	MONOTEJIDO	ALL
GSE57659	GPL6885	179	MONOTEJIDO	ALL
GSE39313	GPL6885	16	MULTITEJIDO	'GSM960692', 'GSM960693', 'GSM960694', 'GSM960695', 'GSM960696', 'GSM960697', 'GSM960698', 'GSM960699', 'GSM960700', 'GSM960701', 'GSM960718', 'GSM960719', 'GSM960720', 'GSM960721', 'GSM960722', 'GSM960723'
GSE118876	GPL16570	12	MONOTEJIDO	ALL
GSE104955	GPL16570	16	MONOTEJIDO	ALL
GSE87661	GPL16570	24	MONOTEJIDO	ALL
GSE79711	GPL16570	12	MONOTEJIDO	ALL
GSE60150	GPL16570	37	MONOTEJIDO	ALL

Tabla B.9: Resultados por plataforma de *Homo sapiens*.

PLATAFORMA: GPL570								
gene	general ranking	median CV	gene	general ranking	median IQR/m	gene	general ranking	median MAD/m
PPIA	47	0,0257	UBC	77	0,0222	PPIA	36	0,0270
RPL19	57	0,0272	RPL19	81	0,0224	UBC	75	0,0323
UBC	92	0,0302	PPIA	106	0,0234	RPL19	82	0,0329
GAPDH	1033	0,0457	GAPDH	1187	0,0369	GAPDH	1069	0,0533
HPRT1	6837	0,0778	HPRT1	9411	0,0719	RNA18S	10750	0,1134
RNA18S	7336	0,0801	RNA18S	10591	0,0762	HPRT1	12586	0,1233

PLATAFORMA: GPL6244								
gene	general ranking	median CV	gene	general ranking	median IQR/m	gene	general ranking	median MAD/m
UBC	22	0,0170	UBC	5	0,0098	UBC	5	0,0152
PPIA	30	0,0175	PPIA	71	0,0170	PPIA	109	0,0268
RPL19	489	0,0236	RPL19	855	0,0229	RPL19	267	0,0297
GAPDH	741	0,0247	GAPDH	2843	0,0277	GAPDH	3225	0,0425
RNA18S5	12986	0,0491	RNA18S5	7780	0,0359	HPRT1	6818	0,0515
HPRT1	15240	0,0559	HPRT1	13153	0,0463	RNA18S5	9267	0,0575

PLATAFORMA: GPL10558								
gene	general ranking	median CV	gene	general ranking	median IQR/m	gene	general ranking	median IQR/m
RPL19	3455	0,0312	PPIA	5299	0,0322	PPIA	5941	0,0488
UBC	5266	0,0343	GAPDH	7256	0,0345	GAPDH	7822	0,0519
PPIA	9866	0,0390	UBC	12155	0,0388	RPL19	20197	0,0692
GAPDH	13201	0,0417	RPL19	19447	0,0458	UBC	20325	0,0695
HPRT1	27186	0,0671	HPRT1	27325	0,0655	HPRT1	27084	0,0951
RNA18SN5	29782	0,0858	RNA18SN5	29850	0,0834	RNA18SN5	28824	0,1082

PLATAFORMA: GPL6947								
gene	general ranking	median CV	gene	general ranking	median IQR/m	gene	general ranking	median MAD/m
RPL19	242	0,0941	RPL19	327	0,0865	UBC	803	0,1446
UBC	422	0,0999	UBC	1288	0,1030	RPL19	1144	0,1513
HPRT1	8766	0,1584	HPRT1	8729	0,1445	HPRT1	6997	0,2039
PPIA	18601	0,1998	GAPDH	17617	0,1793	GAPDH	18894	0,2728
GAPDH	19330	0,2042	PPIA	20218	0,1924	PPIA	21285	0,2918

Tabla B.10: Resultados por plataforma de Mujeres.

PLATAFORMA: GPL570								
gene	general ranking	median CV	gene	general ranking	median IQR/m	gene	general ranking	median MAD/m
RPL19	15	0,0315	UBC	60	0,0206	UBC	51	0,0302
PPIA	137	0,0502	PPIA	69	0,0217	RPL19	72	0,0332
UBC	615	0,0646	RPL19	94	0,0238	PPIA	79	0,0338
GAPDH	1877	0,0822	GAPDH	1257	0,0479	GAPDH	1101	0,0688
HPRT1	7730	0,1234	HPRT1	7805	0,0987	HPRT1	8404	0,1533
RNA18S	22756	0,4176	RNA18S	22869	0,6204	RNA18S	22858	0,6382

PLATAFORMA: GPL6244								
gene	general ranking	median CV	gene	general ranking	median IQR/m	gene	general ranking	median MAD/m
UBC	3760	0,0808	GAPDH	3136	0,0604	UBC	3359	0,0945
HPRT1	4698	0,0889	UBC	3525	0,0646	GAPDH	3864	0,1017
PPIA	5708	0,0951	PPIA	5394	0,0796	PPIA	6567	0,1306
GAPDH	6269	0,0984	HPRT1	7480	0,0919	RPL19	7436	0,1377
RPL19	10365	0,1190	RNA18S5	9065	0,0999	HPRT1	8021	0,1419
RNA18S5	13294	0,1323	RPL19	13393	0,1217	RNA18S5	9073	0,1499

PLATAFORMA: GPL10558								
gene	general ranking	median CV	gene	general ranking	median IQR/m	gene	general ranking	median MAD/m
HPRT1	3647	0,0678	RPL19	2115	0,0459	RPL19	35	0,0118
RPL19	4860	0,0736	HPRT1	7055	0,0667	RNA18SN5	11138	0,1023
RNA18SN5	9537	0,0873	RNA18SN5	14374	0,0834	HPRT1	13510	0,1114
UBC	24675	0,1499	GAPDH	26056	0,1322	GAPDH	26814	0,1998
PPIA	27340	0,1778	PPIA	27914	0,1505	PPIA	28235	0,2263
GAPDH	29604	0,2195	UBC	28103	0,1527	UBC	28893	0,2413

PLATAFORMA: GPL6947								
gene	general ranking	median CV	gene	general ranking	median IQR/m	gene	general ranking	median MAD/m
RPL19	2412	0,1191	RPL19	3344	0,1115	UBC	3160	0,1643
UBC	9355	0,1537	UBC	5373	0,1225	RPL19	3613	0,1683
HPRT1	9885	0,1556	HPRT1	9737	0,1402	HPRT1	8824	0,2037
PPIA	19941	0,2069	GAPDH	20095	0,1952	GAPDH	20045	0,2860
GAPDH	20854	0,2169	PPIA	21233	0,2094	PPIA	21989	0,3231

Tabla B.11: Resultados por plataforma de Hombres.

PLATAFORMA: GPL570								
gene	general ranking	median CV	gene	general ranking	median IQR/m	gene	general ranking	median MAD/m
UBC	39	0,0184	UBC	116	0,0156	UBC	213	0,0238
RPL19	368	0,0288	RPL19	777	0,0246	RPL19	853	0,0354
PPIA	502	0,0310	PPIA	802	0,0249	PPIA	1314	0,0407
GAPDH	3188	0,0540	GAPDH	2099	0,0336	GAPDH	2408	0,0511
RNA18S	9819	0,0897	HPRT1	15922	0,0994	HPRT1	11911	0,1172
HPRT1	15211	0,1209	RNA18S	16740	0,1050	RNA18S	15895	0,1524

PLATAFORMA: GPL6244								
gene	general ranking	median CV	gene	general ranking	median IQR/m	gene	general ranking	median MAD/m
PPIA	41	0,0123	GAPDH	69	0,0088	GAPDH	115	0,0141
GAPDH	161	0,0145	PPIA	417	0,0119	PPIA	724	0,0197
UBC	338	0,0159	UBC	889	0,0137	UBC	1738	0,0236
RPL19	6034	0,0262	HPRT1	5182	0,0208	RPL19	3761	0,0284
RNA18S5	14848	0,0366	RPL19	5447	0,0211	HPRT1	4473	0,0297
HPRT1	15546	0,0379	RNA18S5	5846	0,0216	RNA18S5	7779	0,0355

PLATAFORMA: GPL10558								
gene	general ranking	median CV	gene	general ranking	median IQR/m	gene	general ranking	median MAD/m
HPRT1	1237	0,0737	HPRT1	981	0,0622	HPRT1	792	0,0890
RNA18SN5	2411	0,0834	RNA18SN5	2914	0,0768	RNA18SN5	3109	0,1151
GAPDH	9553	0,1314	GAPDH	5465	0,0917	GAPDH	5504	0,1365
UBC	19315	0,1721	RPL19	15909	0,1402	RPL19	10462	0,1754
RPL19	21429	0,1815	UBC	22995	0,1718	UBC	25188	0,2752
PPIA	26783	0,2163	PPIA	25535	0,1870	PPIA	26474	0,2889

PLATAFORMA: GPL6947								
gene	general ranking	median CV	gene	general ranking	median IQR/m	gene	general ranking	median MAD/m
UBC	6	0,0346	UBC	28	0,0386	UBC	19	0,0542
RPL19	75	0,0473	RPL19	39	0,0413	RPL19	49	0,0632
GAPDH	3778	0,0949	HPRT1	3115	0,0861	HPRT1	2785	0,1225
HPRT1	4169	0,0975	GAPDH	4242	0,0944	GAPDH	4061	0,1361
PPIA	5930	0,1083	PPIA	5694	0,1033	PPIA	6212	0,1553

Tabla B.12: Resultados por plataforma de *Mus musculus*.

PLATAFORMA: GPL1261								
gene	**general ranking**	**median CV**	**gene**	**general ranking**	**median IQR/m**	**gene**	**general ranking**	**median MAD/m**
Ppia	8	0,0269	Ppia	16	0,0235	Ppia	5	0,0264
Ubc	65	0,0354	Ubc	703	0,0387	Ubc	83	0,0435
Rpl19	1462	0,0519	Hprt	3672	0,0510	Rpl19	2756	0,0720
Hprt	4842	0,0643	Rpl19	4313	0,0528	Hprt	7977	0,0935
Gapdh	15569	0,0992	Gapdh	9573	0,0670	Gapdh	10502	0,1040
Rn18s	21399	0,2711	Rn18s	21077	0,1841	Rn18s	19862	0,1901

PLATAFORMA: GPL6246								
gene	**general ranking**	**median CV**	**gene**	**general ranking**	**median IQR/m**	**gene**	**general ranking**	**median MAD/m**
Rn18s	37	0,0072	Ppia	69	0,0070	Rn18s	9	0,0081
Ppia	90	0,0082	Rn18s	72	0,0070	Ppia	116	0,0114
Rpl19	131	0,0088	Rpl19	393	0,0092	Rpl19	510	0,0148
Ubc	631	0,0115	Ubc	925	0,0111	Ubc	1552	0,0187
Gapdh	1063	0,0127	Gapdh	2682	0,0141	Gapdh	2785	0,0217
Hprt	3170	0,0162	Hprt	4659	0,0166	Hprt	3439	0,0230

PLATAFORMA: GPL6885								
gene	**general ranking**	**median CV**	**gene**	**general ranking**	**median IQR/m**	**gene**	**general ranking**	**median MAD/m**
Ppia	11168	0,0897	Ubc	11448	0,0764	Rpl19	10523	0,1044
Ubc	11206	0,0900	Ppia	12102	0,0806	Ubc	10923	0,1078
Gapdh	13824	0,1148	Rpl19	13251	0,0896	Ppia	11951	0,1171
Rpl19	15954	0,1476	Gapdh	14305	0,0994	Gapdh	14992	0,1573
Hprt	17234	0,1969	Hprt	16680	0,1383	Hprt	17121	0,2228

PLATAFORMA: GPL6887								
gene	**general ranking**	**median CV**	**gene**	**general ranking**	**median IQR/m**	**gene**	**general ranking**	**median MAD/m**
Ppia	606	0,0100	Ppia	642	0,0091	Ppia	483	0,0130
Rpl19	21642	0,0217	Rpl19	24126	0,0223	Rpl19	22655	0,0307
Gapdh	24457	0,0254	Gapdh	25152	0,0238	Gapdh	25178	0,0350
Rn18s	24553	0,0255	Hprt	26153	0,0254	Hprt	26637	0,0387
Hprt	28941	0,0398	Rn18s	26851	0,0268	Rn18s	28247	0,0448

PLATAFORMA: GPL16570								
gene	**general ranking**	**median CV**	**gene**	**general ranking**	**median IQR/m**	**gene**	**general ranking**	**median MAD/m**
Ppia	25	0,0084	Ppia	17	0,0071	Ppia	29	0,0117
Gapdh	969	0,0157	Ubc	2171	0,0169	Ubc	2803	0,0271
Rpl19	3182	0,0208	Gapdh	4052	0,0202	Gapdh	2817	0,0271
Ubc	3377	0,0211	Hprt	5115	0,0217	Hprt	6232	0,0348
Hprt	8069	0,0282	Rpl19	6339	0,0232	Rpl19	9450	0,0408

Bibliografía

[1] Albert. Lehninger, David L. Nelson, and Michael M. Cox. *Principles of Biochemistry Lehninger Sixth editionk*. 2014. ISBN: 9781429234146.

[2] J. D. Watson and F. H. Crick. "The structure of DNA." *Cold Spring Harbor symposia on quantitative biology* 18 (1953), pp. 123–131. ISSN: 00917451. DOI: 10.1101/SQB.1953.018.01.020.

[3] Francis Crick. "Central dogma of molecular biology". *Nature* 227.5258 (1970), pp. 561–563. ISSN: 00280836. DOI: 10.1038/227561a0.

[4] Olena Morozova, Martin Hirst, and Marco A. Marra. "Applications of new sequencing technologies for transcriptome analysis". *Annual Review of Genomics and Human Genetics* 10 (2009), pp. 135–151. ISSN: 15278204. DOI: 10.1146/annurev-genom-082908-145957.

[5] John S. Mattick and Igor V. Makunin. "Non-coding RNA." 15 Spec No.suppl_1 (2006). ISSN: 09646906. DOI: 10.1093/hmg/ddl046.

[6] Satya P. Yadav. "The wholeness in suffix -omics, -omes, and the word om". *Journal of Biomolecular Techniques* 18.5 (2007), p. 277. ISSN: 15240215.

[7] Jun Zhang et al. "The impact of next-generation sequencing on genomics". 38.3 (2011), pp. 95–109. ISSN: 16738527. DOI: 10.1016/j.jgg.2011.02.003.

[8] Zhong Wang, Mark Gerstein, and Michael Snyder. "RNA-Seq: A revolutionary tool for transcriptomics". 10.1 (2009), pp. 57–63. ISSN: 14710056. DOI: 10.1038/nrg2484.

[9] Sunitha Kogenaru et al. "RNA-seq and microarray complement each other in transcriptome profiling". *BMC Genomics* 13.1 (2012). ISSN: 14712164. DOI: 10.1186/1471-2164-13-629.

[10] Andrew Watson et al. "Technology for microarray analysis of gene expression". *Current opinion in biotechnology* 9.6 (1998), pp. 609–614.

[11] Antonio Doménech-Sánchez and Jordi Vila. "Fundamento, tipos y aplicaciones de los arrays de ADN en la microbiología médica". *Enfermedades Infecciosas y Microbiologia Clinica* 22.1 (2004), pp. 46–54. ISSN: 0213005X. DOI: 10.1157/13056692.

[12] Mark Schena. "Genome analysis with gene expression microarrays". *BioEssays* 18.5 (1996), pp. 427–431. ISSN: 0265-9247. DOI: 10.1002/bies.950180513.

[13] Amandeep Singh and Naresh Kumar. "A review on DNA microarray technology". 22 (2013), p. 5.

[14] Paolo Tieri et al. "Bioinformatics for Omics Data". *OMICS: A Journal of Integrative Biology* 719 (2011). DOI: 10.1007/978-1-61779-027-0.

[15] John H. Malone and Brian Oliver. "Microarrays, deep sequencing and the true measure of the transcriptome". 9.1 (2011), p. 34. ISSN: 17417007. DOI: 10.1186/1741-7007-9-34.

[16] Brian T. Wilhelm and Josette Renée Landry. "RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing". 48.3 (2009), pp. 249–257. ISSN: 10462023. DOI: 10.1016/j.ymeth.2009.03.016.

[17] Ana Conesa et al. "A survey of best practices for RNA-seq data analysis". 17.1 (2016), pp. 1–19. ISSN: 1474760X. DOI: 10.1186/s13059-016-0881-8.

[18] Claudia Manzoni et al. "Genome, transcriptome and proteome: The rise of omics data and their integration in biomedical sciences". *Briefings in Bioinformatics* 19.2 (2018), pp. 286–302. ISSN: 14774054. DOI: 10.1093/BIB/BBW114.

[19] Ochuko Orakpoghenor. "Diagnostic Techniques in Molecular Biology-An Overview". *MedRead Journal of Family Medicine* (2020).

[20] A. J. Butte, V. J. Dzau, and S. B. Glueck. "Further defining housekeeping, or "maintenance," genes Focus on "A compendium of gene expression in normal human tissues"." 7.2 (2001), pp. 95–96. ISSN: 15312267. DOI: 10.1152/physiolgenomics.2001.7.2.95.

[21] Cheng-Wei Chang et al. "Identification of Human Housekeeping Genes and Tissue-Selective Genes by Microarray Meta-Analysis". *PLoS ONE* 6.7 (2011). ISSN: 1932-6203. DOI: 10.1371/journal.pone.0022859.

[22] O Thellin et al. "Housekeeping genes as internal standards: use and limits". 75 (1999), pp. 291–295.

[23] Jiang Zhu et al. "On the nature of human housekeeping genes". 24.10 (2008), pp. 481–484. ISSN: 01689525. DOI: 10.1016/j.tig.2008.08.004.

[24] Eli Eisenberg and Erez Y. Levanon. "Human housekeeping genes, revisited". *Trends in Genetics* 29.10 (2013). Human Genetics, pp. 569–574. ISSN: 0168-9525. DOI: https://doi.org/10.1016/j.tig.2013.05.010.

[25] Stephen R. Stürzenbaum and Peter Kille. "Control genes in quantitative molecular biological techniques: The variability of invariance". 130.3 (2001), pp. 281–289. ISSN: 10964959. DOI: 10.1016/S1096-4959(01)00440-7.

[26] Keertan Dheda et al. "Validation of housekeeping genes for normalizing RNA expression in real-time PCR". *BioTechniques* 37.1 (2004), pp. 112–119. ISSN: 07366205. DOI: 10.2144/04371rr03.

[27] T. Suzuki, P. J. Higgins, and D. R. Crawford. "Control selection for RNA quantitation". 29.2 (2000), pp. 332–337. ISSN: 07366205. DOI: 10.2144/00292rv02.

[28] Peter D. Lee et al. "Control genes and variability: Absence of ubiquitous reference transcripts in diverse mammalian expression studies". *Genome Research* 12.2 (2002), pp. 292–297. ISSN: 10889051. DOI: 10.1101/gr.217802.

[29] Li Li Hsiao et al. "A compendium of gene expression in normal human tissues". *Physiological Genomics* 2002.7 (2002), pp. 97–104. ISSN: 15312267. DOI: 10.1152/physiolgenomics.00040.2001.

[30] Seram Lee et al. "Identification of novel universal housekeeping genes by statistical analysis of microarray data". *Journal of Biochemistry and Molecular Biology* 40.2 (2007), pp. 226–231. ISSN: 12258687. DOI: 10.5483/bmbrep.2007.40.2.226.

[31] Hendrik J.M. de Jonge et al. "Evidence based selection of housekeeping genes". *PLoS ONE* 2.9 (2007). ISSN: 19326203. DOI: 10.1371/journal.pone.0000898.

[32] Xinwei She et al. "Definition, conservation and epigenetics of housekeeping and tissue-enriched genes". *BMC Genomics* 10 (2009). ISSN: 14712164. DOI: 10.1186/1471-2164-10-269.

[33] Jiang Zhu et al. "How many human genes can be defined as housekeeping with current expression data?" *BMC Genomics* 9.1 (2008), p. 172. ISSN: 14712164. DOI: 10.1186/1471-2164-9-172.

[34] Yijuan Zhang, Ding Li, and Bingyun Sun. "Do Housekeeping Genes Exist?" *PLOS ONE* 10.5 (2015). ISSN: 1932-6203. DOI: 10.1371/journal.pone.0123691.

[35] Jo Vandesompele et al. "Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes" (2002).

[36] Michael W. Pfaffl et al. "Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper - Excel-based tool using pair-wise correlations". *Biotechnology Letters* 26.6 (2004), pp. 509–515. ISSN: 01415492. DOI: 10.1023/B:BILE.0000019559.84305.47.

[37] Fuliang Xie et al. "miRDeepFinder: A miRNA analysis tool for deep sequencing of plant small RNAs". *Plant Molecular Biology* 80.1 (2012), pp. 75–84. ISSN: 01674412. DOI: 10.1007/s11103-012-9885-2.

[38] Claus Lindbjerg Andersen, Jens Ledet Jensen, and Torben Falck Ørntoft. "Normalization of Real-Time Quantitative Reverse Transcription-PCR Data: A Model-Based Variance Estimation Approach to Identify Genes Suited for Normalization, Applied to Bladder and Colon Cancer Data Sets". 64 (2004), pp. 5245–5250.

[39] Maria V. Schneider and Sandra Orchard. "Omics technologies, data and bioinformatics principles." 719 (2011), pp. 3–30. ISSN: 19406029. DOI: 10.1007/978-1-61779-027-0_1.

[40] Daniel J. Rigden and Xosé M. Fernández. "The 27th annual Nucleic Acids Research database issue and molecular biology database collection". *Nucleic Acids Research* 48.D1 (2020), pp. 1–8. ISSN: 13624962. DOI: 10.1093/nar/gkz1161.

[41] H. Parkinson et al. "ArrayExpress - A public database of microarray experiments and gene expression profiles". *Nucleic Acids Research* 35.SUPPL. 1 (2007). ISSN: 03051048. DOI: 10.1093/nar/gkl995.

[42] Tanya Barrett et al. "NCBI GEO: archive for functional genomics data sets—update". *Nucleic Acids Research* 41.D1 (2012), pp. 991–995. ISSN: 0305-1048. DOI: 10.1093/nar/gks1193.

[43] Ron Edgar, Michael Domrachev, and Alex E Lash. "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository". *Nucleic Acids Research* 30.1 (2002), pp. 207–210. ISSN: 0305-1048. DOI: 10.1093/nar/30.1.207.

[44] Emily Clough and Tanya Barrett. "The Gene Expression Omnibus database". 1418 (2016), pp. 93–110. ISSN: 10643745. DOI: 10.1007/978-1-4939-3578-9_5.

[45] Alvis Brazma et al. "Minimum information about a microarray experiment (MIAME)-toward standards for microarray data" (2001).

[46] "Microarray standards at last". *Nature* 419.6905 (2002), p. 323. DOI: 10.1038/419323a.

[47] Mark D. Wilkinson et al. "Comment: The FAIR Guiding Principles for scientific data management and stewardship". *Scientific Data* 3 (2016). ISSN: 20524463. DOI: 10.1038/sdata.2016.18.

[48] Yuelin Zhu et al. "GEOmetadb: powerful alternative search engine for the Gene Expression Omnibus". *Bioinformatics* 24.23 (2008), pp. 2798–2800. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btn520.

[49] Sean Davis and Paul S Meltzer. "GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor". *Bioinformatics* 23.14 (2007), pp. 1846–1847. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btm254.

[50] Zichen Wang, Alexander Lachmann, and Avi Ma'ayan. "Mining data and metadata from the gene expression omnibus". 11.1 (2019), pp. 103–110. ISSN: 18672469. DOI: 10.1007/s12551-018-0490-8.

[51] Sahin Naqvi et al. "Conservation, acquisition, and functional impact of sex-biased gene expression in mammals". *Science* 365.6450 (2019). ISSN: 10959203. DOI: 10.1126/science.aaw7317.

[52] Marianne J. Legato, Paula A. Johnson, and Joann E. Manson. "Consideration of sex differences in medicine to improve health care and patient outcomes". 316.18 (2016), pp. 1865–1866. ISSN: 15383598. DOI: 10.1001/jama.2016.13995.

[53] Nicole C Woitowich, Annaliese Beery, and Teresa Woodruff. "A 10-year follow-up study of sex inclusion in the biological sciences". *eLife* 9 (2020). ISSN: 2050-084X. DOI: 10.7554/eLife.56344.

[54] Vlad Popovici et al. "Selecting control genes for RT-QPCR using public microarray data". *BMC Bioinformatics* 10.1 (2009), p. 42. ISSN: 14712105. DOI: 10.1186/1471-2105-10-42.

[55] R Core Team. "R: a language and environment for statistical computing" (2013).

[56] Alessandro Liberati et al. "The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration". 62.10 (2009). ISSN: 18785921. DOI: 10.1016/j.jclinepi.2009.06.006.

[57] David Moher et al. "Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement". *PLoS Medicine* 6.7 (2009). ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1000097.

[58] Robert C. Gentleman et al. "Bioconductor: open software development for computational biology and bioinformatics." *Genome biology* 5.10 (2004). ISSN: 14656914. DOI: 10.1186/gb-2004-5-10-r80.

[59] Chandima N P G Arachchige, Luke A Prendergast, and Robert G Staudte. "Robust analogues to the Coefficient of Variation" (2019).

[60] Rainer Breitling et al. "Rank products: A simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments". *FEBS Letters* 573.1-3 (2004), pp. 83–92. ISSN: 00145793. DOI: 10.1016/j.febslet.2004.07.055.

[61] Fangxin Hong et al. "RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis". *Bioinformatics* 22.22 (2006), pp. 2825–2827. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btl476.

[62] Francesco Del Carratore et al. "RankProd 2.0: a refactored bioconductor package for detecting differentially expressed features in molecular profiling datasets". *Bioinformatics* 33.17 (2017), pp. 2774–2775. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btx292.

[63] Lawrence Mitchell. "A parallel implementation of the Rank Product method for R" (2011).

[64] Gil Stelzer et al. "The GeneCards suite: From gene data mining to disease genome sequence analyses". *Current Protocols in Bioinformatics* 2016.1 (2016). ISSN: 1934340X. DOI: 10.1002/cpbi.5.

[65] Anna P Pilbrow et al. "Genomic selection of reference genes for real-time PCR in human myocardium". *BMC Medical Genomics* 1.1 (2008), p. 64. ISSN: 1755-8794. DOI: 10.1186/1755-8794-1-64.