



Universidad Nacional de Rosario
Facultad de Ciencias Económicas y Estadística

Carrera de posgrado
Maestría en Estadística Aplicada

Estudio y comparación de métodos estadísticos utilizados
para combinar resultados provenientes de análisis
transcripcionales de microarreglos, para cantidades de
experimentos y tamaños muestrales pequeños

Maestranda: Lic. Laura Piskulic
Director: Dr. Lucas D. Daurelio
Co-directora: Mg. Guillermina B. Harvey

Año 2026

Agradecimientos

Porque toda gran aventura empieza por alguien que cree en ti quisiera comenzar agradeciendo a quienes me dirigieron en esta tesis, a Lucas Daurelio y Guillermina Harvey. También quisiera incluir aquí a Marcos Prunello quien estuvo presente en los inicios de esta maravillosa aventura. Gracias por creer en mí, por su guía, sus consejos y paciencia infinita.

Un reconocimiento especial a todos los que hicieron posible esta Maestría. Gracias al equipo docente por el amor y generosidad con la que compartieron su conocimiento y experiencias haciendo de cada encuentro un momento único e irrepetible. Gracias a mis compañeros de cursado, por los mates, por la contención en los momentos de estrés y por las horas de estudio compartidas que hicieron de este camino una aventura mucho más gratificante.

Quiero agradecer también a todos mis compañeros del Área Estadística de la Facultad de Bioquímica y Farmacia, a los jubilados y a los actuales. A los primeros por abrirme las puertas al maravilloso mundo de la estadística y por impulsarme a ir por más. A los actuales por sus palabras de aliento y por acompañar en todo momento.

Finalmente quiero agradecer infinitamente a mi familia, a mi mamá, mi papá, mi hermana y en especial a Cristián, Mirko y Nuria por motivarme a seguir aún en los momentos más complicados de este recorrido. Su amor y apoyo incondicional fueron y son el motor que me impulsa a seguir adelante y superar los desafíos.

A todos y cada uno, gracias, gracias, gracias por acompañarme en esta hermosa aventura.

Resumen

En muchas ocasiones distintos grupos de investigación estudian fenómenos biológicos en organismos de la misma especie producidos por variaciones similares en las condiciones experimentales. La metodología estadística denominada metaanálisis permite combinar los resultados provenientes de estas diferentes investigaciones permitiendo, entre otras cosas, generalizar conclusiones. Existen múltiples técnicas que permiten combinar estos resultados. La elección de la más apropiada depende del tipo de estudio que se lleva a cabo, así como también de las características de los experimentos y modo en que se presentan los resultados. En el presente trabajo se describen y ponen a prueba una selección de técnicas de metaanálisis para combinar resultados provenientes de experimentos de expresión génica obtenidas por la técnica de microarreglos en situaciones donde tanto la cantidad de estudios a incluir como de réplicas en cada uno de ellos es pequeña. Las técnicas seleccionadas fueron la de “mínimo valor p ”, la de “máximo valor p ”, la técnica de “valor p de orden r ”, el método de Fisher, el de Stouffer, el método basado en la estadística de Fisher con pesos ponderados, la técnica basada en “suma de rangos” y la basada en “producto de rangos”. Luego, con la utilización de datos simulados en distintos escenarios producto de la combinación de distinto número de estudios (entre 2 y 7) y distinta cantidad de réplicas (entre 2 y 5) se evaluó el desempeño de cada una de ellas para tres configuraciones de hipótesis de investigación diferentes. A medida que aumenta la cantidad de información, reflejado en un aumento de la cantidad de réplicas biológicas por experimentos y/o en la cantidad de estudios a combinar, se observó un mejor desempeño de todas las técnicas de metaanálisis estudiadas, aunque se obtuvieron diferencias entre las mismas. Es importante notar que la metodología que presentó mejor desempeño en casi todos los escenarios y bajo las tres configuraciones de hipótesis estudiadas fue la técnica de Fisher. Por otro lado, la técnica de “máximo valor p ” aunque presentó valores de desempeño acordes como para ser considerada una posible alternativa, es la que presentó peor desempeño en la mayoría de las situaciones por lo que sería la menos recomendable para los casos estudiados.

Palabras clave: experimentos de microarreglos, metaanálisis de expresión génica, pequeña cantidad de estudios, número de réplicas reducido.

Índice general

Agradecimientos	i
Resumen	ii
Índice de gráficos	v
Índice de tablas	vi
Glosario de abreviaturas	vii
1. Introducción	1
2. Objetivos	8
3. Metodología	9
3.1. Consideraciones inherentes a las técnicas de MA	9
3.1.1. La configuración de las hipótesis nula y alternativa detrás del objetivo del MA	9
3.1.2. Datos disponibles de los estudios a incluir en el MA	10
3.1.3. Diferenciar o no entre la sub-expresión y sobre-expresión de los genes DE	11
3.2. Técnicas de MA	12
3.2.1. Mínimo valor p	13
3.2.2. Máximo valor p	13
3.2.3. Valor p de orden r	13
3.2.4. Estadística de Fisher	14
3.2.5. Estadística de Stouffer	14
3.2.6. Estadística de Fisher con pesos adaptados	15
3.2.7. Suma de rangos	15
3.2.8. Producto de rangos	16
3.3. Generación de datos	17
3.4. Evaluación del desempeño de las técnicas de MA	18
3.5. Soporte Computacional	20
4. Resultados	21
4.1. Desempeño de las técnicas de MA cuando se combinan dos estudios	21
4.2. Desempeño de las técnicas de MA cuando se combinan tres estudios	24
4.3. Desempeño de las técnicas de MA cuando se combinan cuatro estudios	27
4.4. Desempeño de las técnicas de MA cuando se combinan cinco estudios	29

4.5. Desempeño de las técnicas de MA cuando se combinan seis estudios	31
4.6. Desempeño de las técnicas de MA cuando se combinan siete estudios	34
5. Consideraciones Finales	38
6. Bibliografía	41
Anexo A	44
Anexo B	47

Índice de gráficos

Gráfico 1: Valores del AUC en función de la cantidad de réplicas (r) según técnica de MA utilizada y para cada una de las configuraciones de hipótesis, 23 cuando se combinan dos estudios ($k=2$)

Gráfico 2: Valores del AUC en función de la cantidad de réplicas (r) según técnica de MA utilizada y para cada una de las configuraciones de hipótesis, 25 cuando se combinan tres estudios ($k=3$)

Gráfico 3: Valores del AUC en función de la cantidad de réplicas (r) según técnica de MA utilizada y para cada una de las configuraciones de hipótesis, 27 cuando se combinan cuatro estudios ($k=4$)

Gráfico 4: Valores del AUC en función de la cantidad de réplicas (r) según técnica de MA utilizada y para cada una de las configuraciones de hipótesis, 29 cuando se combinan cinco estudios ($k=5$)

Gráfico 5: Valores del AUC en función de la cantidad de réplicas (r) según técnica de MA utilizada y para cada una de las configuraciones de hipótesis, 32 cuando se combinan seis estudios ($k=6$)

Gráfico 6: Valores del AUC en función de la cantidad de réplicas (r) según técnica de MA utilizada y para cada una de las configuraciones de hipótesis, 35 cuando se combinan siete estudios ($k=7$)

Índice de tablas

Tabla 1: Valores del AUC y resultado del test de comparaciones múltiples entre las técnicas de MA para cada escenario presentado con $k=2$	22
Tabla 2: Valores del AUC y resultado del test de comparaciones múltiples entre las técnicas de MA para cada escenario presentado con $k=3$	26
Tabla 3: Valores del AUC y resultado del test de comparaciones múltiples entre las técnicas de MA para cada escenario presentado con $k=4$	28
Tabla 4: Valores del AUC y resultado del test de comparaciones múltiples entre las técnicas de MA para cada escenario presentado con $k=5$	30
Tabla 5: Valores del AUC y resultado del test de comparaciones múltiples entre las técnicas de MA para cada escenario presentado con $k=6$	33
Tabla 6: Valores del AUC y resultado del test de comparaciones múltiples entre las técnicas de MA para cada escenario presentado con $k=7$	36
Tabla B1: Valores del AUC y resultado del test de comparaciones entre configuraciones de hipótesis para cada técnica de MA cuando se combinan dos estudios	48
Tabla B2: Valores del AUC y resultado del test de comparaciones entre configuraciones de hipótesis para cada técnica de MA cuando se combinan tres estudios	49
Tabla B3: Valores del AUC y resultado del test de comparaciones entre configuraciones de hipótesis para cada técnica de MA cuando se combinan cuatro estudios	50
Tabla B4: Valores del AUC y resultado del test de comparaciones entre configuraciones de hipótesis para cada técnica de MA cuando se combinan cinco estudios	51
Tabla B5: Valores del AUC y resultado del test de comparaciones entre configuraciones de hipótesis para cada técnica de MA cuando se combinan seis estudios	52
Tabla B6: Valores del AUC y resultado del test de comparaciones entre configuraciones de hipótesis para cada técnica de MA cuando se combinan siete estudios	53

Glosario de abreviaturas

A: Adenina

ADN: Ácido desoxirribonucleico

ARNm: Ácido ribonucleico mensajero

AUC: Área debajo de la curva ROC (siglas que provienen del término en inglés *Area Under the Curve*).

AW: Estadística de Fisher con pesos adaptados (siglas que provienen del término en inglés *Adaptively Weighted*)

C: Citocina

CLD: método para visualizar el resultado de las comparaciones múltiples (siglas que provienen del término en inglés *Compact Letter Display*).

curva ROC: acrónimo de *Receiver Operating Characteristic*

DE: Diferencialmente Expresado

E: Especificidad de una prueba

FC: Tasa de cambio (por sus siglas en inglés *Fold Change*)

FDR: método que permite ajustar del valor p en comparaciones múltiple (siglas que provienen del término en inglés *False Discovery Rate*).

G: Guanina

HS_A: Configuración de hipótesis cuando el objetivo es identificar un gen como DE cuando el efecto de su expresión sea distinto de cero en todos los estudios combinados.

HS_B: Configuración de hipótesis cuando el objetivo es identificar un gen como DE cuando el efecto de su expresión sea distinto de cero en por lo menos uno menos h de los K estudios combinados.

HS_r: Configuración de hipótesis cuando el objetivo es identificar un gen como DE cuando el efecto de su expresión sea distinto de cero en por lo menos menos h de los K estudios combinados.

k: cantidad de estudios a combinar en un MA

MA: Metaanálisis

maxP: Máximo valor p

minP: Mínimo valor p

PR: Producto de rangos

r: cantidad de réplicas en un estudio individual de microarreglos

roP: Valor p de orden r

S: Sensibilidad de una prueba

SR: Suma de rangos

T: Timina

1. Introducción

Los seres vivos están compuestos por una o más células dónde guardan las instrucciones que determinan todas sus funciones y características particulares en su material genético, escrito en un lenguaje (código genético) que cualquier forma de vida puede leer. Para la mayoría de los organismos conocidos, el material genético es ácido desoxirribonucleico (ADN) y se encuentra en el interior de las células. El ADN es una molécula con forma de una escalera retorcida o doble hélice conectada. Los escalones de la escalera contienen un par de cuatro moléculas posibles denominadas adenina (A), timina (T), citosina (C) y guanina (G), las cuales son las letras del alfabeto genético. Cada A de una hebra se encuentra unida a una T de la otra y cada C con una G, por esto se dice que las hebras del ADN son complementarias. Las diferentes secuencias específicas de esas cuatro letras determinan las instrucciones de los seres vivos, y contienen el código de cómo crecer, moverse, percibir el entorno, curarse, reproducirse (Alberts *et al.*, 2023).

Cada característica particular está codificada en un segmento del ADN denominado gen. Los seres vivos cuentan con una maquinaria específica que, entre otras cosas, puede producir proteínas utilizando los genes como molde. Este procedimiento se denomina expresión de gen. En forma simplificada se puede decir que, para obtener una proteína, en un primer paso la maquinaria celular usa un determinado gen como molde para producir una molécula de ácido ribonucleico mensajero (ARNm) este proceso se denomina transcripción. Luego, usando esta última como molde obtiene la proteína, este procedimiento se denomina traducción (Alberts *et al.*, 2023).

El ADN contiene toda la información necesaria (todos los genes) para la síntesis de todas las proteínas. Sin embargo, no todos los genes se expresan en todas las células del organismo ni al mismo tiempo. Es por esto que en las células sólo estarán presentes las moléculas de ARNm correspondientes a los genes que se estén expresando en un momento determinado, esto quiere decir que hay genes que se expresan (inducen) o dejan de hacerlo (reprimen) en diferentes momentos de la vida de un organismo o ante un cambio en las condiciones del medio que lo rodea, cambio climático, presencia de un patógeno, entre otros. Analizar estos cambios de expresión de los genes es un área de mucha relevancia biológica, que se estudia en diversos sistemas y enfoques (Lodish *et al.*, 2021).

Los experimentos denominados de expresión génica tienen como objetivo detectar aquellos genes que se expresan diferencialmente (DE), es decir los que cambian su expresión en forma significativa, al comparar dos o más condiciones diferentes. Para lograrlo, utilizan diferentes técnicas que permiten medir la cantidad de las moléculas de ARNm de cada gen específico, abarcando desde uno a varios genes al mismo tiempo. El análisis de las expresiones de múltiples genes simultáneamente tiene la ventaja de aumentar la posibilidad de identificar genes que son inducidos o reprimidos en algún estado biológico concreto, incluso abarcando prácticamente todos los transcritos posibles. Particularmente, los experimentos de expresión génica basados en la tecnología de microarreglos, comúnmente denominados experimentos de microarreglos, son una metodología que permite medir la actividad transcripcional de varios miles de genes a la vez (Lodish *et al.*, 2021).

En los experimentos de análisis transcripcional se compara la intensidad de la señal producida por la cantidad presente de un transcrito, obtenida para cada gen proveniente de un tratamiento en comparación a la correspondiente del control u otro tratamiento. En el caso de los microarreglos la señal emitida es de tipo lumínica, y cada gen va a estar específicamente ubicado en un arreglo de puntos bidimensional. Como se analizan miles de genes al mismo tiempo, una particularidad que presenta este tipo de experimentos viene dada por la gran cantidad de información generada (Moreno y Solé, 2004). Por ejemplo, en un ensayo típico para comparar dos condiciones determinadas suele analizarse el cambio de más de 10.000 genes e incluso este número puede ascender a más de 20.000, valores que serán analizados como variables independientes. Es por esto que se han desarrollado herramientas bioinformáticas que permitan el rápido y correcto procesamiento de los resultados obtenidos. En este sentido, el paquete *Linear Models for Microarray* (LIMMA, Ritchie *et al.*, 2015) de R (R Core Team, 2024) es uno de los más utilizados. LIMMA es una librería para el análisis de datos de experimentos de microarreglos que tiene la capacidad de usar modelos lineales para evaluar la expresión diferencial en el contexto de experimentos diseñados multifactorialmente.

Los mencionados experimentos constan de múltiples pasos, de los cuales la última etapa consiste en un procesamiento de la señal lumínica obtenida para cada gen con el objetivo de asegurar, entre otras cosas, que las diferencias entre

las intensidades observadas sean realmente debidas a diferencias en la expresión y no a artificios propios de la técnica. En el contexto de experimentos para comparar dos condiciones (control y tratado) se trabaja con la intensidad obtenida bajo el tratamiento relativizada a la del control, dando origen a la tasa de cambio denominada FC (siglas que provienen del término en inglés *Fold Change*). El FC permite identificar aquellos genes que cambian su expresión, donde un valor de 1 estaría indicando que no hay diferencia en la expresión del gen entre los tratamientos comparados. Dado que los efectos sobre las intensidades en las señales de los experimentos de microarreglos tienden a ser multiplicativos generalmente se trabaja con el $\log_2$ FC. Esta transformación logarítmica convierte los efectos multiplicativos en efectos aditivos, en este caso un valor igual a cero estaría indicando que no hay diferencia de expresión entre ambos tratamientos (Cui y Churchill, 2003).

No se debe perder de vista la variabilidad inherente a todo experimento aleatorio por lo cual, una vez obtenidos los valores de FC, se realiza un test estadístico específico para este tipo de experimentos, para cada uno de los genes bajo estudio. En situaciones donde se comparan sólo dos tratamientos se utilizan métodos estadísticos basados en el test t, y cuando se trabaja con más de dos condiciones se utilizan test de análisis de la variancia (ANOVA). En la hipótesis nula se plantea que la cantidad de expresión del gen es la misma en ambas condiciones bajo estudio, mientras que la hipótesis alternativa plantea que se observa un cambio significativo en la expresión del mismo en un tratamiento en comparación con el otro, ya sea en forma positiva (sobre-expresión o inducción) o negativa (sub-expresión o represión) (Cui y Churchill, 2003).

Debido a los costos de los experimentos de microarreglos, en la mayoría de las situaciones se cuenta con un número limitado de réplicas biológicas, lo que lleva a un bajo poder de detección para la determinación de expresión diferencial de genes. Por otro lado, en muchas ocasiones distintos grupos de investigación estudian los cambios en el perfil de expresión génica en organismos de la misma especie producidos por variaciones similares en las condiciones experimentales. Como parte del proceso de publicación de los resultados obtenidos, estos grupos depositan sus datos en repositorios internacionales, dando lugar a la posibilidad de acceder fácilmente a los mismos. Sin embargo, los resultados se analizan en forma particular dentro de cada experimento sin comparar con aquellos

obtenidos en otros ensayos por otros investigadores, motivo por el cual se pierde una cantidad muy valiosa de información que se podría obtener de la combinación de todos los resultados. Una alternativa para hacer frente a esta situación surge de aplicar la metodología estadística denominada metaanálisis (MA) (Forero, 2019; Walsh *et al.*, 2015; Kontouet *al.*, 2018), que tiene por objetivo combinar los resultados provenientes de diferentes investigaciones, presentando la ventaja de aumentar la potencia para detectar genes DE y permitiendo generalizar conclusiones.

La metodología del MA involucra varios pasos (Ramasamy *et al.*, 2008; Kontou *et al.*, 2018):

- Definir claramente el objetivo de la investigación y buscar estudios publicados relacionados con este interés. Para esto se deben establecer sitios de búsqueda y especificar los criterios de inclusión/exclusión de los trabajos a considerar. Todo esto requiere de la experiencia y conocimientos específicos en el tema por lo que son tareas que conciernen en particular al investigador especialista en el tema.
- Extraer los datos de los estudios elegidos y realizar un pre-procesamiento de los mismos. En esta etapa se unifica la nomenclatura con la que se identifica cada gen, y se toman decisiones respecto a cómo trabajar en caso que haya genes considerados sólo en alguno de los estudios incluidos.
- Seleccionar las técnicas estadísticas más apropiadas para la combinación de los resultados de los estudios individuales en función del objetivo planteado por los investigadores para analizar, presentar e interpretar los resultados obtenidos.

El presente trabajo se centra en el estudio de las técnicas estadísticas para combinar los resultados provenientes de estudios individuales. En este sentido, en la bibliografía consultada se observan diferentes autores que abordaron esta temática, la mayoría con aplicaciones en medicina. Por ejemplo, entre las aplicaciones presentadas por Chang *et al.* (2013) se cuenta con una en donde se combinan 8 trabajos realizados sobre pacientes con cáncer de mama, cuyos tamaños muestrales varían entre 97 y 286 pacientes. Otra aplicación, también

presentada por los mencionados autores, consiste en la combinación de 7 trabajos sobre cáncer de cerebro, en estos se compara la expresión de genes entre pacientes tomados como control y los que padecen la enfermedad, y los tamaños muestrales considerados van de 29 a 175 pacientes.

Campaign y Yang (2010) también presentaron diferentes casos de estudio, en uno de ellos combinan resultados de tres trabajos donde se estudian dos diferentes tipos de linfoma, uno de estos estudios incluye información sobre 60 pacientes (37 con linfoma folicular y 23 con linfoma difuso de células B grandes), un segundo con 77 pacientes (19 y 58 respectivamente) y el tercero con 46 pacientes (19 y 27 respectivamente).

En este contexto se puede decir que los autores antes mencionados y otros como Tseng *et al.* (2012) y Makinde *et al.* (2021), se enfocaron en el estudio de MA de micorreglos en humanos utilizando distintas técnicas para situaciones en donde la cantidad de estudios incluidos y/o el tamaño muestral dentro de los mismos es considerado grande. Sin embargo, en experimentos donde se trabaja con otros sistemas biológicos como plantas, sometidas a distintas condiciones experimentales, la situación es diferente y el tamaño muestral suele ser bastante más pequeño. Por ejemplo, en un trabajo donde se estudia la respuesta de hojas de *Citrus sinensis* (naranja dulce) al patógeno *Xanthomonas campestris* pv. *vesicatoria*, Daurelio *et al.* (2013) realizan experimentos de microarreglos por triplicado utilizando plantas idénticas como réplicas biológicas.

En la actualidad, no se cuenta con suficiente información para decidir acerca de las mejores técnicas de MA a utilizar en casos donde se quiera combinar un número de estudios (k) en donde la cantidad de réplicas (r) evaluadas en cada uno de ellos son pequeños.

Según Tseng *et al.* (2012) hay diversas particularidades que deben ser tenidas en cuenta al momento de combinar los resultados provenientes de estudios individuales. A continuación se mencionan las tres principales que luego son desarrolladas en la sección Metodología.

- La configuración de las hipótesis nula y alternativa detrás del objetivo del MA. Esto generalmente depende de la situación biológica considerada por los investigadores. Cuando el objetivo es buscar, por

ejemplo, un marcador genético asociado a una determinada enfermedad, el MA se centra en buscar genes DE en todos los estudios incluidos en el mismo. En cambio, cuando el objetivo es estudiar genes relacionados, por ejemplo, a una ruta metabólica, los investigadores comenzarán por buscar aquellos genes que se identifiquen como DE en por lo menos uno de los trabajos incluidos en el MA para luego seguir su estudio.

- Los datos disponibles de los estudios a incluir en el MA. Esta información puede encontrarse como una lista de genes que resultaron DE, el valor p del respectivo análisis estadístico para cada gen incluido en el estudio, o incluso se puede contar con los datos crudos.
- Diferenciar o no entre la sub-expresión y sobre-expresión de los genes DE. Este punto suele depender no sólo del objetivo de los investigadores, sino también de los datos disponibles para llevar adelante el MA.

Una vez definidos estos aspectos quedan naturalmente seleccionadas un conjunto de técnicas de MA capaces de dar respuesta a las inquietudes del investigador. Pero todavía queda por definir cuál de todas ellas es la más apropiada.

Para evaluar el desempeño de métodos estadísticos se pueden utilizar datos reales y/o datos simulados. Cuando se utilizan datos reales se suelen comparar los resultados obtenidos al aplicar la técnica bajo estudio con los obtenidos por alguna otra técnica ya conocida y bien estudiada. Esta situación tiene como desventaja que los dos métodos están sujetos a error dado que no se conoce con certeza el comportamiento o naturaleza de los datos considerados. Cuando se utilizan datos simulados se consigue conocer (o controlar) las características más importantes de los datos. Se considera a las simulaciones como representaciones artificiales de la realidad puesto que se basan en modelos matemáticos para tratar de reproducir la realidad de un modo controlado, sujeto a los supuestos de los modelos. Dado que esto lleva en general a tener una descripción simplificada de la realidad, la desventaja radica justamente en que se trabaja con una situación ideal, por lo que la clave aquí es encontrar un algoritmo de simulación que se parezca (ajuste) lo más posible a lo que sucede en la vida real.

Para el presente trabajo se utilizó como base el algoritmo de simulación de datos de experimentos de microarreglos publicado por Chang *et al.* (2013), para evaluar el desempeño de las técnicas de MA seleccionadas. Si bien el algoritmo comienza simulando intensidades de señal lumínica característica de los mencionados experimentos, como producto final se obtienen valores p (un valor p para cada gen) producto de suponer que se está trabajando sólo con dos condiciones (control y tratamiento).

En la Sección 3, dedicada a la metodología se comienza detallando algunas consideraciones inherentes a las técnicas de MA para luego enumerar aquellas incluidas en este trabajo. Se continúa mencionando cómo se procedió a generar los datos simulados y las herramientas empleadas para evaluar el desempeño de las técnicas de MA. En la Sección 4 se exponen los resultados obtenidos de esta evaluación en los diferentes escenarios propuestos. Por último, se dedica una sección a las consideraciones finales en donde se realiza un comentario de los resultados hallados aquí en comparación con los encontrados por otros autores y se proponen análisis a realizar para futuras líneas de investigación.

2. Objetivos

Objetivo general:

Evaluar y comparar el desempeño de las técnicas estadísticas más comúnmente utilizadas en la combinación de resultados (MA) provenientes de experimentos de microarreglos en situaciones donde tanto la cantidad de estudios involucrados como el número de réplicas realizadas en cada estudio son pequeños.

Objetivos específicos:

1. Compendiar las técnicas estadísticas de MA más comúnmente utilizadas para el caso de combinación de resultados provenientes de experimentos de microarreglos.
2. Generar bases de datos a partir de simulaciones para distintas combinaciones de valores de k (*número de estudios*) y de r (*número de réplicas*) pequeños, con valores de k entre 2 y 7, y valores de r entre 2 y 5.
3. Presentar una medida del desempeño de cada una de las técnicas de MA en cada uno de los escenarios y para cada configuración de hipótesis.

3. Metodología

En esta sección se presentarán las técnicas de MA seleccionadas para desarrollar el presente trabajo. Primero se detallarán algunas consideraciones que se tuvieron en cuenta al momento de acotar las técnicas con las cuales se trabajará para luego enumerarlas. A continuación se menciona como se procedió a generar los datos simulados para los distintos escenarios propuestos y las herramientas empleadas para evaluar el desempeño de las técnicas de MA en cada una de esas situaciones.

3.1. Consideraciones inherentes a las técnicas de MA

La elección de las técnicas estadísticas para la combinación de resultados depende principalmente del objetivo del MA propuesto por el investigador especialista y de los datos disponibles para cumplir con el objetivo planteado. Como se mencionó anteriormente, Tseng *et al.* (2012) detallan ciertas particularidades a tener en cuenta en este paso. A continuación se mencionan algunas de ellas que sirvieron de marco para seleccionar las técnicas de MA incluidas en este trabajo.

3.1.1. La configuración de las hipótesis nula y alternativa detrás del objetivo del MA

Si bien existen métodos basados en el enfoque bayesiano, las técnicas estadísticas para la combinación de resultados en investigaciones de MA más utilizadas para estudios de diferenciación de genes utilizan el enfoque frecuentista. Siguiendo la convención propuesta por Birnbaum (1954), Li y Tseng (2011) y luego con los aportes de Song y Tseng (2014), los objetivos de los MA se pueden clasificar en base a tres configuraciones de hipótesis complementarias.

En la primera configuración de hipótesis (denotada como HS_A), el objetivo es identificar un gen como DE cuando el efecto de su expresión sea distinto de cero en todos los estudios, esto es, cuando el valor de probabilidad asociada a la estadística del test (valor p) indique que hubo un cambio significativo en su expresión en los k estudios. Así esta configuración puede expresarse como:

$$H_0) \cap_{k=1}^K \{\theta_k = 0\} \quad \text{versus} \quad H_1) \cap_{k=1}^K \{\theta_k \neq 0\} \quad (\text{HS}_A)$$

donde θ_k es el efecto de la expresión del gen en el k -ésimo estudio, con $k=1, \dots, K$. Como se mencionó anteriormente, esta sería la situación elegida cuando se desea buscar un marcador genético, por ejemplo, un gen asociado a una determinada condición o enfermedad.

La segunda configuración de hipótesis (denotada como HS_B), apunta a identificar un gen como DE si tiene un efecto distinto de cero en por lo menos un estudio, si el valor p indica que hubo un cambio significativo de su expresión en uno o más estudios. En otras palabras, si el gen es clasificado como DE en por lo menos uno de los K estudios.

$$H_0) \cap_{k=1}^K \{\theta_k = 0\} \quad \text{versus} \quad H_1) \cup_{k=1}^K \{\theta_k \neq 0\} \quad (\text{HS}_B)$$

Esta situación se presenta por ejemplo en estudios donde se buscan genes involucrados en una ruta metabólica particular. Así los investigadores seleccionan un grupo de genes potenciales que luego serán estudiados con otras técnicas para decidir respecto a su verdadera participación en esta.

Dado que HS_A frecuentemente es demasiado conservadora cuando se combinan muchos estudios, Song y Tseng (2014) propusieron una configuración de hipótesis más práctica y robusta (denotada como HS_r). Esta se enfoca en los genes DE con efecto distinto a cero en por lo menos h de los K estudios combinados. Los mencionados autores proponen expresar esta configuración de hipótesis más robusta como:

$$H_0) \cap_{k=1}^K \{\theta_k = 0\} \quad \text{versus} \quad H_1) \cup_{h=1}^K \theta_h \quad (\text{HS}_r)$$

donde $\theta_h = \{\sum_{k=1}^K I(\theta_k \neq 0) \geq h\}$ representa la situación en donde el gen es considerado DE en por lo menos h de los K estudios.

3.1.2. Datos disponibles de los estudios a incluir en el MA

La elección de la mejor técnica estadística para combinar los resultados en un MA depende de cómo se presentan los datos en los diferentes trabajos a

incluir en el MA (Tseng *et al.*, 2012). Básicamente, la información referente a los genes identificados como DE se pueden presentar como:

- Una **lista de genes** donde se presentan los genes que se declaran DE en un estudio determinado. A pesar de que usualmente se encuentran en el texto principal o anexo de los estudios publicados y, por lo tanto, son fáciles de obtener, estos datos son de uso limitado para el MA. Esto se debe a que representan sólo un subconjunto de los genes realmente estudiados, y la información de muchos genes estará completamente ausente.
- Los **datos crudos**. Los datos así presentados no dependen del pre-procesamiento de la información utilizado por cada grupo de investigación y sobre todo, permiten diferenciar los genes DE que se sobre-expresan de los que se sub-expresan. Esta información es muy valiosa para los investigadores especialistas, pero no siempre se encuentra disponible.
- El **valor p** obtenido para cada gen. Este tipo de dato tiene como ventaja la de brindar más información que la lista de genes, sin embargo, depende del procesamiento de los datos crudos que realiza cada grupo de investigación al publicar sus resultados. Otra particularidad es que generalmente no permite diferenciar los genes que se sobre-expresan de los que se sub-expresan. Más allá de las desventajas mencionadas es más factible encontrar disponible este tipo de información en los trabajos de investigación. Es así que se optó por este tipo de datos para desarrollar el presente trabajo.

3.1.3. Diferenciar o no entre la sub-expresión y sobre-expresión de los genes DE

Autores como Chang *et al.* (2013) y Makinde *et al.* (2021) clasifican a las técnicas de MA en varias categorías dependiendo de cómo combinan la información para detectar genes DE. Entre estas se encuentran las que combinan valores p , las que combinan tamaños de efecto y aquellas que utilizan asignación de rangos en lugar de la información original. Tseng *et al.* (2012) presentan dos categorías adicionales. Una de ellas abarca a las estrategias basadas en fusionar

directamente datos crudos normalizados y la otra se encuentra formada por aquellas que utilizan un enfoque más sofisticado de variable latente.

Dentro de estas técnicas, la que consiste en unir directamente los datos crudos normalizados y las que pertenecen al grupo que combinan tamaños de efecto identifican automáticamente los genes que se sobre-expresan de los que se sub-expresan en forma consistente en todos los estudios. Por otro lado, las técnicas que combinan valores p o rangos si estos son obtenidos de test de hipótesis a dos colas no permiten diferenciar entre los dos tipos de expresión.

Como se mencionó en el apartado anterior se seleccionaron técnicas que trabajan con valores p para el desarrollo del presente trabajo. Esta especificación condiciona la no diferenciación entre genes sub-expresados y sobre-expresados. Es por esto que simplemente se mencionarán en general como genes DE.

3.2. Técnicas de MA

En situaciones experimentales donde sólo se desea comparar la expresión de genes entre dos condiciones, en general un tratamiento con un control, para cada gen g ($g = 1, \dots, G$), luego de realizar el experimento de microarreglos y las correcciones correspondientes a la señal lumínica, se obtiene un valor de FC para cada gen. Como se mencionó anteriormente, este valor de FC es sometido a un test estadístico, test t modificado específico, obteniendo así un valor p para cada gen. Si este valor p es menor o igual a un nivel de significación determinado, en general fijado en 0,05, se considera que hay una diferencia estadísticamente significativa en la expresión de ese gen entre las dos condiciones consideradas, generalmente del estado tratado en relación al control, y se clasificará al gen como DE.

En las situaciones de MA, donde se busca combinar la información de k estudios ($k = 2, \dots, K$), para cada gen se tendrán varios valores p , uno por cada estudio considerado. Existen diversas técnicas de MA que permiten utilizar apropiadamente estos valores p de modo de poder clasificar a cada gen como DE o no DE en el MA teniendo en cuenta la información proveniente de los k estudios en forma conjunta. En todos los casos, la hipótesis nula a evaluar es que no hay diferencia en las medias de expresión del gen entre las dos condiciones

consideradas. Entre las técnicas más frecuentemente utilizadas en la bibliografía consultada se seleccionaron las que se presentan a continuación.

3.2.1. Mínimo valor p

El método del mínimo valor p (minP) fue presentado por Tippett en 1931. La estadística de este método consiste en tomar el valor p más pequeño entre los K estudios considerados como el estadístico de prueba. Esta estadística, bajo la hipótesis nula, sigue una distribución beta con grados de libertad $\alpha = 1$ y $\beta = K$ (Tippett, 1931). Por las características propias de esta estadística es fácil ver que esta técnica detecta un gen como DE en el MA siempre que exista un valor p pequeño en por lo menos uno de los K estudios.

3.2.2. Máximo valor p

La estadística máximo valor p (maxP) de Wilkinson (1951) consiste en tomar el valor p más grande observado entre los K estudios como estadística de prueba para el gen. Esta estadística, bajo la hipótesis nula, sigue una distribución beta con grados de libertad $\alpha = K$ y $\beta = 1$. Es así que este método clasifica al gen como DE en el MA si este presenta valores p pequeños en los K estudios.

3.2.3. Valor p de orden r

El método valor p de orden r (roP) toma como estadística el valor p ubicado en el r -ésimo lugar entre los valores p ordenados de los K estudios. Bajo la hipótesis nula, esta estadística sigue una distribución beta con grados de libertad $\alpha = r$ y $\beta = K - r + 1$ (Song y Tseng, 2014). La elección del valor de r depende principalmente del interés del investigador especialista en el tema, y en particular del objetivo de su trabajo.

En el presente trabajo se estableció el número r de modo de considerar un gen como DE en el MA cuando presenta valores p pequeños en por lo menos la mitad de los K estudios considerados. Esto es:

- en situaciones donde K es un número par, $r = 0,5 * K$,
- en los casos donde K es un número impar, $r = (0,5 * K) + 0,5$.

Puede observarse que los métodos minP y maxP son casos especiales de roP. Como se mencionó en los apartados anteriores, el método minP clasificará al gen como DE si este presenta valores p pequeños en por lo menos uno de los K estudios considerados, esto equivale a establecer $r = 1$ para la técnica roP. Para el método maxP, que clasifica al gen como DE si este presenta valores p pequeños en todos los K estudios por lo que la equivalencia con el método roP se establece cuando $r = K$.

3.2.4. Estadística de Fisher

El método de Fisher es uno de los primeros utilizados para combinar valores p aplicados al MA de microarreglos. Fue desarrollado por Fisher en 1925 y se basa en la combinación de los valores p de conjuntos de datos independientes. Utiliza como estadística de prueba la suma de los valores p obtenidos de los estudios individuales transformados logarítmicamente. Considerando que p_{gk} es el valor p para el gen g en el k -ésimo estudio, la estadística de esta técnica podría definirse como:

$$S_g = -2 \sum_{k=1}^K \log(p_{gk})$$

Esta estadística, bajo la hipótesis nula, sigue una distribución χ^2 con $2K$ grados de libertad, donde K es el número total de estudios (Campain y Yang, 2010). La desventaja de este método es que un valor p extremadamente pequeño en un solo estudio puede ser suficiente para clasificar al gen como DE aún cuando ese gen no sea significativo en ninguno de los otros estudios (Chang *et al.*, 2013).

3.2.5. Estadística de Stouffer

La estadística utilizada por el método de Stouffer suma los valores p transformados normales inversos. Esta estadística, introducido por Stouffer en 1949 queda definida entonces como:

$$T_{Stouffer} = \sum_{k=1}^K \frac{z_k}{\sqrt{K}}$$

donde $z_k = \Phi^{-1}(p_k)$, siendo Φ la función de distribución acumulada de una variable normal estándar.

Esta estadística sigue una distribución normal estándar bajo la hipótesis nula (Stouffer, 1949). En modo similar al método de Fisher, los valores p más pequeños contribuyen más a dar un resultado significativo, pero en una menor magnitud (Chang *et al.*, 2013).

3.2.6. Estadística de Fisher con pesos adaptados

La estadística de Fisher con pesos adaptados (AW, siglas que provienen del término en inglés *Adaptively Weighted*) fue desarrollada por Li y Tseng en 2011. Los autores proponen una estadística basada en la estadística de Fisher pero que asigna diferentes pesos a cada estudio individual. Así la estadística de este método queda definida como:

$$T_{AW} = - \sum_{k=1}^K w_k \log(p_{gk})$$

donde $w_k = 0$ o 1 y busca en todos los pesos posibles para encontrar el mejor peso adaptativo con el valor p derivado más pequeño. Si se asume que p_{gk} tiene distribución Uniforme $[0,1]$ bajo la hipótesis nula. La estadística tendrá distribución Gamma $(\sum_{k=1}^K w_k, 1)$. Sin embargo, el supuesto de valor p uniforme suele no ser cierto en aplicaciones reales por lo que se prefiere la alternativa de análisis por permutaciones (Li y Tseng, 2011). Una ventaja significativa de este método es su capacidad para indicar qué estudios contribuyen más al MA (Chang *et al.*, 2013).

3.2.7. Suma de rangos

La técnica suma de rangos (SR) no trabaja directamente con valores p , sino que se basa en la asignación de rangos en función de estos. Al ordenar en cada estudio los genes según su valor p , si un gen está repetidamente en la parte

superior de esta lista ordenada, es más probable que el gen se encuentre DE en el MA. La estadística de esta técnica consiste simplemente en sumar los rangos asignados a los valores p en cada estudio (Chang *et al.*, 2013). Si R_{gk} representa el rango que obtuvo el valor p del gen g entre todos los genes en el estudio k , la estadística de prueba queda definida como:

$$SR_g = \sum_{k=1}^K R_{gk}$$

El valor p de esta estadística de prueba puede ser calculada en forma analítica o ser obtenida de un análisis de permutaciones.

3.2.8. Producto de rangos

La técnica denominada producto de rangos (PR) es análoga a la presentada en el ítem anterior, su diferencia radica en multiplicar los rangos asignados a los valores p en lugar de sumarlos (Chang *et al.*, 2013) por lo que en este caso la estadística de prueba se expresa como:

$$PR_g = \prod_{k=1}^K R_{gk}$$

Una desventaja importante es que los métodos de combinación de rangos como SR y PR, sólo consideran los rangos de genes en lugar de sus correspondientes valores p . Sin embargo, esto también le concede la ventaja de su robustez frente a la heterogeneidad en diferentes estudios (Makinde *et al.*, 2021).

Si bien los valores p de cada una de las estadísticas de prueba antes mencionadas pueden calcularse analíticamente utilizando sus correspondientes distribuciones, también pueden obtenerse a través de un análisis de permutación (Chang *et al.* 2013). Para este desarrollo se optó por el segundo método para seguir con los lineamientos de los autores antes mencionados y así poder

comparar los resultados obtenidos de este trabajo con los publicados por dichos autores.

3.3. Generación de datos

Se simularon intensidades de señal lumínica, como las esperadas al realizar experimentos de microarreglos. Se utilizó como base el algoritmo de simulación propuesto por Chang *et al.* (2013). En el artículo publicado los autores proponen generar niveles de expresión de genes como señales lumínicas para 2.000 genes por experimento, 800 de ellos en *clusters* con 40 grupos de genes (20 genes en cada grupo) y otros 1.200 genes que no pertenecen a ningún grupo. Trabajaron considerando 5 estudios a combinar. De su simulación resultaban 1.000 genes DE entre los 5 estudios de modo de tener aproximadamente 200 genes DE comunes entre los estudios (Chang *et al.*, 2013).

En la generación de datos se consideró un total de 10.000 genes por experimento de los cuales 500 fueron DE. Esta modificación realizada sobre el algoritmo de los autores antes mencionados pretendió considerar situaciones más parecidas a lo que se observa en experimentos biológicos con plantas (Daurelio *et al.*, 2015). Es así que también se consideraron situaciones experimentales con diferentes cantidades de réplicas biológicas (r entre 2 y 5).

Una vez simuladas las señales lumínicas, se procesaron, en analogía con lo que ocurre en experimentos reales. Utilizando el paquete LIMMA se aplican modelos lineales para procesar las señales teniendo en cuenta entre otras cosas la “corrección por fondo” y la “normalización” de las señales. De este modo se obtuvo un valor de FC para cada gen que luego se sometió a un test estadístico, test t modificado específico para este tipo de experimentos (Ritchie *et al.*, 2015; Moreno y Solé, 2004). Como resultado final se obtuvo un valor p para cada gen que permitió clasificarlo como DE o como no DE dentro de cada estudio individual.

Utilizando las simulaciones de estos estudios independientes se construyeron posibles escenarios de MA donde la cantidad de estudios a incluir también era pequeña, esto es para k entre 2 y 7, quedando así definidas 24 situaciones o escenarios diferentes producto de la combinación $K \times R$.

Paso seguido se procedió a unir la información proveniente de diferentes experimentos simulados para cada escenario propuesto, aplicando las técnicas de MA antes presentadas, estas son: minP, maxP, roP, Fisher, Stouffer, AW, PR y SR. Así, para cada escenario, se obtuvo un valor p para cada gen con cada una de las técnicas de MA. Como se adelantó en la sección anterior, para obtener estos valores p se utilizó un análisis de permutaciones en lugar de la correspondiente evaluación paramétrica para cada uno de los casos antes mencionados.

3.4. Evaluación del desempeño de las técnicas de MA

El desempeño de las técnicas de MA dentro de cada situación planteada se evaluó bajo cada combinación de hipótesis mencionadas en el apartado 3.1.1 (HS_A , HS_B y HS_r), lo que llevó a considerar un total de 72 escenarios posibles. Para cada técnica y en cada escenario, se calculó el área debajo de la correspondiente curva ROC (acrónimo de *Receiver Operating Characteristic*), conocido como AUC (siglas que provienen del término en inglés *Area Under the Curve*).

Para construir la curva ROC correspondiente a una determinada técnica de MA en un escenario particular se cuenta para cada uno de los genes, con el valor p obtenido por esa técnica en el correspondiente escenario y con la información de su estado verdadero (DE o no DE) dado por lo especificado en la simulación. En el estudio de técnicas como las aquí estudiadas donde se obtienen resultados en un rango continuo (valores p) es necesario establecer un punto de corte que divida los resultados en dos grupos: (positivo y negativo) para poder calcular su sensibilidad (S) y su especificidad (E) frente al método de referencia (simulaciones).

La sensibilidad mide la capacidad de la técnica para detectar a un gen DE cuando realmente lo es. Se calcula haciendo el cociente de la cantidad de genes considerados DE por la técnica sobre el total de genes verdaderamente DE. La especificidad mide la capacidad de la técnica para detectar a un gen no DE cuando ese es su estado verdadero. Se calcula haciendo el cociente de la cantidad de genes considerados no DE por la técnica sobre el total de genes verdaderamente no DE.

Las pruebas no tienen un solo par sensibilidad/especificidad (S/E), sino muchos, uno por cada nivel de decisión o punto de corte que se adopte. La curva ROC es un gráfico en el que se observan todos los pares sensibilidad/especificidad resultantes de la variación continua de los puntos de corte en todo el rango de resultados observados. En el eje y de coordenadas se sitúa la sensibilidad o fracción de verdaderos positivos/DE y calculada en el grupo de genes definido como DE. En el eje x se sitúa la fracción de falsos positivos o 1-especificidad, calculada en el subgrupo de genes no DE (Burgueño *et al.*, 1995).

Una prueba con discriminación perfecta, sin solapamiento de resultados en las dos poblaciones, tiene una curva ROC que pasa por la esquina superior izquierda, donde S y E toman valores máximos (S y E = 1). Una prueba sin discriminación, con igual distribución de resultados en los dos subgrupos, da lugar a una línea diagonal de 45°, desde la esquina inferior izquierda hasta la superior derecha.

El área debajo de la curva ROC (AUC) da una medida global de la exactitud de la técnica. Toma el valor de 0,5 cuando no existen diferencias en la distribución de los resultados dado por la técnica entre los subgrupos DE y no DE, y el valor de 1,0 cuando brinda una separación perfecta entre los dos grupos.

Para evaluar qué técnica presenta mejor desempeño en cada uno de los escenarios propuestos se procedió a realizar un test de comparaciones múltiples entre los valores de AUC. Como primer paso se realizaron las correspondientes comparaciones de a dos utilizando el método de DeLong (DeLong *et al.*, 1988) para muestras apareadas. El nivel de significación fue ajustado por el método FDR (siglas que provienen del término en inglés *False Discovery Rate*). En todos los casos se considera que la diferencia entre las AUC es estadísticamente significativa cuando el valor *p* es menor o igual a 0,05. Para visualizar el resultado de estas comparaciones se aplicó CLD (siglas que provienen del término en inglés *Compact Letter Display*). Esta técnica de visualización consiste en etiquetar, en este caso a las técnicas de MA, con una o más letras de modo que se pueda decir que no hay diferencia estadísticamente significativa entre las AUC de las técnicas que comparten al menos una letra.

3.5. Soporte Computacional

En el desarrollo del presente trabajo se utilizó el *software* R (R Core Team, 2024) tanto para la generación de datos como para la aplicación de las distintas técnicas de MA consideradas y su posterior evaluación de desempeño. Se trabajó con valores iniciales para la generación de valores pseudo aleatorios de manera de garantizar la reproducibilidad de los resultados.

Tal como fue mencionado, para la generación de los datos en el contexto de cada uno de los escenarios propuestos, se utilizó el algoritmo propuesto por Chang *et al.* (2013). Así se generaron señales lumínicas según lo esperado para experimentos de microarreglos teniendo en cuenta las especificaciones descriptas en el inciso 3.3. Para el procesamiento y obtención final de los valores p para cada gen dentro de cada estudio se utilizó el paquete LIMMA de Bioconductor (Ritchie, 2015)

Para la aplicación de las técnicas de MA para combinar la información proveniente de k estudios en cada uno de los escenarios bajo estudio se utilizó la función `MetaDE.pvalue()` del paquete `MetaDE` (Wang *et al.*, 2012). Los métodos de MA especificados fueron: maxP, minP, roP, Fisher, Stouffer, SR y PR y se utilizó análisis de permutaciones para la obtención de los valores p . Para la técnica de MA denominada AW se utilizó el algoritmo de Chang *et al.* (2013), cuyo *script* fue provisto por dichos autores.

En lo que se refiere a la evaluación del desempeño de cada una de las técnicas de MA para cada uno de los escenarios, se calculó el AUC utilizando las funciones `roc()` y `auc()` del paquete `pROC` (Robinet *et al.*, 2011). También se utilizó la función `roc.test()` del mismo paquete para realizar la comparación de a pares aplicando el test de DeLong para muestras apareadas. A los valores p obtenidos de las comparaciones de a pares entre todas las técnicas se les aplicó la corrección FDR utilizando la función `p.adjust()` del paquete `base stats` (R Core Team, 2024). Para una mejor visualización en formato tabla de estos resultados se aplicó la función `make_cld()` del paquete `biostat` (Gegzna, 2020).

Por último, los gráficos presentados en este escrito fueron realizados mediante el paquete `ggplot2` (Wickham, 2016). Los códigos utilizados en este caso se encuentran disponibles en el Anexo A.

4. RESULTADOS

A continuación, se presentan los resultados obtenidos para cada uno de los 72 escenarios propuestos. Los resultados se ordenan en función de la cantidad de estudios a combinar, comenzando por $k=2$.

En términos generales se puede decir que las técnicas de MA estudiadas en todos los escenarios considerados bajo las tres configuraciones de hipótesis analizadas presentaron valores del AUC mayores a 0,5, indicando que el uso de cualquiera de ellas es apropiado para combinar resultados de experimentos de microarreglos provenientes de distintos estudios relacionados. Sin embargo, al realizar el test para comparar los valores del AUC entre las diferentes técnicas para cada uno de los escenarios propuestos se puede considerar que algunas técnicas se comportaron mejor que otras.

4.1. Desempeño de las técnicas de MA cuando se combinan dos estudios

Para los escenarios en los cuales la cantidad de estudios a combinar es igual a dos es importante notar que trabajar bajo los pares de hipótesis HS_B y HS_r resulta equivalente. Esto es, clasificar a un gen como DE para el MA cuando el efecto de su expresión sea distinto de cero en por lo menos un estudio, en este caso uno de dos estudios, es lo mismo que identificar a un gen como DE cuando el efecto es distinto de cero en por lo menos la mitad de los estudios combinados. Además, dada la naturaleza de las técnicas minP y roP, los resultados obtenidos con ambas son iguales. La estadística de la técnica minP consiste en tomar el valor p más pequeño entre los estudios considerados por lo que para este caso, indicará que un gen se encuentra DE cuando presente un valor p pequeño en por lo menos uno de los dos estudios. En esta situación donde k es igual a dos lo mismo ocurre para la técnica roP ya que esta toma como estadística el valor p ubicado en el r -ésimo lugar, clasificando a un gen como DE cuando se observan valores p pequeños en por lo menos la mitad de los estudios, para este caso en por lo menos uno de los dos estudios considerados. (Grafico 1, Tabla 1)

Tabla 1: Valores del AUC y resultado del test de comparaciones múltiples entre las técnicas de MA para cada escenario presentado con $k=2$

	HS _B /HS _r			HS _A		
	Técnica	Test*	AUC	Técnica	Test*	AUC
$r = 2$	Fisher	b	0,8430	Fisher	c	0,8879
	AW	b	0,8414	AW	bc	0,8831
	minP/ roP	b	0,8411	minP/ roP	b	0,8780
	SR	c	0,7796	SR	d	0,8502
	PR	c	0,7637	Stouffer	a	0,8287
	Stouffer	c	0,7630	PR	a	0,8225
	maxP	a	0,7172	maxP	a	0,8133
$r = 3$	Fisher	b	0,8471	Fisher	c	0,9216
	minP/ roP	bc	0,8443	AW	bc	0,9187
	AW	c	0,8408	minP/ roP	b	0,9158
	Stouffer	d	0,7848	Stouffer	d	0,8756
	PR	de	0,7721	SR	d	0,8633
	SR	e	0,7647	PR	ad	0,8611
	maxP	a	0,7357	maxP	a	0,8444
$r = 4$	Fisher	b	0,9163	Fisher	c	0,9605
	minP/ roP	bc	0,9142	AW	b	0,9557
	AW	c	0,9117	minP/ roP	b	0,9518
	Stouffer	d	0,8480	Stouffer	d	0,9252
	SR	d	0,8344	SR	d	0,9126
	PR	a	0,7803	maxP	a	0,8709
	maxP	a	0,7613	PR	a	0,8655
$r = 5$	minP/ roP	b	0,9101	Fisher	b	0,9614
	Fisher	b	0,9092	minP/ roP	b	0,9607
	AW	b	0,9091	AW	b	0,9602
	Stouffer	c	0,8540	Stouffer	c	0,9226
	PR	d	0,8133	SR	d	0,9034
	SR	d	0,8015	PR	ad	0,8948
	maxP	a	0,7644	maxP	a	0,8755

*Si dos o más técnicas de MA comparten la misma letra, se dice que no hay diferencia estadísticamente significativa en los respectivos valores del AUC

Al evaluar el desempeño de las técnicas de MA para $k=2$ se observó para las dos configuraciones de hipótesis que, en general, a medida que aumentó la cantidad de réplicas por estudio, aumentaron los valores del AUC; esto es, las

técnicas de MA presentaron mejor desempeño. También se vió que para todas las técnicas de MA y para todos los r , los valores del AUC fueron mayores cuando se trabajó bajo la configuración de hipótesis HS_A que cuando se hizo bajo HS_B/HS_r .

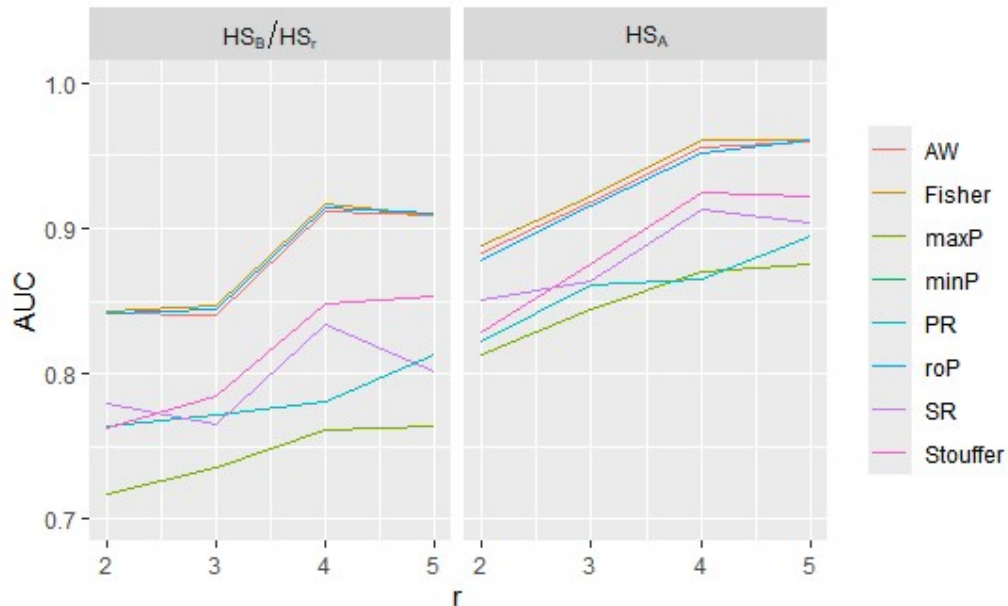


Gráfico 1: Valores del AUC en función de la cantidad de réplicas (r) según técnica de MA utilizada y para cada una de las configuraciones de hipótesis, cuando se combinan dos estudios ($k=2$)

Al analizar los resultados obtenidos para la configuración de hipótesis HS_B/HS_r , las técnicas que presentaron mejor desempeño fueron: Fisher, AW y minP/roP. En la Tabla 1 se pueden observar que estas técnicas presentaron los valores de AUC más altos. Además al analizar el resultado de aplicar CLD para $r = 2$ y $r = 5$, las tres están etiquetadas con la misma letra indicando que no se observó diferencia estadísticamente significativa entre ellas, pero sí entre estas y las restantes. En cambio, al observar las letras con las que se etiquetaron estas técnicas para $r = 3$ y $r = 4$ se puede decir que no se observó diferencia significativa entre Fisher y minP/roP, ni entre minP/roP y AW, pero sí entre Fisher y AW ya que estas dos se encuentran etiquetadas con letras diferentes.

Respecto a la configuración de hipótesis HS_A en particular, las técnicas con mejor desempeño fueron las mismas que las mencionadas en el párrafo anterior,

sin embargo, al comparar las AUC para $r = 2$ y $r = 3$ no se observó diferencia estadísticamente significativa entre Fisher y AW, ni entre AW y minP/roP pero sí entre Fisher y minP/roP. Para $r = 4$ no se observó diferencia estadísticamente significativa entre AW y minP/roP pero sí entre Fisher y estas. En el caso donde $r = 5$ no se observó diferencia estadísticamente significativa entre Fisher, minP/roP y AW.

Resumiendo, la técnica de MA que presentó mejor desempeño en casi todos los casos es la técnica de Fisher. En una sola situación no presentó el mayor valor del AUC entre las técnicas consideradas, esto se observó cuando $k = 2$ y $r = 5$ para la combinación de hipótesis HS_B/HS_r , pero en esta situación no hubo diferencia estadísticamente significativa entre esta y las técnicas de MA con valores del AUC más altos.

Respecto a la técnica con peor desempeño se observó que maxP presentó menor valor del AUC en casi todos los escenarios, excepto cuando se trabaja bajo HS_A con $r = 4$. En este caso la técnica con menor AUC es PR, sin embargo, al realizar el test de comparaciones múltiples no se observó diferencia estadísticamente significativa de las AUC entre ambas.

4.2. Desempeño de las técnicas de MA cuando se combinan tres estudios

Al evaluar el desempeño de las técnicas de MA en los casos de combinación de resultados provenientes de tres estudios ($k = 3$) se observó que, en general, para cada una de las configuraciones de hipótesis, a medida que aumentó la cantidad de réplicas por estudio, aumentaron los valores del AUC para cada técnica de MA; esto es, las técnicas de MA presentaron mejor desempeño (Gráfico 2). En particular, para cada técnica MA y para todos los r , los valores del AUC fueron mayores cuando se trabajó bajo la configuración de hipótesis HS_A que cuando se hizo bajo HS_r o HS_B donde se observaron los valores del AUC más pequeños (Gráfico 2, Tabla 2).

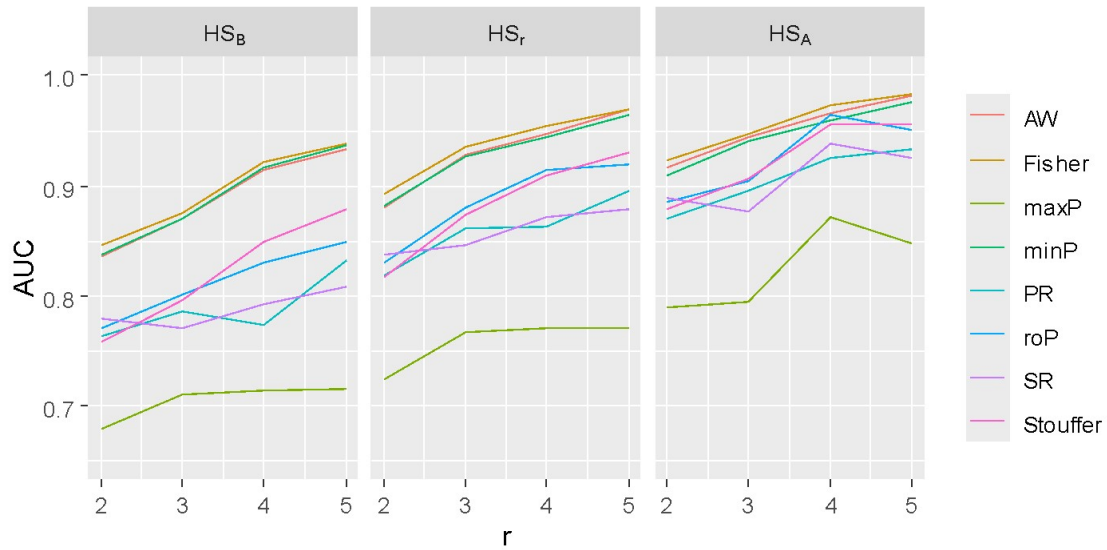


Gráfico 2: Valores del AUC en función de la cantidad de réplicas (r) según técnica de MA utilizada y para cada una de las configuraciones de hipótesis, cuando se combinan tres estudios ($k = 3$)

Al considerar todas las combinaciones de hipótesis se vió que la técnica de MA que presentó mejor desempeño en casi todos los escenarios fue Fisher. La excepción se presentó cuando se trabajó con $r = 5$ bajo HS_r , en este caso la mejor técnica fue AW y Fisher quedó en segundo lugar; sin embargo el test de comparaciones múltiples de las AUC indicó que no hubo diferencia estadísticamente significativa entre estas (Tabla 2).

Respecto a la técnica con peor desempeño, se observó que maxP es la que presentó menor valor del AUC en todos los escenarios (Gráfico 2). A esto se suma que al realizar el test de comparaciones múltiples de las AUC se observó diferencia estadísticamente significativa entre maxP y las restantes técnicas de MA (Tabla 2).

Tabla 2: Valores del AUC y resultado del test de comparaciones múltiples entre las técnicas de MA para cada escenario presentado con $k = 3$

	HS _B			HS _r			HS _A		
	Técnica	Test*	AUC	Técnica	Test*	AUC	Técnica	Test*	AUC
$r = 2$	Fisher	e	0,8456	Fisher	e	0,8929	Fisher	e	0,9242
	minP	b	0,8373	minP	b	0,8825	AW	c	0,9167
	AW	b	0,8347	AW	b	0,8801	minP	bc	0,9100
	SR	d	0,7794	SR	d	0,8369	SR	bd	0,8887
	roP	cd	0,7702	roP	cd	0,8303	roP	bd	0,8853
	PR	cd	0,7633	PR	cd	0,8188	Stouffer	d	0,8783
	Stouffer	c	0,7577	Stouffer	c	0,8165	PR	d	0,8700
	maxP	a	0,6781	maxP	a	0,7244	maxP	a	0,7898
$r = 3$	Fisher	b	0,8747	Fisher	d	0,9358	Fisher	d	0,9476
	minP	b	0,8701	AW	b	0,9282	AW	bd	0,9442
	AW	b	0,8698	minP	b	0,9260	minP	b	0,9411
	roP	c	0,8012	roP	c	0,8806	Stouffer	c	0,9054
	Stouffer	c	0,7964	Stouffer	c	0,8736	roP	c	0,9042
	PR	cd	0,7858	PR	ce	0,8621	PR	ce	0,8958
	SR	d	0,7698	SR	e	0,8452	SR	e	0,8772
	maxP	a	0,7091	maxP	a	0,7660	maxP	a	0,7937
$r = 4$	Fisher	d	0,9217	Fisher	d	0,9541	Fisher	c	0,9733
	minP	b	0,9160	AW	b	0,9481	AW	d	0,9666
	AW	b	0,9146	minP	b	0,9440	roP	bcd	0,9645
	Stouffer	e	0,8485	roP	c	0,9152	minP	b	0,9589
	roP	c	0,8295	Stouffer	c	0,9093	Stouffer	b	0,9554
	SR	f	0,7927	SR	e	0,8721	SR	e	0,9391
	PR	g	0,7742	PR	e	0,8627	PR	e	0,9246
	maxP	a	0,7137	maxP	a	0,7699	maxP	a	0,8716
$r = 5$	Fisher	b	0,9380	AW	d	0,9694	Fisher	e	0,9843
	minP	b	0,9369	Fisher	bd	0,9692	AW	e	0,9826
	AW	b	0,9340	minP	b	0,9649	minP	b	0,9763
	Stouffer	d	0,8793	Stouffer	c	0,9298	Stouffer	c	0,9556
	roP	c	0,8497	roP	c	0,9204	roP	cd	0,9515
	PR	c	0,8325	PR	e	0,8958	PR	df	0,9331
	SR	e	0,8073	SR	e	0,8782	SR	f	0,9258
	maxP	a	0,7152	maxP	a	0,7707	maxP	a	0,8478

*Si dos o más técnicas de MA comparten la misma letra, se dice que no hay diferencia estadísticamente significativa en los respectivos valores del AUC.

4.3. Desempeño de las técnicas de MA cuando se combinan cuatro estudios

Cuando se estudió qué técnica de MA es más apropiada en el caso de que se quiera combinar los resultados provenientes de cuatro estudios ($k = 4$) se observó en general que para cada configuración de hipótesis, a medida que aumentó la cantidad de réplicas por estudio, aumentaron los valores del AUC; esto es, las técnicas de MA presentaron mejor desempeño a medida que aumentó r (Gráfico 3). Además, para todas las técnicas de MA y para todos los r , los valores del AUC fueron mayores cuando se trabajó bajo la configuración de hipótesis HS_A que cuando se hizo bajo HS_r o HS_B donde se observan los valores del AUC más pequeños (Gráfico 3, Tabla 3).

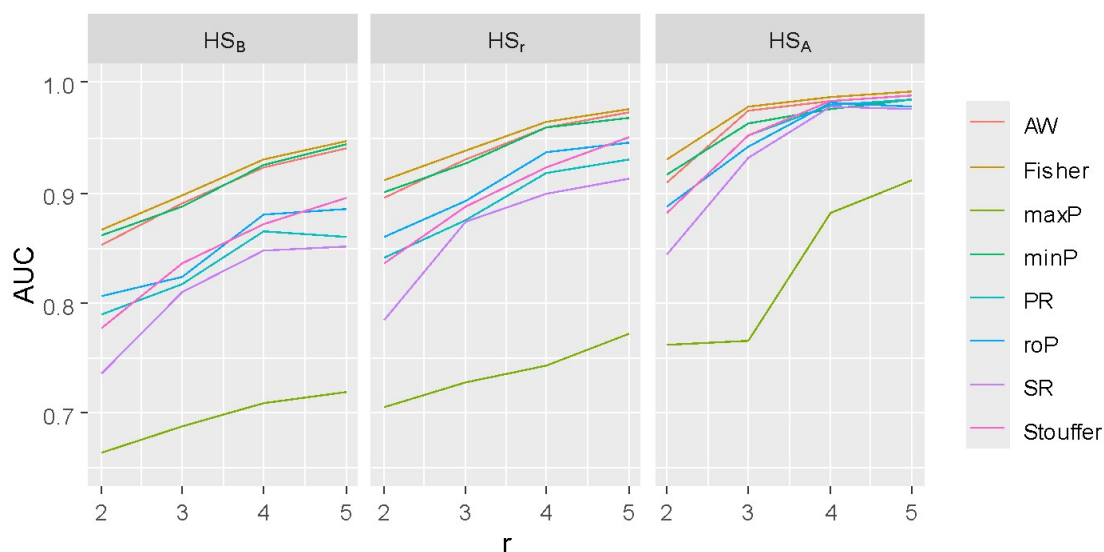


Gráfico 3: Valores del AUC en función de la cantidad de réplicas (r) según técnica de MA utilizada y para cada una de las configuraciones de hipótesis, cuando se combinan cuatro estudios ($k = 4$)

Cuando se consideran todas las combinaciones de hipótesis, se observó que la técnica de MA que presentó mejor desempeño en todos los escenarios fue la técnica de Fisher (Tabla 3).

Tabla 3: Valores del AUC y resultado del test de comparaciones múltiples entre las técnicas de MA para cada escenario presentado con k=4

	HS _B			HS _r			HS _A		
	Técnica	Test*	AUC	Técnica	Test*	AUC	Técnica	Test*	AUC
r = 2	Fisher	b	0,8659	Fisher	d	0,9119	Fisher	d	0,9300
	minP	bc	0,8613	minP	b	0,9002	minP	b	0,9156
	AW	c	0,8530	AW	b	0,8956	AW	b	0,9090
	roP	d	0,8061	roP	c	0,8596	roP	bc	0,8876
	PR	de	0,7899	PR	e	0,8403	PR	c	0,8825
	Stouffer	e	0,7775	Stouffer	e	0,8363	Stouffer	c	0,8823
	SR	f	0,7363	SR	f	0,7839	SR	e	0,8438
	maxP	a	0,6635	maxP	a	0,7052	maxP	a	0,7621
r = 3	Fisher	e	0,8977	Fisher	d	0,9396	Fisher	d	0,9781
	AW	b	0,8908	AW	b	0,9306	AW	d	0,9750
	minP	b	0,8878	minP	b	0,9266	minP	b	0,9631
	Stouffer	c	0,8350	roP	c	0,8918	PR	bc	0,9520
	roP	cd	0,8242	Stouffer	c	0,8877	Stouffer	bc	0,9518
	PR	df	0,8173	PR	ce	0,8746	roP	bc	0,9428
	SR	f	0,8101	SR	e	0,8731	SR	c	0,9326
	maxP	a	0,6873	maxP	a	0,7278	maxP	a	0,7647
r = 4	Fisher	b	0,9298	Fisher	d	0,9648	Fisher	d	0,9870
	minP	bc	0,9255	AW	b	0,9601	AW	c	0,9831
	AW	c	0,9229	minP	b	0,9594	Stouffer	c	0,9827
	roP	d	0,8807	roP	c	0,9369	roP	bc	0,9813
	Stouffer	d	0,8722	Stouffer	ce	0,9225	PR	bcd	0,9801
	PR	d	0,8653	PR	e	0,9184	SR	bc	0,9781
	SR	e	0,8479	SR	f	0,8993	minP	b	0,9766
	maxP	a	0,7083	maxP	a	0,7418	maxP	a	0,8821
r = 5	Fisher	b	0,9473	Fisher	d	0,9773	Fisher	c	0,9919
	minP	bc	0,9433	AW	b	0,9727	Stouffer	bc	0,9891
	AW	c	0,9403	minP	b	0,9689	AW	b	0,9885
	Stouffer	d	0,8952	Stouffer	c	0,9505	minP	b	0,9856
	roP	d	0,8853	roP	c	0,9461	PR	bc	0,9849
	PR	e	0,8599	PR	f	0,9309	roP	bc	0,9782
	SR	e	0,8517	SR	e	0,9133	SR	bc	0,9774
	maxP	a	0,7184	maxP	a	0,7720	maxP	a	0,9110

*Si dos o más técnicas de MA comparten la misma letra, se dice que no hay diferencia estadísticamente significativa en los respectivos valores del AUC.

Respecto a la técnica con peor desempeño, se observó que maxP es la que presentó menor valor del AUC en todos los escenarios (Gráfico 3). Al aplicar test de comparaciones múltiples se observó diferencia estadísticamente significativa del AUC entre maxP y las restantes técnicas de MA (Tabla 3).

4.4. Desempeño de las técnicas de MA cuando se combinan cinco estudios

Cuando se analizó el comportamiento de las técnicas de MA para el caso en que se combinaron los resultados provenientes de cinco estudios se observó que para todas las técnicas de MA y considerando todos los r , en general los valores del AUC fueron mayores cuando se trabajó bajo la configuración de hipótesis HS_A que cuando se hizo bajo HS_r o HS_B , observándose los menores valores del AUC bajo esta última configuración de hipótesis (Gráfico 4, Tabla 4).

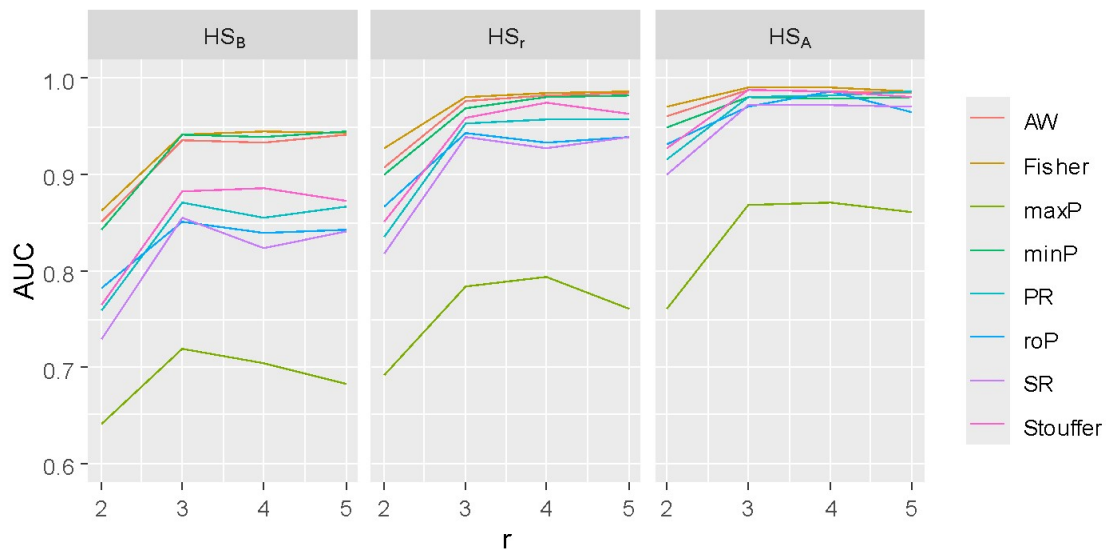


Gráfico 4: Valores del AUC en función de la cantidad de réplicas (r) según técnica de MA utilizada y para cada una de las configuraciones de hipótesis, cuando se combinan cinco estudios ($k = 5$)

Tabla 4: Valores del AUC y resultado del test de comparaciones múltiples entre las técnicas de MA para cada escenario presentado con $k = 5$

	HS _B			HS _r			HS _A		
	Técnica	Test*	AUC	Técnica	Test*	AUC	Técnica	Test*	AUC
$r = 2$	Fisher	d	0,8625	Fisher	d	0,9272	Fisher	d	0,9713
	AW	b	0,8508	AW	b	0,9093	AW	e	0,9610
	minP	b	0,8443	minP	b	0,9001	minP	b	0,9490
	roP	c	0,7832	roP	c	0,8673	roP	bc	0,9313
	Stouffer	ce	0,7650	Stouffer	ce	0,8508	Stouffer	bc	0,9282
	PR	e	0,7595	PR	ef	0,8363	PR	cf	0,9162
	SR	f	0,7294	SR	f	0,8176	SR	f	0,9003
	maxP	a	0,6402	maxP	a	0,6918	maxP	a	0,7611
$r = 3$	Fisher	b	0,9428	Fisher	e	0,9805	Fisher	c	0,9911
	minP	bc	0,9415	AW	g	0,9764	AW	d	0,9893
	AW	c	0,9350	minP	b	0,9703	Stouffer	cd	0,9887
	Stouffer	e	0,8833	Stouffer	bf	0,9594	minP	b	0,9821
	PR	e	0,8713	PR	df	0,9536	PR	bcd	0,9809
	SR	d	0,8557	roP	cd	0,9437	SR	b	0,9740
	roP	d	0,8508	SR	c	0,9395	roP	b	0,9717
	maxP	a	0,7198	maxP	a	0,7842	maxP	a	0,8699
$r = 4$	Fisher	b	0,9449	Fisher	d	0,9859	Fisher	c	0,9901
	minP	bc	0,9407	AW	b	0,9825	AW	d	0,9868
	AW	c	0,9343	minP	b	0,9806	Stouffer	bcd	0,9868
	Stouffer	e	0,8861	Stouffer	b	0,9747	roP	bcd	0,9867
	PR	d	0,8562	PR	e	0,9572	PR	bcd	0,9840
	roP	d	0,8398	roP	c	0,9335	minP	b	0,9802
	SR	f	0,8242	SR	c	0,9284	SR	bcd	0,9742
	maxP	a	0,7048	maxP	a	0,7948	maxP	a	0,8703
$r = 5$	minP	b	0,9458	Fisher	d	0,9876	Fisher	e	0,9875
	Fisher	b	0,9447	AW	b	0,9852	PR	de	0,9868
	AW	b	0,9428	minP	b	0,9825	AW	cd	0,9849
	Stouffer	d	0,8722	Stouffer	e	0,9642	minP	bcd	0,9822
	PR	d	0,8669	PR	e	0,9575	Stouffer	cde	0,9821
	roP	c	0,8430	roP	c	0,9401	SR	bc	0,9715
	SR	c	0,8421	SR	c	0,9390	roP	b	0,9647
	maxP	a	0,6813	maxP	a	0,7614	maxP	a	0,8617

*Si dos o más técnicas de MA comparten la misma letra, se dice que no hay diferencia estadísticamente significativa en los respectivos valores del AUC.

Considerando todas las combinaciones de hipótesis se observó que la técnica de MA que presentó mejor desempeño en casi todos los escenarios fue la técnica de Fisher. La excepción se presentó al trabajar con $r = 5$ bajo HS_B , en este caso la técnica que presentó el AUC más grande fue minP, pero al realizar el test de comparaciones múltiple, no se observó diferencia estadísticamente significativa en el AUC entre estas dos técnicas (Tabla 4).

Es interesante notar que cuando se trabajó bajo HS_B , en los casos en que el número de réplicas fue mayor a dos, no se observó diferencia estadísticamente significativa en los valores del AUC entre las técnicas de Fisher y minP (para el caso en que $r = 5$ tampoco se observó diferencia con el método AW). Mientras que cuando se trabajó bajo HS_A , en los casos en que $r = 3, 4$ ó 5 las técnicas que no presentaron diferencias estadísticamente significativas en sus valores del AUC fueron las técnicas de Fisher, Stouffer y PR (para $r = 4$ tampoco se observó diferencia con el método roP) (Tabla 4).

Respecto a la técnica con peor desempeño, se observó que el método maxP presentó valores del AUC considerablemente más pequeños en todos los escenarios. Esto fue confirmado por los resultados del test de comparaciones múltiples donde se observó diferencia estadísticamente significativa del AUC de esta técnica con las restantes (Gráfico 4, Tabla 4).

4.5. Desempeño de las técnicas de MA cuando se combinan seis estudios

Cuando se estudió el desempeño de las técnicas de MA en los casos de combinar los resultados provenientes de seis estudios ($k = 6$) se observó que en general para todas las configuraciones de hipótesis, hubo un aumento en los valores del AUC cuando se pasó de trabajar con dos réplicas por estudio a cuando se trabajó con 3, 4 ó 5 (Gráfico 5, Tabla 5).

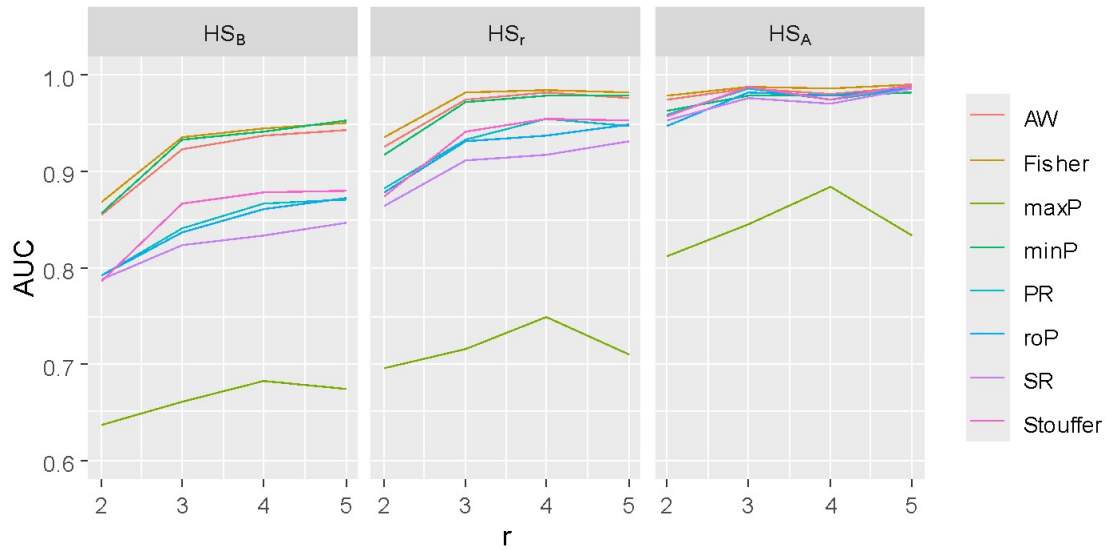


Gráfico 5: Valores del AUC en función de la cantidad de réplicas (r) según técnica de MA utilizada y para cada una de las configuraciones de hipótesis, cuando se combinan seis estudios ($k = 6$)

Además se pudo ver que para todas las técnicas de MA y considerando todos los r , en general los valores del AUC fueron mayores cuando se trabajó bajo la configuración de hipótesis HS_A que cuando se lo hizo bajo HS_r o HS_B observándose para esta última valores de AUC más pequeños. En este punto es interesante notar que cuando se trabajó bajo HS_A todas las técnicas de MA, excepto maxP, presentaron valores del AUC mayores a 0,94 (Tabla 5).

Considerando todas las combinaciones de hipótesis la técnica de MA que presentó mejor desempeño en casi todas las situaciones fue la técnica de Fisher. Se presentaron dos excepciones, ambas cuando se consideran experimentos con $r = 5$. Una ocurre bajo HS_B donde la técnica que presentó el valor del AUC más alto fue minP aunque este no resultó estadísticamente diferente del valor del AUC para el método de Fisher. La otra ocurre bajo HS_A , donde la técnica que presentó el valor del AUC más alto fue la del método de Stouffer aunque al realizar el test tampoco se observó diferencia estadísticamente significativa con respecto al de Fisher (Tabla 5).

Tabla 5: Valores del AUC y resultado del test de comparaciones múltiples entre las técnicas de MA para cada escenario presentado con $k = 6$

	HS _B			HS _r			HS _A		
	Técnica	Test*	AUC	Técnica	Test*	AUC	Técnica	Test*	AUC
$r = 2$	Fisher	d	0,8681	Fisher	d	0,9351	Fisher	d	0,9800
	minP	b	0,8574	AW	b	0,9259	AW	cd	0,9752
	AW	b	0,8556	minP	b	0,9191	minP	bc	0,9628
	PR	c	0,7928	PR	ce	0,8820	PR	bcd	0,9591
	roP	c	0,7922	roP	c	0,8795	Stouffer	bc	0,9568
	SR	c	0,7884	Stouffer	ce	0,8750	SR	bc	0,9545
	Stouffer	c	0,7866	SR	e	0,8644	roP	b	0,9480
	maxP	a	0,6369	maxP	a	0,6951	maxP	a	0,8124
$r = 3$	Fisher	b	0,9356	Fisher	d	0,9828	Fisher	c	0,9901
	minP	b	0,9331	AW	b	0,9757	Stouffer	cd	0,9884
	AW	e	0,9238	minP	b	0,9734	AW	d	0,9870
	Stouffer	d	0,8663	Stouffer	c	0,9420	PR	bcd	0,9862
	PR	c	0,8419	PR	c	0,9342	roP	bcd	0,9827
	roP	c	0,8379	roP	c	0,9321	minP	b	0,9789
	SR	c	0,8248	SR	e	0,9116	SR	bcd	0,9768
	maxP	a	0,6609	maxP	a	0,7163	maxP	a	0,8447
$r = 4$	Fisher	b	0,9461	Fisher	d	0,9858	Fisher	de	0,9864
	minP	bc	0,9419	AW	bd	0,9830	AW	bcde	0,9821
	AW	c	0,9382	minP	b	0,9784	minP	bc	0,9790
	Stouffer	e	0,8790	PR	e	0,9554	roP	bcde	0,9785
	PR	de	0,8676	Stouffer	e	0,9551	Stouffer	bcde	0,9763
	roP	d	0,8607	roP	c	0,9384	PR	ce	0,9759
	SR	f	0,8334	SR	f	0,9191	SR	bd	0,9705
	maxP	a	0,6823	maxP	a	0,7487	maxP	a	0,8847
$r = 5$	minP	b	0,9541	Fisher	d	0,9834	Stouffer	c	0,9919
	Fisher	b	0,9515	minP	b	0,9786	Fisher	c	0,9915
	AW	d	0,9436	AW	b	0,9779	AW	d	0,9892
	Stouffer	c	0,8807	Stouffer	c	0,9546	PR	bcd	0,9872
	roP	c	0,8729	roP	c	0,9501	SR	bcd	0,9867
	PR	c	0,8708	PR	c	0,9472	roP	bcd	0,9862
	SR	e	0,8469	SR	e	0,9311	minP	b	0,9829
	maxP	a	0,6736	maxP	a	0,7107	maxP	a	0,8344

*Si dos o más técnicas de MA comparten la misma letra, se dice que no hay diferencia estadísticamente significativa en los respectivos valores del AUC.

Es interesante notar que cuando se trabaja bajo HS_B , en los casos en que se consideró $r = 3, 4$ ó 5 , al comparar los valores del AUC, no se observó diferencia estadísticamente significativa entre Fisher y minP. Mientras que cuando se trabaja bajo HS_A , en estas mismas situaciones, las técnicas que no presentaron diferencias estadísticamente significativas en sus valores del AUC con el método de Fisher fueron las técnicas de Stouffer, PR, roP y SR. Se debe incluir en este último grupo a la técnica AW para el caso en que $r = 4$ (Tabla 5).

Respecto a la técnica con peor desempeño, se observó que el método maxP fue la que presentó menor valor del AUC y este fue estadísticamente diferente a las correspondientes AUC de las demás técnicas de MA en todos los escenarios (Tabla 5).

4.6. Desempeño de las técnicas de MA cuando se combinan siete estudios

Al evaluar el desempeño de las técnicas MA para el caso que se combinaron los resultados provenientes de siete estudios ($k = 7$) se observó que en general para todas las configuraciones de hipótesis, los valores del AUC aumentaron cuando se pasó de trabajar con dos réplicas por estudio a cuando se trabajó con 3, 4 ó 5 (Gráfico 6).

También se vio que para todas las técnicas de MA y considerando todos los valores de r , en general los valores del AUC fueron mayores cuando se trabajó bajo la configuración de hipótesis HS_A que cuando se hizo bajo HS_r o HS_B , observándose los menores valores del AUC bajo esta última (Gráfico 6). En este punto es interesante notar que cuando se trabaja bajo HS_A todas las técnicas de MA, excepto maxP, presentaron valores del AUC muy altos, mayores a 0,930. En especial para $r = 5$ los valores del AUC van de 0,985 a 0,994, excepto para maxP que presentó un valor del AUC igual a 0,901 (Tabla 6).

Considerando todas las combinaciones de hipótesis, la técnica de MA que presentó mejor desempeño en casi todas las situaciones fue la técnica de Fisher. Se presentaron dos excepciones. Una ocurre bajo HS_B cuando $r = 4$, donde el valor del AUC más alto lo presentó la técnica minP aunque este valor no fue estadísticamente diferente al observado para el método de Fisher. La otra excepción se observó cuando se trabajó bajo HS_A considerando $r = 5$, en este caso la técnica que presentó el valor del AUC más alto fue la de Stouffer, seguida

por el valor del AUC obtenido con el método PR que, aunque sus valores no presentaron diferencia estadísticamente significativa, sí lo hicieron con el valor del AUC obtenido con el método de Fisher (que presentó el tercer valor de AUC más alto) (Tabla 6).

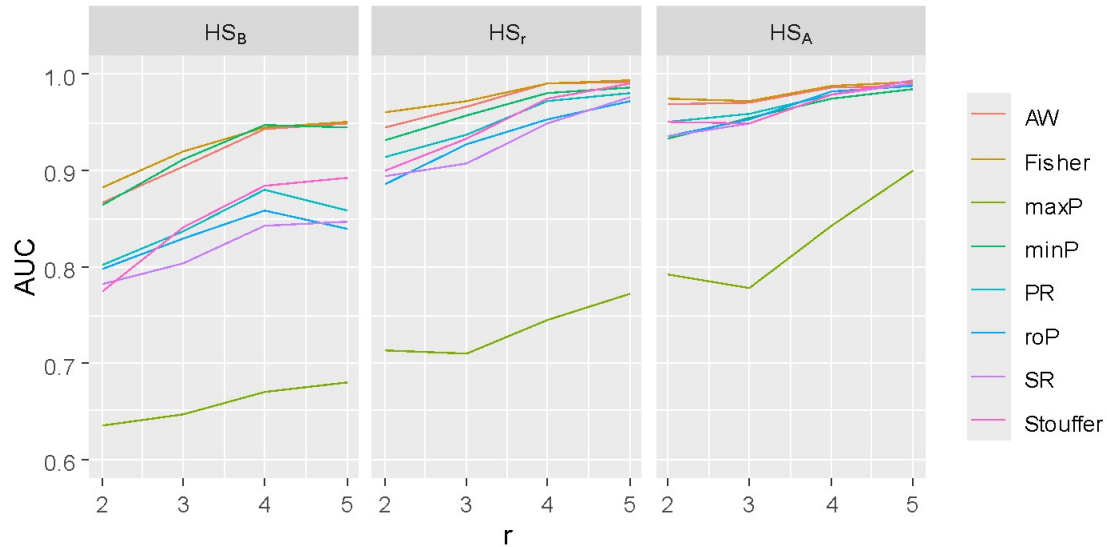


Gráfico 6: Valores del AUC en función de la cantidad de réplicas (r) según técnica de MA utilizada y para cada una de las configuraciones de hipótesis, cuando se combinan siete estudios ($k = 7$)

Respecto a la técnica con peor desempeño, se observó que el método maxP presentó el menor valor del AUC en todos los casos y además este resultó estadísticamente diferente al compararlo con las AUC de las restantes técnicas de MA en todos los escenarios considerados. (Tabla 6)

Tabla 6: Valores del AUC y resultado del test de comparaciones múltiples entre las técnicas de MA para cada escenario presentado con $k = 7$

	HS _B			HS _r			HS _A		
	Técnica	Test*	AUC	Técnica	Test*	AUC	Técnica	Test*	AUC
$r = 2$	Fisher	e	0,8819	Fisher	d	0,9624	Fisher	d	0,9762
	AW	b	0,8672	AW	f	0,9450	AW	cd	0,9703
	minP	b	0,8644	minP	b	0,9312	Stouffer	bc	0,9518
	PR	d	0,8026	PR	be	0,9151	PR	bcd	0,9512
	roP	cd	0,7979	Stouffer	ce	0,9011	roP	bc	0,9359
	SR	cf	0,7830	SR	c	0,8940	SR	bc	0,9353
	Stouffer	f	0,7750	roP	c	0,8866	minP	b	0,9335
	maxP	a	0,6352	maxP	a	0,7147	maxP	a	0,7930
$r = 3$	Fisher	d	0,9203	Fisher	d	0,9741	Fisher	d	0,9731
	minP	b	0,9116	AW	e	0,9678	AW	cd	0,9712
	AW	b	0,9052	minP	b	0,9575	PR	bcd	0,9602
	Stouffer	c	0,8416	PR	c	0,9373	minP	b	0,9554
	PR	c	0,8373	Stouffer	c	0,9334	roP	bc	0,9532
	roP	c	0,8289	roP	c	0,9272	Stouffer	bcd	0,9507
	SR	e	0,8051	SR	f	0,9087	SR	bc	0,9490
	maxP	a	0,6465	maxP	a	0,7090	maxP	a	0,7786
$r = 4$	minP	b	0,9477	Fisher	d	0,9920	Fisher	c	0,9885
	Fisher	b	0,9465	AW	e	0,9908	AW	d	0,9863
	AW	b	0,9446	minP	b	0,9820	roP	bcd	0,9824
	Stouffer	d	0,8840	Stouffer	b	0,9749	Stouffer	bcd	0,9793
	PR	d	0,8800	PR	b	0,9736	PR	bcd	0,9786
	roP	c	0,8586	roP	c	0,9538	SR	bcd	0,9784
	SR	e	0,8432	SR	c	0,9493	minP	b	0,9753
	maxP	a	0,6696	maxP	a	0,7445	maxP	a	0,8443
$r = 5$	Fisher	d	0,9509	Fisher	d	0,9951	Stouffer	d	0,9942
	AW	bd	0,9504	AW	e	0,9927	PR	d	0,9942
	minP	b	0,9460	Stouffer	e	0,9913	Fisher	c	0,9927
	Stouffer	e	0,8927	minP	b	0,9870	SR	cde	0,9917
	PR	f	0,8591	PR	bc	0,9822	AW	e	0,9897
	SR	cf	0,8482	SR	bc	0,9781	roP	bcde	0,9883
	roP	c	0,8398	roP	c	0,9729	minP	b	0,9851
	maxP	a	0,6804	maxP	a	0,7726	maxP	a	0,9013

*Si dos o más técnicas de MA comparten la misma letra, se dice que no hay diferencia estadísticamente significativa en los respectivos valores del AUC.

Hasta aquí se presentaron los resultados obtenidos al comparar los valores del AUC entre las diferentes técnicas de MA para cada uno de los 72 escenarios estudiados ($k \times r$) y bajo cada una de las tres configuraciones de hipótesis antes explicadas. En el Anexo B se presentan los resultados obtenidos de comparar entre los valores de AUC obtenidos bajo cada configuración de hipótesis, para cada una de las técnicas de MA y en cada uno de los escenarios bajo estudio.

5. Consideraciones finales

Como se mencionó en la introducción, las técnicas de MA suelen clasificarse en función de la información que utilizan como base para la construcción de la estadística de prueba. Los tres grupos principales son: métodos que combinan valores p , métodos que combinan rangos y métodos que combinan tamaños de efecto (Ramasamy *et al.*, 2008; Makinde *et al.*, 2021; Tseng *et al.*, 2012; Chang *et al.*, 2013). De las técnicas estudiadas en este trabajo, maxP, minP, roP, Fisher, Stouffer y AW pertenecen al primer grupo. Mientras que SR y PR pertenecen al grupo de métodos que combinan rangos, sin embargo, vale recordar que esos rangos son asignados a valores p . Así es que las técnicas aquí estudiadas se podrían decir que son aquellas que utilizan valores p como información inicial para su aplicación.

La elección de la mejor técnica de MA para la combinación de resultados de estudios provenientes de experimentos de microarreglos depende, entre otras cosas, de la hipótesis de trabajo (Chang *et al.*, 2013; Makinde *et al.*, 2021). Estos autores clasifican a las técnicas de MA dentro de cada set de hipótesis en función de las particularidades en la construcción de la estadística propia de cada método. Es así que cuando el objetivo de los investigadores se encuentra relacionado con hipótesis del tipo HS_A recomiendan utilizar maxP, SR o PR, cuando se hace bajo HS_r , aconsejan utilizar por ejemplo roP y cuando se plantean objetivos de investigación más en concordancia con HS_B recomiendan técnicas como Fisher, Stouffer, AW y minP. En este sentido, Chang *et al.* (2013) recomiendan también tener en cuenta las características propias de los estudios a incluir, si se espera una mayor heterogeneidad a partir de, por ejemplo, un protocolo experimental o si los estudios combinados provienen de diferentes tejidos, sugieren considerar el set de hipótesis HS_B . Por el contrario, si los estudios recopilados son relativamente homogéneos (por ejemplo, utilizan la misma plataforma de microarreglos o son del mismo laboratorio), recomienda trabajar bajo HS_r . Es por esto que resultó interesante estudiar el desempeño de cada una de las técnicas de MA bajo cada una de las combinaciones de hipótesis.

En el presente trabajo se observó que para cada una de las configuraciones de hipótesis a medida que aumenta la cantidad de estudios combinados, para igual número de réplicas, los valores de las AUC para cada una de las técnicas aumentan, indicando un mejor desempeño de las mismas. Lo mismo se observó,

en general, al aumentar el número de réplicas manteniendo constante la cantidad de estudios. Como es de esperarse, a medida que aumenta la cantidad de información, reflejado en un aumento de la cantidad de réplicas biológicas por experimentos y/o aumento en la cantidad de estudios a combinar, se observa un mejor desempeño de las técnicas de MA.

En un trabajo presentado por Chang *et al.* (2013) se recomienda elegir la técnica PR antes que maxP cuando se trabaja bajo HS_A y la técnica AW sobre Fisher y Stouffer cuando se trabaja bajo HS_B . Teniendo en cuenta los resultados obtenidos en el presente trabajo se puede decir que la técnica que presentó mejor desempeño en casi todos los escenarios y bajo las tres configuraciones de hipótesis estudiadas fue Fisher. En este sentido es importante resaltar que cuando se trabajó con valores pequeño de k y de r no se observaron grandes cambios en el orden de las técnicas de MA según sus AUC. Sin embargo, para los valores de k y r más grandes los resultados aquí presentados parecerían tener un comportamiento que tiende a lo publicado por los autores antes mencionados.

Por otro lado, si bien la técnica maxP presentó valores para el AUC mayores a 0,60 en todos los casos y valores más altos cuando se trabajó bajo la configuración de hipótesis HS_B , es la técnica que presentó peor desempeño en casi todas las situaciones. Esto se condice con lo observado por Song y Tseng (2014) que adjudican este fenómeno a que el planteo de las hipótesis HS_A y el método maxP son muy estrictos a la luz de la naturaleza generalmente ruidosa de los experimentos de microarreglos por lo que inevitablemente detectará muy pocos genes.

Para todas las técnicas el mejor desempeño se presentó cuando se trabajó bajo la combinación de hipótesis HS_A . En algunas situaciones se observó mayores valores del AUC para algunas técnicas cuando se trabajó bajo HS_r , sin embargo, en todos estos casos no se observó diferencia estadísticamente significativa con las AUC bajo HS_A (Anexo B)

Otra estrategia a considerar en estudios de MA es la propuesta por de Abreu Neto y Frei (2016). Estos autores establecen que como hoy en día no resulta complicado, ni lleva mucho tiempo aplicar las diferentes técnicas de MA, se podrían aplicar todas ellas y luego clasificar a un gen como DE en el MA en caso que la mayoría de las técnicas lo clasifiquen como tal. Esta estrategia fue la

utilizada en un trabajo de aplicación cuyo objetivo fue utilizar técnicas estadísticas de MA para detectar genes de cítricos involucrados en la interacción planta-patógeno (genes DE) en experimentos de infección por *Xanthomonas citri* pv. *citri* (Piskulic, Prunello y Daurelio, trabajo no publicado). En esta aplicación se combinó la información proveniente de cinco estudios utilizando las técnicas de MA minP, maxP, roP, Fisher, Stouffer, SR y PR. De los 7167 genes estudiados, 232 fueron clasificados como DE por lo menos con 5 de las 7 técnicas aplicadas (68 en todas las técnicas y 43 en 6 de ellas). Estos genes fueron considerados para un posterior análisis funcional.

Es importante notar que las conclusiones aquí presentadas están basadas en una única simulación para cada escenario y que en todos ellos la cantidad de réplicas consideradas es la misma entre los estudios combinados. Sería interesante extender este trabajo al análisis de situaciones en donde la cantidad de réplicas por estudio no fuera la misma, así como también realizar más de una simulación por escenario y volver a evaluar el comportamiento de cada una de las técnicas de MA.

6. Bibliografía

- Alberts, B., Heald, R., Hopkin, K., Johnson, A., Morgan, D., Roberts, K., Walter, P. (2023). *Essential cell biology* (6th ed.). W. W. Norton & Company
- Birnbaum, A. (1954). Combining Independent Tests of Significance. *Journal of the American Statistical Association*, 49(267), 559–574. doi:10.2307/2281130
- Burgueño, M.J., García-Bastos, J.L., González-Buitrago, J.M. (1995). Las curvas ROC en la evaluación de las pruebas diagnósticas. *Medicina clínica*, 104 (17) 661–670
- Campaign, A., Yang, Y. H. (2010). Comparison study of microarray meta-analysis methods. *BMC bioinformatics*, 11, 408. <https://doi.org/10.1186/1471-2105-11-408>
- Chang, L. C., Lin, H. M., Sibille, E., Tseng, G. C. (2013). Meta-analysis methods for combining multiple expression profiles: comparisons, statistical characterization and an application guideline. *BMC bioinformatics*, 14, 368. <https://doi.org/10.1186/1471-2105-14-368>
- Cui, X., Churchill, G. A. (2003). Statistical tests for differential expression in cDNA microarray experiments. *Genome Biology*, 4(4), 210. doi: 10.1186/gb-2003-4-4-210.
- Daurelio, L. D., Romero, M. S., Petrocelli, S., Merelo, P., Cortadi, A. A., Talón, M., Tadeo, F. R., Orellano, E. G. (2013). Characterization of *Citrus sinensis* transcription factors closely associated with the non-host response to *Xanthomonas Campestris* P.v. *vesicatoria*. *Journal of plant physiology*, 170(10), 934–942. <https://doi.org/10.1016/j.jplph.2013.01.011>
- Daurelio, L. D., Tondo, M. L., Romero, M. S., Merelo, P., Cortadi, A. A., Talón, M., Tadeo, F. R., Orellano, E. G. (2015). Novel insights into the *Citrus sinensis non host* response suggest photosynthesis decline, abiotic stress networks and secondary metabolism modifications. *Functional plant biology: FPB*, 42(8), 758–769. <https://doi.org/10.1071/FP14307>
- de Abreu Neto, J., Frei, M. (2016). Microarray Meta-Analysis Focused on the Response of Genes Involved in Redox Homeostasis to Diverse Abiotic Stresses in Rice. *Frontiers in Plant Science*. 6. 1–14. <https://doi.org/10.3389/fpls.2015.01260>.
- DeLong, E. R., DeLong, D. M., Clarke-Pearson, D. L. (1988). Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A

Nonparametric Approach. *Biometrics*, 44, 837–845.
<http://dx.doi.org/10.2307/2531595>

Forero, D. A. (2019). Available software for meta-analyses of genome-wide expression studies. *PeerJ Preprints* 7: e27708v1 <https://doi.org/10.7287/peerj.preprints.27708v1>

Gegzna, V. (2020). Biostat: Routines for Basic (Bio) Statistics_. R package version 0.0.20, commit 1868f7f57cd2869edfdd9f21ffe80fe49d5acb7f, <https://github.com/GegznaV/BioStat>.

Kontou, P. I., Pavlopoulou, A., Bagos, P. G. (2018). Methods of Analysis and Meta-Analysis for Identifying Differentially Expressed Genes. *Methods Mol Biol*, 1793, 183–210. https://doi.org/10.1007/978-1-4939-7868-7_12

Li, J., Tseng, G.C. (2011). An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *The Annals of Applied Statistics*, 5, 994–1019. <https://doi.org/10.1214/10-AOAS393>

Lodish, H. F., Berk, A., Kaiser, C., Krieger, M., Bretscher, A., Ploegh, H. L., Martin, K. C., Yaffe, M. B., & Amon, A. (2021). *Molecular cell biology* (9th ed.). Macmillan Learning.

Makinde, F. L., Tchamga, M. S. S., Jafari, J., Fatumo, S., Chimusa, E. R., Mulder, N., Mazandu, G. K. (2021). Reviewing and assessing existing meta-analysis models and tools. *Briefings in bioinformatics*, 22(6), bbab324. <https://doi.org/10.1093/bib/bbab324>

Moreno, V., Solé, X. (2004). Uso de chips de ADN (*microarrays*) en medicina: fundamentos técnicos y procedimientos básicos para el análisis estadístico de resultados. *Medicina Clínica (Barcelona)*; 122 (Supl 1):73–9

R Core Team (2024). R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>

Ramasamy, A., Mondry, A., Holmes, C. C., Altman, D. G. (2008). Key issues in conducting a metaanalysis of gene expression microarray datasets. *PLoS Med*, 5(9), e184. <https://doi.org/10.1371/journal.pmed.0050184>

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., Smyth, G.K. (2015). *limma* powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7), e47. <https://doi.org/10.1093/nar/gkv007>.

- Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J., Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12, p. 77. <https://doi.org/10.1186/1471-2105-12-77>
- Song, C., Tseng, G. C. (2014). Hypothesis setting and order statistics for robust genomic Meta-analysis. *The annals of applied statistics*, 8(2), 777–800. <https://doi.org/10.1214/13-aos683>
- Stouffer, S. A. (1949). A study of attitudes. *Scientific American Magazine*; 180(5): 11–5. <https://doi.org/10.1038/scientificamerican0549-11>
- Tippett, L. H. C. (1931). The methods of statistics: An introduction mainly for workers in the biological sciences. Williams & Norgate Ltd.
- Tseng, G. C., Ghosh, D., Feingold, E. (2012). Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic acids research*, 40(9), 3785–3799. <https://doi.org/10.1093/nar/gkr1265>
- Walsh, C. J., Hu, P., Batt, J., Santos, C. C. (2015). Microarray Meta-Analysis and Cross-Platform Normalization: Integrative Genomics for Robust Biomarker Discovery. *Microarrays (Basel)*, 4(3), 389–406. <https://doi.org/10.3390/microarrays4030389>
- Wang, X., Kang, D. D., Shen, K., Song, C., Lu, S., Chang, L. C., Liao, S. G., Huo, Z., Tang, S., Ding, Y., Kaminski, N., Sibille, E., Lin, Y., Li, J., Tseng, G. C. (2012). An R package suite for microarray meta-analysis in quality control, differentially expressed gene analysis and pathway enrichment detection. *Bioinformatics (Oxford, England)*, 28(19), 2534–2536. <https://doi.org/10.1093/bioinformatics/bts485>
- Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.
- Wilkinson B. (1951). A statistical consideration in psychological research. *Psychological Bulletin*, 48(2):156–158. <https://doi.org/10.1037/h0059111>

Anexo A

Se presenta a modo de ejemplo los pasos seguidos para la generación de los datos y su posterior análisis para el caso $k=4$ y $r=3$, siendo los demás casos análogos.

1. Utilizando como base el *script* provisto por Chang *et al.* (2013) se simularon intensidades de señal lumínica como las esperadas al realizar experimentos de microarreglos cuando se tienen 3 réplicas biológicas por experimento ($r = 3$). Este paso se realizó 4 veces por considerar $k = 4$.

2. Se procesaron las señales de cada una de las 4 situaciones anteriores utilizando el paquete LIMMA de Bioconductor (Ritchie, 2015). De este modo se obtuvo un valor de FC para cada gen que luego se sometió a un test estadístico específico para este tipo de experimentos. Como resultado final se obtuvo una matriz de valores p , para cada gen en cada uno de los 4 estudios simulados. Esto permitió clasificar a cada gen, como DE cuando el valor p fue menor a 0,05 o como no DE en caso contrario.

3. Se procedió a combinar la información proveniente de estos 4 estudios aplicando las técnicas minP, maxP, roP, Fisher, Stouffer, PR y SR del paquete MetaDE (Wang *et al.*, 2012) utilizando análisis de permutaciones para la obtención del valor p . Para la técnica de MA denominada AW se utilizó el algoritmo de Chang *et al.* (2013). Como resultado de este paso se obtuvo una matriz con 10.000 filas (una por cada gen) y 12 columnas (las 4 primeras correspondientes a si el gen fue clasificado como DE o no DE en cada uno de los estudios individuales simulados y las restantes 8 conteniendo los valores p obtenidos al aplicar cada una de las técnicas de MA)

4. Se calculó el área debajo de la correspondiente curva ROC para cada técnica y considerando cada una de las configuraciones de hipótesis (HS_A , HS_B y HS_r). El estándar de oro para cada uno de los pares de hipótesis se especificó utilizando la información de las 4 primeras columnas de la matriz mencionada en el punto anterior.

En forma análoga se procedió para las restantes combinaciones de k y r , obteniéndose los valores de AUC para cada técnica en cada uno de los 72 escenarios posibles. Esta información se utilizó para construir las tablas y los gráficos presentados en la sección Resultados.

Para construir los gráficos 1 a 6 se utilizó el paquete ggplot2 (Wickham, 2016). Los datos se encuentran en un archivo de Excel y están organizados en 73 filas (la primera de ellas correspondiente a rótulos, las 72 restantes a cada uno de los escenarios en estudio) y 12 columnas. La información contenida en cada columna se detalla a continuación:

“ k ”: cantidad de estudios combinados (toma valores del 2 a 7)

“ r ”: cantidad de réplicas por estudio (toma valores del 2 al 5)

“verdad”: hace referencia al par de hipótesis considerada (las posibles categorías son: “Uno”, “Mitad” y “Todos” correspondientes a los pares de hipótesis HS_B , HS_r y HS_A a respectivamente)

“maxP”, “minP”, “roP”, “Fisher”, “Stouffer”, “AW”, “SR”, “PR”: Contiene valores de las AUC para cada uno de los escenarios bajo estudio.

“medias”: contiene el valor medio de las AUC entre todas las técnicas para cada uno de los escenarios considerados.

A modo de ejemplo se transcribe el *script* correspondiente al **Gráfico 3**.

```
## PAQUETES ##
```

```
library(readxl)
```

```
library(dplyr)
```

```
library(ggplot2)
```

```
library(tidyr)
```

```
## DATOS ##
```

```
tablas<- read_excel("AUC_2025_conMedias.xlsx")
```

```

# Grafico para k=4
tablask4 <- tablas[tablas$k == "4", ]

tablask4$verdad<-factor(tablask4$verdad,
levels=c("Uno","Mitad","Todos"),
labels= c("HSB","HSr","HSA"))

tablask4_graf <- gather(tablask4, key = "metodoMA", value = "auc", -1, -2, -3, -12)

levels(tablask4_graf$verdad) <- c("HS[B]","HS[r]", "HS[A]")
ggplot(tablask4_graf, aes(x=r, y=auc, color = metodoMA)) +
  geom_line() +
  facet_wrap(~ verdad,labeller = label_parsed) +
  ylab("AUC")+
  scale_y_continuous(limits = c(0.65, 1))+
  theme(legend.title = element_blank())

```

Anexo B

Se presentan los resultados obtenidos al comparar las AUC de las distintas configuraciones de hipótesis para cada técnica de MA en cada uno de los 24 ($k \times r$) escenarios propuestos. Vale recordar que para los escenarios en donde la cantidad de estudios a combinar es igual a dos ($k = 2$), trabajar bajo los pares de hipótesis HS_B y HS_r resulta equivalente; y además, dada la naturaleza de las técnicas minP y roP, los valores de las AUC obtenidos con ambas son iguales para este caso.

Tabla B1: Valores del AUC y resultado del test de comparaciones entre configuraciones de hipótesis para cada técnica de MA cuando se combinan dos estudios

	Hipótesis	Valores del AUC (Test*)						
		maxP	minP/roP	Fisher	Stouffer	AW	SR	PR
$r=2$	HS_B/HS_r	0,7172 (a)	0,8411 (a)	0,8430 (a)	0,7630 (a)	0,8414 (a)	0,7796 (a)	0,7637 (a)
	HS_A	0,8133 (b)	0,8780 (a)	0,8879 (b)	0,8287 (b)	0,8831 (a)	0,8502 (b)	0,8225 (b)
$r=3$	HS_B/HS_r	0,7357 (a)	0,8443 (a)	0,8471 (a)	0,7848 (a)	0,8408 (a)	0,7647 (a)	0,7721 (a)
	HS_A	0,8444 (b)	0,9158 (b)	0,9216 (b)	0,8756 (b)	0,9187 (b)	0,8633 (b)	0,8611 (b)
$r=4$	HS_B/HS_r	0,7613 (a)	0,9142 (a)	0,9163 (a)	0,8480 (a)	0,9117 (a)	0,8344 (a)	0,7803 (a)
	HS_A	0,8709 (b)	0,9518 (b)	0,9605 (b)	0,9252 (b)	0,9557 (b)	0,9126 (b)	0,8655 (b)
$r=5$	HS_B/HS_r	0,7644 (a)	0,9101 (a)	0,9092 (a)	0,8540 (a)	0,9091 (a)	0,8015 (a)	0,8133 (a)
	HS_A	0,8755 (b)	0,9607 (b)	0,9614 (b)	0,9226 (b)	0,9602 (b)	0,9034 (b)	0,8948 (b)

*Si las configuraciones de hipótesis comparten la misma letra, se dice que no hay diferencia estadísticamente significativa en los respectivos valores del AUC.

Tabla B2: Valores del AUC y resultado del test de comparaciones múltiples entre configuraciones de hipótesis para cada técnica de MA cuando se combinan tres estudios

	Hipótesis	Valores del AUC (Test*)							
		maxP	minP	roP	Fisher	Stouffer	AW	SR	PR
r=2	HS_B	0,6781 (a)	0,8373 (a)	0,7702 (a)	0,8456 (a)	0,7577 (a)	0,8347 (a)	0,7794 (a)	0,7633 (a)
	HS_r	0,7244 (b)	0,8825 (b)	0,8303 (b)	0,8929 (b)	0,8165 (b)	0,8801 (b)	0,8369 (b)	0,8188 (b)
	HS_A	0,7898 (c)	0,9100 (b)	0,8853 (c)	0,9242 (b)	0,8783 (c)	0,9167 (c)	0,8887 (c)	0,8700 (c)
r=3	HS_B	0,7091 (a)	0,8701 (a)	0,8012 (a)	0,8747 (a)	0,7964 (a)	0,8698 (a)	0,7698 (a)	0,7858 (a)
	HS_r	0,7660 (b)	0,9260 (b)	0,8806 (b)	0,9358 (b)	0,8736 (b)	0,9282 (b)	0,8452 (b)	0,8621 (b)
	HS_A	0,7937 (b)	0,9411 (b)	0,9042 (b)	0,9476 (b)	0,9054 (b)	0,9442 (b)	0,8772 (b)	0,8958 (b)
r=4	HS_B	0,7137 (a)	0,9160 (a)	0,8295 (a)	0,9217 (a)	0,8485 (a)	0,9146 (a)	0,7927 (a)	0,7742 (a)
	HS_r	0,7699 (b)	0,9440 (b)	0,9152 (b)	0,9541 (b)	0,9093 (b)	0,9481 (b)	0,8721 (b)	0,8627 (b)
	HS_A	0,8716 (c)	0,9589 (b)	0,9645 (c)	0,9733 (c)	0,9554 (c)	0,9666 (b)	0,9391 (c)	0,9246 (c)
r=5	HS_B	0,7152 (a)	0,9369 (a)	0,8497 (a)	0,9380 (a)	0,8793 (a)	0,9340 (a)	0,8073 (a)	0,8325 (a)
	HS_r	0,7707 (b)	0,9649 (b)	0,9204 (b)	0,9692 (b)	0,9298 (b)	0,9694 (b)	0,8782 (b)	0,8958 (b)
	HS_A	0,8478 (c)	0,9763 (b)	0,9515 (c)	0,9843 (c)	0,9556 (b)	0,9826 (c)	0,9258 (c)	0,9331 (c)

*Si dos o más configuraciones de hipótesis comparten la misma letra, se dice que no hay diferencia estadísticamente significativa en los respectivos valores del AUC.

Tabla B3: Valores del AUC y resultado del test de comparaciones múltiples entre configuraciones de hipótesis para cada técnica de MA cuando se combinan cuatro estudios

	Hipótesis	Valores del AUC (Test*)							
		maxP	minP	roP	Fisher	Stouffer	AW	SR	PR
r=2	HS_B	0,6635 (a)	0,8613 (a)	0,8061 (a)	0,8659 (a)	0,7775 (a)	0,8530 (a)	0,7363 (a)	0,7899 (a)
	HS_r	0,7052 (b)	0,9002 (b)	0,8596 (b)	0,9119 (b)	0,8363 (b)	0,8956 (b)	0,7839 (b)	0,8403 (b)
	HS_A	0,7621 (c)	0,9156 (b)	0,8876 (b)	0,9300 (b)	0,8823 (b)	0,9090 (b)	0,8438 (c)	0,8825 (c)
r=3	HS_B	0,6873 (a)	0,8878 (a)	0,8242 (a)	0,8977 (a)	0,8350 (a)	0,8908 (a)	0,8101 (a)	0,8173 (a)
	HS_r	0,7278 (a)	0,9266 (b)	0,8918 (b)	0,9396 (b)	0,8877 (b)	0,9306 (b)	0,8731 (b)	0,8746 (b)
	HS_A	0,7647 (a)	0,9631 (c)	0,9428 (c)	0,9781 (c)	0,9518 (c)	0,9750 (c)	0,9326 (c)	0,9520 (c)
r=4	HS_B	0,7083 (a)	0,9255 (a)	0,8807 (a)	0,9298 (a)	0,8722 (a)	0,9229 (a)	0,8479 (a)	0,8653 (a)
	HS_r	0,7418 (a)	0,9594 (b)	0,9369 (b)	0,9648 (b)	0,9225 (b)	0,9601 (b)	0,8993 (b)	0,9184 (b)
	HS_A	0,8821 (b)	0,9766 (c)	0,9813 (c)	0,9870 (c)	0,9827 (c)	0,9831 (c)	0,9781 (c)	0,9801 (c)
r=5	HS_B	0,7184 (a)	0,9433 (a)	0,8853 (a)	0,9473 (a)	0,8952 (a)	0,9403 (a)	0,8517 (a)	0,8599 (a)
	HS_r	0,7720 (b)	0,9689 (b)	0,9461 (b)	0,9773 (b)	0,9505 (b)	0,9727 (b)	0,9133 (b)	0,9309 (b)
	HS_A	0,9110 (c)	0,9856 (c)	0,9782 (c)	0,9919 (c)	0,9891 (c)	0,9885 (c)	0,9774 (c)	0,9849 (c)

*Si dos o más configuraciones de hipótesis comparten la misma letra, se dice que no hay diferencia estadísticamente significativa en los respectivos valores del AUC.

Tabla B4: Valores del AUC y resultado del test de comparaciones múltiples entre configuraciones de hipótesis para cada técnica de MA cuando se combinan cinco estudios

	Hipótesis	Valores del AUC (Test*)							
		maxP	minP	roP	Fisher	Stouffer	AW	SR	PR
r=2	HS_B	0,6402 (a)	0,8443 (a)	0,7832 (a)	0,8625 (a)	0,7650 (a)	0,8508 (a)	0,7294 (a)	0,7595 (a)
	HS_r	0,6918 (b)	0,9001 (b)	0,8673 (b)	0,9272 (b)	0,8508 (b)	0,9093 (b)	0,8176 (b)	0,8363 (b)
	HS_A	0,7611 (c)	0,9490 (c)	0,9313 (c)	0,9713 (c)	0,9282 (c)	0,9610 (c)	0,9003 (c)	0,9162 (c)
r=3	HS_B	0,7198 (a)	0,9415 (a)	0,8508 (a)	0,9428 (a)	0,8833 (a)	0,9350 (a)	0,8557 (a)	0,8713 (a)
	HS_r	0,7842 (b)	0,9703 (b)	0,9437 (b)	0,9805 (b)	0,9594 (b)	0,9764 (b)	0,9395 (b)	0,9536 (b)
	HS_A	0,8699 (c)	0,9821 (c)	0,9717 (c)	0,9911 (c)	0,9887 (c)	0,9893 (c)	0,9740 (c)	0,9809 (c)
r=4	HS_B	0,7048 (a)	0,9407 (a)	0,8398 (a)	0,9449 (a)	0,8861 (a)	0,9343 (a)	0,8242 (a)	0,8562 (a)
	HS_r	0,7948 (b)	0,9806 (b)	0,9335 (b)	0,9859 (b)	0,9747 (b)	0,9825 (b)	0,9284 (b)	0,9572 (b)
	HS_A	0,8703 (c)	0,9802 (b)	0,9867 (c)	0,9901 (b)	0,9868 (b)	0,9868 (c)	0,9742 (c)	0,9840 (c)
r=5	HS_B	0,6813 (a)	0,9458 (a)	0,8430 (a)	0,9447 (a)	0,8722 (a)	0,9428 (a)	0,8421 (a)	0,8669 (a)
	HS_r	0,7614 (b)	0,9825 (b)	0,9401 (b)	0,9876 (b)	0,9642 (b)	0,9852 (b)	0,9390 (b)	0,9575 (b)
	HS_A	0,8617 (c)	0,9822 (b)	0,9647 (b)	0,9875 (b)	0,9821 (b)	0,9849 (b)	0,9715 (c)	0,9868 (c)

*Si dos o más configuraciones de hipótesis comparten la misma letra, se dice que no hay diferencia estadísticamente significativa en los respectivos valores del AUC.

Tabla B5: Valores del AUC y resultado del test de comparaciones múltiples entre configuraciones de hipótesis para cada técnica de MA cuando se combinan seis estudios

	Hipótesis	Valores del AUC (Test*)							
		maxP	minP	roP	Fisher	Stouffer	AW	SR	PR
r=2	HS_B	0,6369 (a)	0,8574 (a)	0,7922 (a)	0,8681 (a)	0,7866 (a)	0,8556 (a)	0,7884 (a)	0,7928 (a)
	HS_r	0,6951 (b)	0,9191 (b)	0,8795 (b)	0,9351 (b)	0,8750 (b)	0,9259 (b)	0,8644 (b)	0,8820 (b)
	HS_A	0,8124 (c)	0,9628 (c)	0,9480 (c)	0,9800 (c)	0,9568 (c)	0,9752 (c)	0,9545 (c)	0,9591 (c)
r=3	HS_B	0,6609 (a)	0,9331 (a)	0,8379 (a)	0,9356 (a)	0,8663 (a)	0,9238 (a)	0,8248 (a)	0,8419 (a)
	HS_r	0,7163 (b)	0,9734 (b)	0,9321 (b)	0,9828 (b)	0,9420 (b)	0,9757 (b)	0,9116 (b)	0,9342 (b)
	HS_A	0,8447 (c)	0,9789 (b)	0,9827 (c)	0,9901 (b)	0,9884 (c)	0,9870 (c)	0,9768 (c)	0,9862 (c)
r=4	HS_B	0,6823 (a)	0,9419 (a)	0,8607 (a)	0,9461 (a)	0,8790 (a)	0,9382 (a)	0,8334 (a)	0,8676 (a)
	HS_r	0,7487 (b)	0,9784 (b)	0,9384 (b)	0,9858 (b)	0,9551 (b)	0,9830 (b)	0,9191 (b)	0,9554 (b)
	HS_A	0,8847 (c)	0,9790 (b)	0,9785 (c)	0,9864 (b)	0,9763 (b)	0,9821 (b)	0,9705 (c)	0,9759 (b)
r=5	HS_B	0,6736 (a)	0,9541 (a)	0,8729 (a)	0,9515 (a)	0,8807 (a)	0,9436 (a)	0,8469 (a)	0,8708 (a)
	HS_r	0,7107 (a)	0,9786 (b)	0,9501 (b)	0,9834 (b)	0,9546 (b)	0,9779 (b)	0,9311 (b)	0,9472 (b)
	HS_A	0,8344 (b)	0,9829 (b)	0,9862 (c)	0,9915 (b)	0,9919 (c)	0,9892 (c)	0,9867 (c)	0,9872 (c)

*Si dos o más configuraciones de hipótesis comparten la misma letra, se dice que no hay diferencia estadísticamente significativa en los respectivos valores del AUC.

Tabla B6: Valores del AUC y resultado del test de comparaciones múltiples entre configuraciones de hipótesis para cada técnica de MA cuando se combinan siete estudios

	Hipótesis	Valores del AUC (Test*)							
		maxP	minP	roP	Fisher	Stouffer	AW	SR	PR
r=2	HS_B	0,6352 (a)	0,8644 (a)	0,7979 (a)	0,8819 (a)	0,7750 (a)	0,8672 (a)	0,7830 (a)	0,8026 (a)
	HS_r	0,7147 (b)	0,9312 (b)	0,8866 (b)	0,9624 (b)	0,9011 (b)	0,9450 (b)	0,8940 (b)	0,9151 (b)
	HS_A	0,7930 (c)	0,9335 (b)	0,9359 (c)	0,9762 (b)	0,9518 (c)	0,9703 (c)	0,9353 (b)	0,9512 (b)
r=3	HS_B	0,6465 (a)	0,9116 (a)	0,8289 (a)	0,9202 (a)	0,8416 (a)	0,9052 (a)	0,8051 (a)	0,8373 (a)
	HS_r	0,7090 (b)	0,9575 (b)	0,9272 (b)	0,9741 (b)	0,9334 (b)	0,9678 (b)	0,9086 (b)	0,9373 (b)
	HS_A	0,7786 (b)	0,9554 (b)	0,9532 (b)	0,9731 (b)	0,9506 (b)	0,9712 (b)	0,9490 (c)	0,9602 (b)
r=4	HS_B	0,6696 (a)	0,9477 (a)	0,8586 (a)	0,9465 (a)	0,8840 (a)	0,9446 (a)	0,8432 (a)	0,8800 (a)
	HS_r	0,7445 (b)	0,9820 (b)	0,9537 (b)	0,9920 (b)	0,9749 (b)	0,9908 (b)	0,9493 (b)	0,9736 (b)
	HS_A	0,8443 (c)	0,9753 (b)	0,9824 (c)	0,9885 (b)	0,9792 (b)	0,9863 (b)	0,9784 (c)	0,9786 (b)
r=5	HS_B	0,6803 (a)	0,9459 (a)	0,8398 (a)	0,9509 (a)	0,8927 (a)	0,9504 (a)	0,8482 (a)	0,8591 (a)
	HS_r	0,7726 (b)	0,9870 (b)	0,9729 (b)	0,9951 (b)	0,9913 (b)	0,9927 (b)	0,9781 (b)	0,9822 (b)
	HS_A	0,9013 (c)	0,9851 (b)	0,9883 (c)	0,9927 (b)	0,9942 (b)	0,9897 (b)	0,9917 (c)	0,9942 (c)

*Si dos o más configuraciones de hipótesis comparten la misma letra, se dice que no hay diferencia estadísticamente significativa en los respectivos valores del AUC.