

Assessing the properties of asymptotic PERMANOVA test through comprehensive simulations in the context of genetic studies



Universitat Oberta
de Catalunya

Aitor Invernón de Campos

Análisis de datos Ómicos

Máster universitario de Bioinformática y
Bioestadística (UOC, UB)

Miquel Calvo Llorca
Diego Garrido-Martín

David Merino Arranz

17/01/2024



UNIVERSITAT DE
BARCELONA



Assessing the properties of asymptotic PERMANOVA test through comprehensive simulations in the context of genetic studies ©2023 by **Aitor Invernón de Campos** is licensed under Attribution 4.0 International.

To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

Ficha Del Trabajo Final

Título del trabajo:	Assessing the properties of asymptotic PERMANOVA test through comprehensive simulations in the context of genetic studies
Nombre del autor/a:	Aitor Invernón de Campos
Nombre del tutor/a de TF:	Miquel Calvo Llorca Diego Garrido-Martín
Nombre del/de la PRA:	David Merino Arranz
Fecha de entrega:	17/01/2024
Titulación o programa:	Máster universitario de Bioinformática y Bioestadística (UOC, UB)
Área del trabajo final:	Análisis de datos Ómicos
Idioma del trabajo:	Castellano
Palabras clave:	Multivariate statistics; MANOVA; Asymptotic theory; PERMANOVA; Parametric; Non-parametric; Distance measure; Quadratic forms; Experimental design; Data visualization; Partitioning; Permutation tests; Canonical analysis; Resampling; Alternative Splicing; GWAS; RNA-Seq; QTL; sQTLs

Resumen del trabajo

...

Abstract

...

Índice general

Índice general	1
Índice de figuras	3
Índice de tablas	4
0 PEC3 - Desarrollo del trabajo - Fase 2	8
0.1 Identificación del trabajo y fecha del informe	8
0.2 Avance del proyecto	8
0.3 Relación de las actividades realizadas	8
0.4 Desviaciones y acciones de mitigación	10
0.5 Resultados parciales obtenidos	10
0.6 Comentarios de los directores del TFM	11
1 Introducción	14
1.1 Contexto y justificación del trabajo	14
1.2 Objetivos del trabajo	16
1.3 Enfoque y método seguido	17
1.4 Planificación del trabajo	18
1.5 Hitos	20
1.6 Desviaciones en la planificación y acciones de mitigación	22
1.7 Análisis de riesgos	22
1.8 Breve resumen de productos obtenidos	23
1.9 Comentarios de los directores del TFM	23
1.10 Descripción de otros capítulos	23
2 Estado del arte	24
2.1 Contexto biotecnológico	24
2.2 Estadística multivariante aplicada a estudios <i>GWAS</i> basados en datos <i>RNA-seq</i>	27
2.2.1 Métodos univariantes y bivalentes	27
2.2.2 Métodos multivariantes: características y aplicaciones	28
2.3 El modelo <i>PERMANOVA</i>	29
2.4 MANTA - Una implementación de la versión asintótica y no paramétrica de <i>PERMANOVA</i>	29
3 Metodología y Resultados	30
3.1 Objetivos finales del estudio	30
3.2 Metodología	31
3.3 Exposición de los resultados obtenidos	34
3.3.1 Resultados del primer objetivo	34
3.3.2 Resultados del segundo objetivo	42
4 Discusión	48
5 Conclusiones y trabajos futuros	49
5.1 Conclusiones	49
5.2 Líneas de futuro	49
5.3 Seguimiento de la planificación	49
Glosario	50
Referencias	53

A Anexo de tablas

I

B Anexo de figuras

VII

Índice de figuras

1	Planificación ideada de las tareas necesarias para la consecución de la memoria y presentación del presente TFM.	12
1.1	Planificación ideada de las tareas necesarias para la consecución de la memoria y presentación del presente TFM.	19
3.1	OBJ1mynormStatsHomog	36
3.2	A.	37
3.3	A.	38
3.4	A.	39
3.5	A.	40
3.6	A.	41
3.7	OBJ1mynormStatsHomog	44
3.8	A.	45
3.9	A.	45
3.10	A.	46
3.11	A.	46
3.12	A.	47
3.13	A.	47
B.1	OBJ1mynormStatsHomog	VII
B.2	A.	VIII
B.3	A.	IX
B.4	A.	X
B.5	A.	XI
B.6	A.	XII
B.7	A.	XIII
B.8	A.	XIII
B.9	A.	XIV
B.10	A.	XIV
B.11	A.	XV
B.12	A.	XV
B.13	A.	XVI
B.14	A.	XVI

Índice de tablas

3.1	Simulaciones comparativas MANTA-MANOVA bajo el modelo de distribución <i>mvnorm</i> (<i>Objetivo I</i>), calculando la potencia estadística \mathbb{P} bajo un nivel de significación $\alpha = 0.05$ y con: $S = 1000$; $n = 300$; $q = 3$. . .	32
3.2	Simulaciones comparativas MANTA-MANOVA bajo el modelo de distribución <i>mvnorm</i> (<i>Objetivo I</i>), calculando la potencia estadística \mathbb{P} bajo unos niveles de significación estadística menores ($\alpha \in [0.01, 0.001]$) y con: $S = 1000$; $n = 300$; $q = 3$. . .	33
3.3	Simulaciones para el estudio de la posible invarianza frente a la transformación de los datos del método asintótico <i>PERMANOVA</i> , implementado en <i>MANTA</i> , con respecto a su potencia estadística (\mathbb{P}). Teniendo en cuenta diferentes situaciones de simulación del conjunto de datos, mediante el uso de un <i>algoritmo simplex</i> con $n = 3$. . .	33
3.4	Muestra aleatoria de 15 de los 5040 resultados obtenidos para los cálculos de la potencia estadística \mathbb{P} bajo las condiciones detalladas en las simulaciones 3.1a y 3.2a. . .	34
3.5	Muestra aleatoria de 15 de los 1008 resultados obtenidos para los cálculos de la potencia estadística \mathbb{P} bajo las condiciones detalladas en las simulaciones 3.1b y 3.2b. . .	34
3.6	Descripción de los estadísticos potencia (\mathbb{P}) y tiempo de computación ($t_{comp.}$), para los métodos MAN-TA y MANOVA , bajo una distribución <i>mvnorm</i> , con una matriz de correlación <i>homogénea</i> , y considerando diferentes niveles de significación. . .	35
3.7	Descripción de los estadísticos potencia (\mathbb{P}) y tiempo de computación ($t_{comp.}$), para los métodos MANTA y MANOVA , bajo una distribución <i>mvnorm</i> , con una matriz de correlación <i>inhomogénea</i> , y considerando diferentes niveles de significación. . .	37
3.8	Muestra aleatoria de 15 de los 2144 resultados obtenidos para los cálculos de la potencia estadística \mathbb{P} bajo las condiciones detalladas en las simulaciones 3.3a y 3.3b. . .	42
3.9	Sumario estadístico (para diferentes valores del nivel de significación α) de la potencia estadística (\mathbb{P}) calculada, y del tiempo de computación empleado en las simulaciones <i>3-simplex</i> , sin aplicar al conjunto de datos ninguna transformación. . .	43
3.10	Sumario estadístico (para diferentes valores del nivel de significación α) de la potencia estadística (\mathbb{P}) calculada, y el tiempo de computación empleado en las simulaciones <i>3-simplex</i> , aplicando al conjunto de datos una transformación logarítmica. . .	43
3.11	Sumario estadístico (para diferentes valores del nivel de significación α) de la potencia estadística (\mathbb{P}) calculada, y el tiempo de computación empleado en las simulaciones <i>3-simplex</i> , aplicando al conjunto de datos una transformación <i>Centered Log Ratio</i> (<i>clr</i>). . .	43
3.12	Sumario estadístico (para diferentes valores del nivel de significación α) de la potencia estadística (\mathbb{P}) calculada, y el tiempo de computación empleado en las simulaciones <i>3-simplex</i> , aplicando al conjunto de datos una transformación de <i>raíz cuadrada</i> (<i>sqr</i> t). . .	43
A.1	Simulaciones comparativas MANTA-MANOVA bajo el modelo de distribución <i>mvnorm</i> (<i>Objetivo I</i>), calculando la potencia estadística \mathbb{P} bajo un nivel de significación $\alpha = 0.05$ y con: $S = 1000$; $n = 300$; $q = 3$. . .	I

A.2	Simulaciones comparativas MANTA-MANOVA bajo el modelo de distribución <i>mvnorm</i> (<i>Objetivo I</i>), calculando la potencia estadística \mathbb{P} bajo unos niveles de significación estadística menores ($\alpha \in [0.01, 0.001]$) y con: $S = 1000$; $n = 300$; $q = 3$	I
A.3	Muestra aleatoria de 15 de los 5040 resultados obtenidos para los cálculos de la potencia estadística \mathbb{P} bajo las condiciones detalladas en las simulaciones A.1a y A.2a.	II
A.4	Muestra aleatoria de 15 de los 1008 resultados obtenidos para los cálculos de la potencia estadística \mathbb{P} bajo las condiciones detalladas en las simulaciones A.1b y A.2b.	II
A.5	Descripción de los estadísticos potencia (\mathbb{P}) y tiempo de computación (<i>t comp.</i>), para el modelo MANTA , bajo una distribución <i>mvnorm</i> , con una matriz de correlación <i>homogénea</i> , y considerando diferentes niveles de significación.	II
A.6	Descripción de los estadísticos potencia (\mathbb{P}) y tiempo de computación (<i>t comp.</i>), para el modelo MANOVA , bajo una distribución <i>mvnorm</i> , con una matriz de correlación <i>homogénea</i> , y considerando diferentes niveles de significación.	III
A.7	Descripción de los estadísticos potencia (\mathbb{P}) y tiempo de computación (<i>t comp.</i>), para el modelo MANTA , bajo una distribución <i>mvnorm</i> , con una matriz de correlación <i>inhomogénea</i> , y considerando diferentes niveles de significación.	III
A.8	Descripción de los estadísticos potencia (\mathbb{P}) y tiempo de computación (<i>t comp.</i>), para el modelo MANOVA , bajo una distribución <i>mvnorm</i> , con una matriz de correlación <i>inhomogénea</i> , y considerando diferentes niveles de significación.	IV
A.9	Descripción de los estadísticos potencia (\mathbb{P}) y tiempo de computación (<i>t comp.</i>), para los métodos MANTA y MANOVA , bajo una distribución <i>mvnorm</i> , con una matriz de correlación <i>homogénea</i> , y considerando diferentes niveles de significación.	IV
A.10	Descripción de los estadísticos potencia (\mathbb{P}) y tiempo de computación (<i>t comp.</i>), para los métodos MANTA y MANOVA , bajo una distribución <i>mvnorm</i> , con una matriz de correlación <i>inhomogénea</i> , y considerando diferentes niveles de significación.	IV
A.11	Simulaciones para el estudio de la posible invarianza frente a la transformación de los datos del método asintótico PERMANOVA , implementado en MANTA , con respecto a su potencia estadística (\mathbb{P}). Teniendo en cuenta diferentes situaciones de simulación del conjunto de datos, mediante el uso de un <i>algoritmo simplex</i> con $n = 3$	V
A.12	Muestra aleatoria de 15 de los 2144 resultados obtenidos para los cálculos de la potencia estadística \mathbb{P} bajo las condiciones detalladas en las simulaciones A.11a y A.11b.	V
A.13	Sumario estadístico (para diferentes valores del nivel de significación α) de la potencia estadística (\mathbb{P}) calculada, y del tiempo de computación empleado en las simulaciones <i>3-simplex</i> , sin aplicar al conjunto de datos ninguna transformación.	V
A.14	Sumario estadístico (para diferentes valores del nivel de significación α) de la potencia estadística (\mathbb{P}) calculada, y el tiempo de computación empleado en las simulaciones <i>3-simplex</i> , aplicando al conjunto de datos una transformación logarítmica.	V
A.15	Sumario estadístico (para diferentes valores del nivel de significación α) de la potencia estadística (\mathbb{P}) calculada, y el tiempo de computación empleado en las simulaciones <i>3-simplex</i> , aplicando al conjunto de datos una transformación <i>Centered Log Ratio</i> (<i>clr</i>).	VI
A.16	Sumario estadístico (para diferentes valores del nivel de significación α) de la potencia estadística (\mathbb{P}) calculada, y el tiempo de computación empleado en las simulaciones <i>3-simplex</i> , aplicando al conjunto de datos una transformación de <i>raíz cuadrada</i> (<i>sqr</i> t).	VI

Índice de Ecuaciones

PEC

Capítulo 0

PEC3 - Desarrollo del trabajo - Fase 2


0.1. Identificación del trabajo y fecha del informe

Ver la *Portada* y la *Ficha Del Trabajo Final*.

0.2. Avance del proyecto

La planificación ideada de las tareas que conforman la totalidad del proyecto puede encontrarse en 1.

Tras analizar los objetivos marcados al inicio del proyecto, y considerando el ritmo de avance en la obtención de los resultados previstos en este plan de trabajo, todo y que se puede dar por completado el estudio del primer y segundo objetivo, se ha decidido posponer la implementación de las simulaciones necesarias para el estudio del tercer objetivo, ya que, previendo el trabajo de redacción que queda por delante, no se dispondría del tiempo suficiente para cumplir las fechas de entrega establecidas.

Tanto para el primer como para el segundo objetivo, y después de implementar diversos escenarios, algunos fallidos y otros exitosos, se han introducido ciertos cambios de criterio para establecer las simulaciones que debían ser relevantes para obtener los resultados que se buscan en el proyecto actual. Este hecho ha aumentado el tiempo dedicado a las simulaciones en , en detrimento de la redacción de la memoria, conllevando la alteración en el orden de algunas tareas, e incluso la anulación de otras.

Recapitulando, se estima que la alteración y reestructuración de dicha planificación puede traducirse en que el proyecto final representará entorno al 80 % de los objetivos del original.



0.3. Relación de las actividades realizadas

Un detalle de las actividades del plan de trabajo actualizado y su compleción se muestran en el siguiente listado:



✓ **Primera entrega: PEC1 - Definición y plan de trabajo**

- ✓ Primera tutoría: propuesta oficial de TFM (título, objetivos principales, etc.).
- ✓ Redacción de la *PEC1 - Definición y plan de trabajo*.
- ✓ Redacción de la memoria: adaptación del contenido de la *PEC1* a los correspondientes apartados.
- ✓ Enmiendas basadas en las sugerencias realizadas por los tutores.
- ✓ Entrega de la *PEC1*.




✓ **Segunda entrega: PEC2 - Desarrollo del trabajo - Fase 1**

- ✎ Redacción de la memoria: inicio de la redacción del capítulo *Estado del arte*.
- ✓ Abordar las pruebas en , primer objetivo: estudiar las propiedades de *MANTA* en algunos escenarios comparando diferentes transformaciones de los datos.
- ✎ Redacción de la memoria: inclusión de los resultados obtenidos para el primer objetivo.
- ✓ Continuar con las pruebas en , segundo objetivo: estudiar la pérdida de potencia de la versión asintótica de *PERMANOVA* con respecto a *MANOVA* y otros métodos.
- ✎ Redacción de la memoria: inclusión de los resultados obtenidos para el segundo objetivo.
- ✗ Valorar, tras los resultados obtenidos para el segundo objetivo, la posibilidad de implementar mejoras en *MANTA*.
- ✓ Entrega de la *PEC2*.

□ **Tercera entrega: PEC3 - Desarrollo del trabajo - Fase 2**

- ✎ Redacción de la memoria: acabar, si es necesario, la escritura del capítulo *Estado del arte*.
- ✗ Seguir con las pruebas en , tercer objetivo: comparar los resultados obtenidos con respecto al cálculo de la distribución de las formas cuadráticas entre el método *Farebrother* (implementado para la versión asintótica de *PERMANOVA* con *MANTA*) y el de *Saddlepoint*.
- ✗ Redacción de la memoria: inclusión de los resultados obtenidos para el tercer objetivo.
- ✗ Pruebas secundarias en , otros objetivos: extender el tercer objetivo, ampliando la comparativa *Farebrother* vs. *Saddlepoint* a otros métodos (*Davies*, *Imhof*, *Liu*, etc.).
- ✗ Redacción de la memoria: inclusión, si cabe, de los resultados secundarios obtenidos.
- ✎ Redacción de la memoria: inicio de los capítulos *Discusión* y *Conclusiones y trabajos futuros*.
- ✓ Entrega de la *PEC3*.

☐ **Cuarta entrega:** *PEC₄ - Cierre de la memoria y de la presentación*

-  Finalizar la redacción de la memoria: adaptación del contenido generado en las diferentes *PEC* a los correspondientes apartados, finalizando las secciones anexas (*Bibliografía*, *Glosario*, etc.).
-  Creación, bajo los criterios establecidos, de la presentación basada en la memoria final.
-  Grabación en vídeo de la presentación.
- ☐ Entrega de la memoria, la presentación y el vídeo final obtenido (sendas copias en el *REC* y la aplicación *Present@*).

☐ **Defensa:** *Preparación para la defensa pública del TFM*

- ☐ Preparación de la defensa a la espera de la asignación definitiva de fecha.
- ☐ Defensa pública síncrona del TFM ante el tribunal asignado.


0.4. Desviaciones y acciones de mitigación

Como ya se ha avanzado en la sección 0.2, han habido ciertas desviaciones en la temporización, en particular, con el tiempo necesario para completar correctamente el primer y segundo objetivo, lo que ha hecho necesario implementar las acciones de mitigación pertinentes:

- Cambiar el orden de prioridad de las tareas relacionadas con el estudio del primer y segundo objetivo.
- Priorizar la obtención de resultados concretos en detrimento del estudio de escenarios, a priori, no tan relevantes.
- Transferir parte del tiempo previsto a la redacción en favor de la implementación correcta del código necesario, y de la posterior simulación de los escenarios considerados. Aplazando, así, la parte teórica a las dos últimas semanas programadas.
- Posponer *sine die* el estudio del tercer objetivo y sus subsiguientes derivadas por haber valorado que el tiempo de dedicación estimado pondría en peligro la compleción del proyecto en su globalidad dentro de las fechas establecidas.

0.5. Resultados parciales obtenidos

Resultados obtenidos hasta el momento:

- **Plan de trabajo:** documento donde se incluye una distribución de tareas según los objetivos determinados, puntos clave y tiempos necesarios (disponible en la sección *Planificación del trabajo*). Este ha ido sufriendo cambios a lo largo del proyecto (especificados en la sección *Desviaciones y acciones de mitigación*), debido principalmente a la alteración en el tiempo dedicado a las simulaciones necesarias, en detrimento de la realización de uno de los objetivos iniciales, y de la redacción de la memoria.
- **Memoria (provisional):** Capítulos *Introducción*, *Estado del arte* y *Resultados*.
- **Producto:** Todos los archivos, producto de la realización de este TFM, pueden encontrarse en el repositorio de GitHub [1]. Como se aprecia en el resumen de los lenguajes más usados, los diferentes *scripts* se han realizado principalmente en  y en \LaTeX . Mientras que los del primer tipo implementan tanto las funciones necesarias, como las diversas simulaciones de los escenarios considerados en cada caso a estudio, los segundos han sido necesarios para producir la memoria final del presente trabajo final de máster.

0.6. Comentarios de los directores del TFM

Durante todo el proceso se ha ido manteniendo un contacto periódico con los directores del proyecto con el fin de encauzar el trabajo iniciado, corregir algunos errores, y recibir recomendaciones tanto teóricas como prácticas. Esto ha permitido clarificar diversos aspectos del proyecto, determinar más claramente algunos enfoques del mismo, y redefinir las prioridades a la hora de obtener unos resultados en detrimento de otros debido a la reestructuración temporal del proyecto.

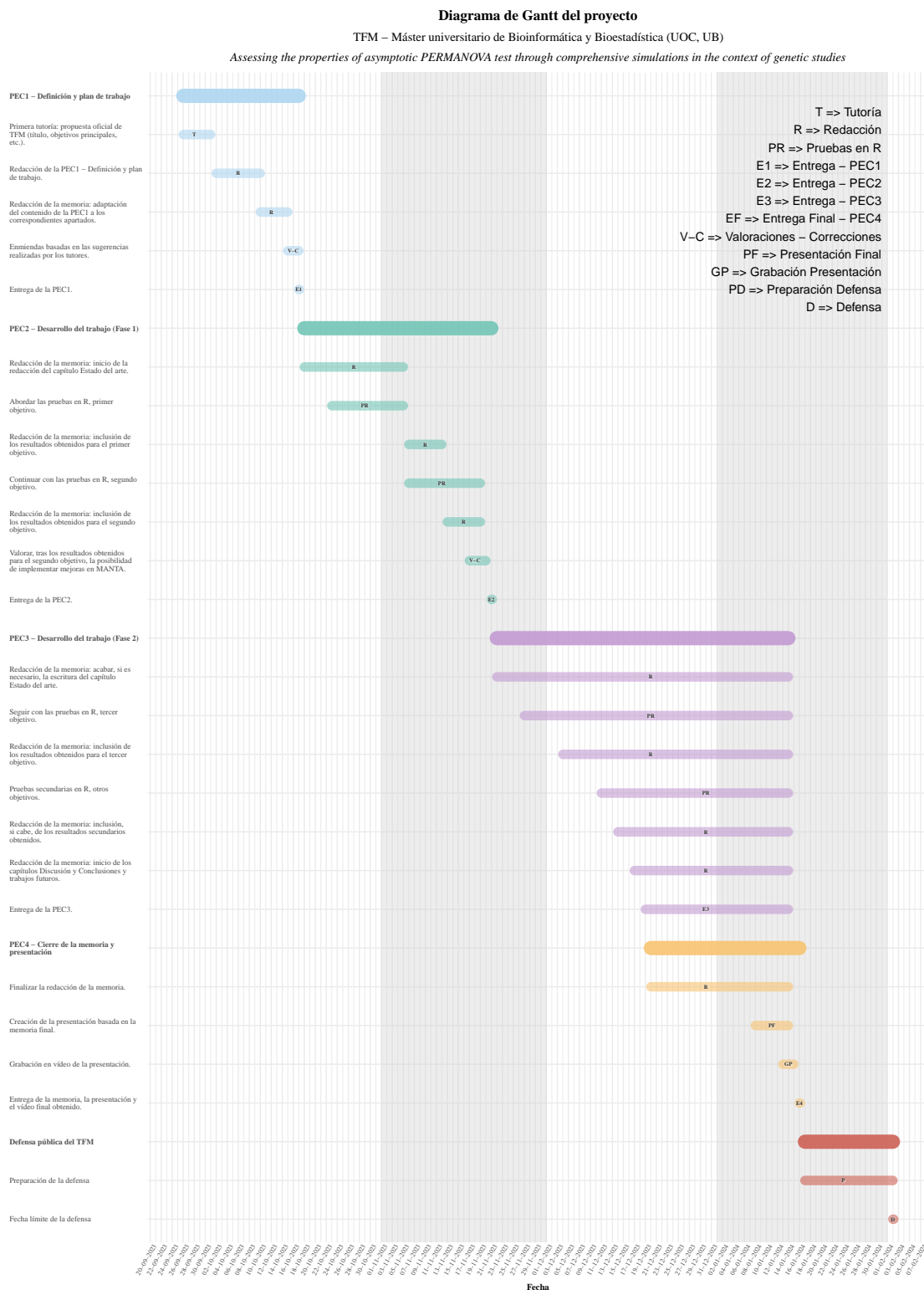


Figura 1: Planificación ideada de las tareas necesarias para la consecución de la memoria y presentación del presente TFM.

Memoria

Capítulo 1

Introducción

1.1. Contexto y justificación del trabajo

El tema escogido para la realización del TFM se enmarca en el análisis de datos ómicos mediante el uso de la *estadística multivariante*, principalmente la versión asintótica de *PERMANOVA*, aplicada al estudio de asociaciones entre los polimorfismos de un solo nucleótido (*SNPs*) del genoma completo (estudios tipo *GWAS*) y algunos rasgos característicos, como son las principales enfermedades humanas, así como en la detección de *sQTLs* utilizando datos *RNA-seq*.

Originalmente, las investigaciones basadas en *GWAS*, ya sea integrando *sQTLs* o no, se han realizado con la finalidad de comprobar la asociación entre los *SNPs* con diferentes variantes genéticas mediante el estudio de un único rasgo (única variable o *trait*), con lo que los análisis estadísticos correspondientes llevados a cabo suelen utilizar, lógicamente, los principales métodos univariantes disponibles (sumario estadístico basado en tablas de distribución de frecuencias, estadísticos de centralización o dispersión, etc.).

De este modo, este tipo de estudios, al centrarse en un solo *trait* de todos los disponibles en el gran volumen de datos sobre fenotipos utilizable, no permiten tratar la posible relación causa-efecto subyacente, obteniendo un análisis meramente descriptivo.

Alternativamente, gracias a la gran cantidad de datos disponibles últimamente con perfiles genómicos complejos (alta diversidad de rasgos moleculares), la necesidad de encontrar correlaciones entre las diferentes variables analizables y los rasgos de interés, ha resultado en un crecimiento en la utilización de métodos multivariantes para su análisis estadístico.

Las principales ventajas con respecto al enfoque univariante clásico, para poder determinar la posible estructura de correlaciones subyacente en los datos, pueden enumerarse como sigue:

- Mayor *potencia estadística* (\mathbb{P}) para detectar asociaciones genéticas.
- Ofrece ventajas en el estudio de la *pleiotropía* (cuando el gen o alelo considerado es responsable de efectos fenotípicos o caracteres distintos y, a priori, no relacionados).
- Resulta de utilidad incluso cuando solo un pequeño grupo de los rasgos se ve afectado por el genotipo de interés.


- Permite el análisis a través de múltiples capas fenotípicas en bloque, dando luz sobre los mecanismos moleculares activados por las variantes genéticas consideradas.
- Posibilita la caracterización de los efectos genéticos sobre un mismo rasgo cuando este es medido en diferentes condiciones ambientales o entornos.
- Requiere de menos pruebas individuales, lo que disminuye las de carácter múltiple.

Contrariamente, del uso de los métodos más habitualmente utilizados para estudiar estas asociaciones genéticas multirasgo emergen diversos inconvenientes, entre los cuales destacan:

- Los métodos que modelan el genotipo como variable dependiente comprobando a su vez la asociación con una suma ponderada de fenotipos (*MV-PLINK* ([2]) o análisis de correlación canónica, y *MultiPhen* [3] que utiliza la regresión ordinal) adolecen de la posibilidad de evaluar diseños complejos que presentan múltiples interacciones entre el genotipo y otras covariables.
- Tanto el análisis multivariante de la varianza (*MANOVA*), como el de los modelos multivariantes lineales mixtos (*mvLMMs*) [4], resultan ser más tolerantes a estos diseños complejos al tratar los fenotipos como variables dependientes, introduciendo de forma natural el posible parentesco genético entre los individuos analizados. Esta ventaja se torna inconveniente para grandes conjuntos de datos, sobre todo para el método *mvLMMs*, cuya continua mejora en su implementación computacional sigue requiriendo de tiempos excesivamente altos.
- La pluralidad de los métodos de regresión multivariante presuponen una normalidad en la distribución de los errores del modelo que puede no llegar a cumplirse. Todo y que pueden aplicarse transformaciones individuales a cada rasgo estudiado, no puede garantizarse la normalidad multivariante, lo que resulta en una reducción de la potencia estadística en comparación con el modelo aplicado a los rasgos no transformados.
- Hasta el momento, las diversas implementaciones de *métodos bayesianos* para el estudio de asociaciones multirasgo no han sido satisfactorias, requiriendo siempre un tiempo elevado de cálculo debido al coste computacional que implican.
- Para los métodos *MTAR* [5] o *MOSTest* [6] [7] existe la necesidad de garantizar la normalidad multivariante asintótica cuando se utilizan los sumarios estadísticos univariantes, lo que no es trivial, sumado a que evitar la aparición de sesgos en la estimación de correlaciones de rasgos a partir de esta clase de estadísticos no es sencillo (afectaciones de heredabilidad de los rasgos, patrones de desequilibrio de ligamiento, etc.).

Con todo lo anterior, resulta evidente la necesidad de disponer de un método no paramétrico adecuado tanto para los estudios basados en (*GWAS*) como en *sQTLs*. El modelo de *PERMANOVA* ([8]) amplía el modelo lineal factorial univariante a múltiples dimensiones sin requerir una distribución de probabilidad conocida de las variables dependientes, introduciendo un enfoque basado en la distancia, poniendo a prueba la hipótesis de ausencia de efectos mediante un procedimiento de permutación basado en un estadístico *pseudo-F*, en el que las sumas de cuadrados del *ANOVA* se sustituyen por sumas de interdistancias entre observaciones.

Pese a ser exitoso en muchos estudios, dando buenos resultados en un tiempo de cálculo reducido para diseños fijos unidireccionales, resulta inviable en los estudios actuales, donde el mayor tamaño y complejidad de los conjuntos de datos requiere una precisión para el cálculo del valor *p* que este procedimiento permutacional no puede alcanzar en las condiciones requeridas.

El punto de partida del presente trabajo radica en los diversos estudios realizados con el fin de superar esta limitación. En concreto: sendos artículos de Garrido-Martín, D. *et al.* ([9] y [10]), y el trabajo de Monlong, J. *et al.* [11]. Donde, gracias al programa *MANTA* ([12], desarrollado principalmente en ) , se estudia mediante diversas simulaciones ([13]) de diseños complejos la distribución asintótica de la estadística de pruebas *PERMANOVA* en el caso de la distancia euclídea (*valores p* de carácter no paramétrico y asintótico para modelos lineales multivariados), obteniendo resultados igualmente válidos tras cualquier transformación de los datos que preserve la independencia de las observaciones.

La finalidad principal será ahondar en estos estudios, yendo más allá en al menos los siguientes aspectos:

- Estudiar las propiedades de *MANTA* en algunos escenarios, determinando cómo los diferentes tipos de transformaciones de datos afectan a los resultados obtenidos, y dilucidar si existe algún protocolo privilegiado en las simulaciones implementadas.
- Estudiar la pérdida de potencia de la versión asintótica de *PERMANOVA* con respecto a *MANOVA* y otros métodos, profundizando en la afectación de la variación del nivel de significación considerado sobre la potencia de cada uno.
- Comparar los resultados obtenidos con respecto al cálculo de la distribución de las formas cuadráticas entre el método Farebrother (implementado para la versión asintótica de *PERMANOVA* con *MANTA*) y el de Saddlepoint.
- Extender el punto anterior, ampliando la comparativa Farebrother vs. Saddlepoint a otros métodos: métodos exactos como el de Davies, R. B. ([14], [15]), o aproximaciones numéricas como la de Liu–Tang–Zhang ([16]), el método de Imhof, etc.
- Partiendo del caso de estudio anterior, y secundariamente, se llevaría a cabo la implementación del método más óptimo en el paquete *MANTA* ya existente, en caso de que este exista.

1.2. Objetivos del trabajo

De los diferentes puntos detallados en el apartado anterior, se extrae que el presente trabajo deberá permitirnos profundizar en aspectos concretos de los estudios ya referenciados ([9], [10]), [11]), con el objetivo último de determinar la validez de la versión asintótica del método *PERMANOVA* (implementado en el paquete *MANTA*) con respecto a otros métodos similares bajo un mismo conjunto de simulaciones computacionales complejas basadas en datos de escenarios reales.

Según las bases generales establecidas, y para una consecución satisfactoria del estudio propuesto, se han considerado los siguientes objetivos principales:

- Estudiar las propiedades de *MANTA* en algunos escenarios, determinando cómo los diferentes tipos de transformaciones de datos afectan a los resultados obtenidos, y dilucidar si existe algún protocolo privilegiado en las simulaciones implementadas.
- Estudiar la pérdida de potencia de la versión asintótica de *PERMANOVA* con respecto a *MANOVA* y otros métodos, profundizando en la afectación de la variación del nivel de significación considerado sobre la potencia de cada uno.
- Comparar los resultados obtenidos con respecto al cálculo de la distribución de las formas cuadráticas entre el método Farebrother (implementado para la versión asintótica de *PERMANOVA* con *MANTA*) y el de Saddlepoint.

Como extensión de los mismos, resulta también conveniente establecer otros objetivos secundarios:

- Extender el tercer objetivo principal, ampliando la comparativa Farebrother vs. Saddlepoint a otros métodos: métodos exactos como el de Davies, R. B. ([14], [15]), o aproximaciones numéricas como la de Liu–Tang–Zhang ([16]), el método de Imhof, etc.
- Partiendo del caso de estudio anterior, se llevaría a cabo la implementación del método más óptimo en el paquete *MANTA* ya existente, en caso de que los resultados obtenidos indiquen que alguno de ellos resulta ser más eficiente tanto computacional como estadísticamente hablando.


1.3. Enfoque y método seguido

En cuanto al tiempo de dedicación, se ha enfocado el trabajo adaptando las tareas ideadas para la consecución de los diferentes objetivos marcados a las pautas establecidas por las diferentes entregas programadas por la UOC (*Pruebas de Evaluación Continua* o *PEC*), pudiendo destacar los siguientes grandes bloques de trabajo y la estimación del peso que estos han tenido en el tiempo dedicado al proyecto:

- **Primera fase o entrega (PEC1):** definición y plan de trabajo.
⇒ *Dedicación estimada:* 3 % del tiempo total final destinado al trabajo.
- **Segunda fase o entrega (PEC2):** primera fase del desarrollo del trabajo.
⇒ *Dedicación estimada:* 35 % del tiempo total final destinado al trabajo.
- **Tercera fase o entrega (PEC3):** segunda fase del desarrollo del trabajo.
⇒ *Dedicación estimada:* 45 % del tiempo total final destinado al trabajo.
- **Cuarta fase o entrega (PEC4):** cierre de la memoria y preparación de los documentos pertinentes.
⇒ *Dedicación estimada:* 15 % del tiempo total final destinado al trabajo.
- **Defensa pública:** Preparación de la defensa pública en base a la memoria final del proyecto.
⇒ *Dedicación estimada:* 2 % del tiempo total final destinado al trabajo.


Se puede encontrar una planificación más detallada en las secciones *Planificación del trabajo* e *Hitos*, y una recopilación de los contratiempos encontrados y sus correspondientes enmiendas en la sección *Desviaciones en la planificación y acciones de mitigación*.

Con respecto al enfoque práctico, y a la metodología específica de trabajo seguida, caben destacar los siguientes aspectos:

- Estudio inicial detallado de los principales artículos de referencia: [9], [10], [11].
- Estudio en profundidad del método *PERMANOVA* implementado en el paquete *MANTA* ([12]), haciendo hincapié en la comprensión de la estructura del algoritmo implementado, y en la labor de cada una de las funciones implicadas.
- Estudio en profundidad de las diferentes simulaciones utilizadas en [9] ([13]), con la finalidad de comprender el propósito de cada una de ellas en dicho trabajo e identificar el escenario en el cual el presente trabajo puede aportar más información.
- Implementación en  del algoritmo necesario para realizar las simulaciones requeridas en el estudio de cada uno de los objetivos especificados, partiendo de la estructura original presente en [13], pero adaptando pertinentemente los escenarios de simulación adecuados.
- Pruebas del código implementado con diferentes finalidades: estudio de los resultados obtenidos bajo diversas combinaciones de las variables implicadas, salvaguardado oportuno de los datos producidos para su posterior análisis y utilización en la memoria, determinación de la visualización gráfica más apropiada en cada caso
- Subsanado continuo de los errores obtenidos en la ejecución del código implementado.
- Evaluación y análisis con respecto al propósito original del proyecto de los resultados finales logrados.

1.4. Planificación del trabajo

Tras analizar los objetivos marcados al inicio del proyecto, y considerar el ritmo de avance en la obtención de los resultados previstos en el plan de trabajo (1.1), cabe destacar que se puede dar por completado el estudio del primer y segundo objetivo. Por otro lado, y pese a que se hicieron algunas pruebas preliminares, la implementación definitiva de las simulaciones necesarias para el estudio del tercer objetivo y su posterior evaluación ha tenido que ser pospuesta a futuros trabajos, ya que la fase de redacción final del trabajo no deja el tiempo suficiente para su realización sin poner en peligro la finalización del proyecto dentro de las fechas establecidas.

En particular, tanto para el primer como para el segundo objetivo, y después de implementar diversos escenarios, algunos fallidos y otros exitosos, se han introducido ciertos cambios de criterio para establecer las simulaciones que debían ser relevantes para obtener los resultados que se buscaban en el proyecto actual. Este hecho ha aumentado el tiempo dedicado a las simulaciones en , en detrimento de la redacción de la memoria, conllevando la alteración final del orden de algunas tareas, e incluso la anulación de otras inicialmente previstas.

Recapitulando, se estima que la alteración y reestructuración de dicha planificación puede traducirse en que el proyecto final representará entorno al 80 % de los objetivos del original. La planificación final de las tareas que han conformado cada bloque de trabajo específico puede encontrarse en 1.1.

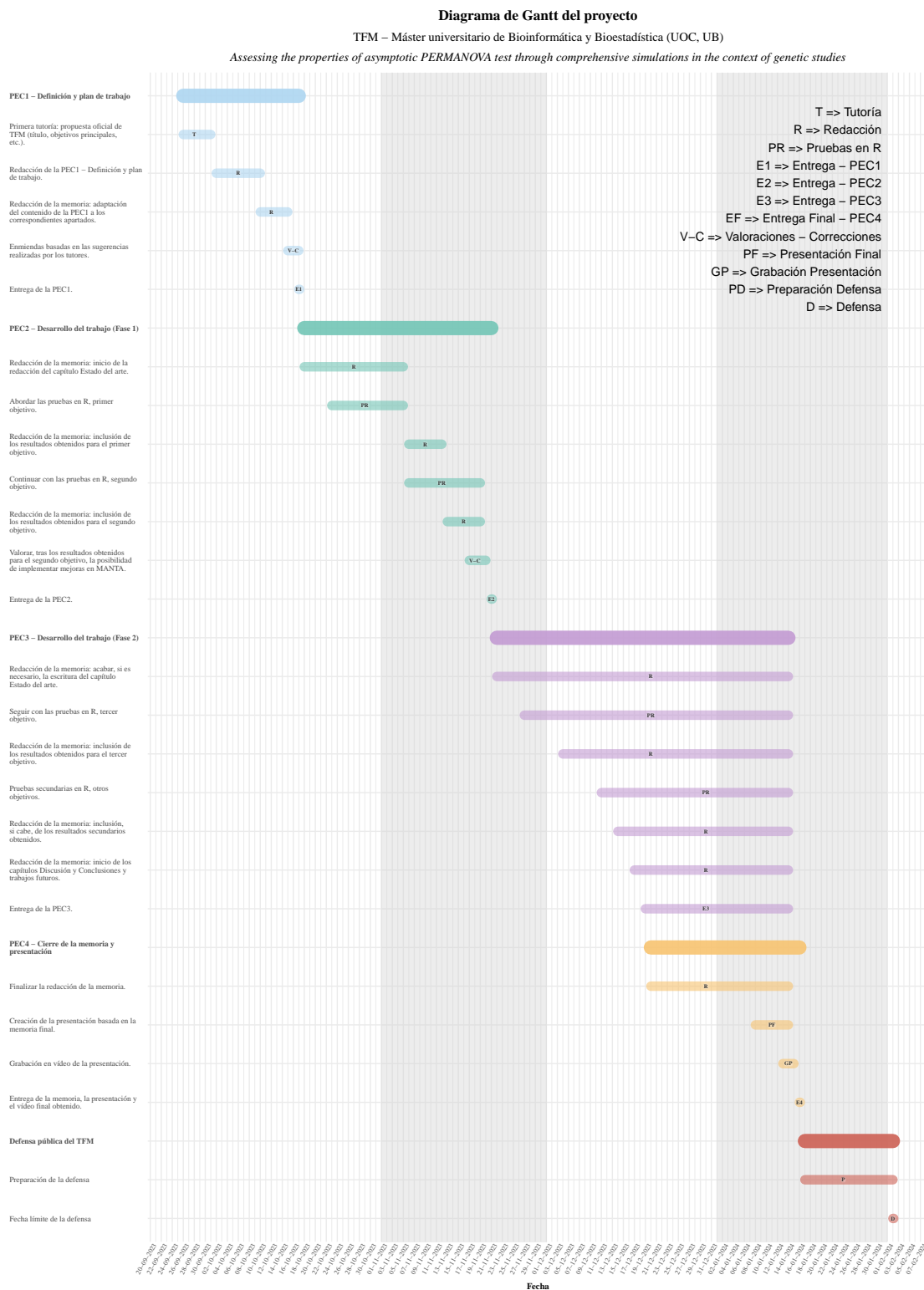


Figura 1.1: Planificación ideada de las tareas necesarias para la consecución de la memoria y presentación del presente TFM.



1.5. Hitos

A continuación se muestran las distintas fases del proyecto según su compleción, indicando los posibles retrasos o problemas inesperados o no que, al surgir, pueden haber puesto en riesgo la consecución de las tareas previstas, y los objetivos establecidos:



✓ **Primera entrega:** *PEC1 - Definición y plan de trabajo*

- ✓ Primera tutoría: propuesta oficial de TFM (título, objetivos principales, etc.).
- ✓ Redacción de la *PEC1 - Definición y plan de trabajo*.
- ✓ Redacción de la memoria: adaptación del contenido de la *PEC1* a los correspondientes apartados.
- ✓ Enmiendas basadas en las sugerencias realizadas por los tutores.
- ✓ Entrega de la *PEC1*.

✓ **Segunda entrega:** *PEC2 - Desarrollo del trabajo - Fase 1*

- ✎ Redacción de la memoria: inicio de la redacción del capítulo *Estado del arte*.
- ✓ Abordar las pruebas en , primer objetivo: estudiar las propiedades de *MANTA* en algunos escenarios comparando diferentes transformaciones de los datos. EL tiempo de dedicación fue ligeramente más elevado del estimado originalmente.
- ✎ Redacción de la memoria: inclusión de los resultados obtenidos para el primer objetivo. Pospuesta a la parte final del proyecto, una vez acabadas las simulaciones y obtenidos los resultados.
- ✓ Continuar con las pruebas en , segundo objetivo: estudiar la pérdida de potencia de la versión asintótica de *PERMANOVA* con respecto a *MANOVA* y otros métodos. En este caso, el tiempo de dedicación creció más de lo esperado con respecto al estimado en un principio, ya que surgieron diversos errores en el desarrollo de las simulaciones.
- ✎ Redacción de la memoria: inclusión de los resultados obtenidos para el segundo objetivo. Pospuesta a la parte final del proyecto, una vez acabadas las simulaciones y obtenidos los resultados.
- ✗ Valorar, tras los resultados obtenidos para el segundo objetivo, la posibilidad de implementar mejoras en *MANTA*. Para no poner en peligro la consecución de los plazos establecidos, se ha decidido posponer a futuros trabajos este objetivo secundario, derivado directamente del anteriormente mencionado.
- ✓ Entrega de la *PEC2*.

✓ **Tercera entrega: PEC3 - Desarrollo del trabajo - Fase 2**

- ✎ Redacción de la memoria: acabar, si es necesario, la escritura del capítulo *Estado del arte*. Pospuesta a la parte final del proyecto, una vez acabadas las simulaciones y obtenidos los resultados.
- ✗ Seguir con las pruebas en , tercer objetivo: comparar los resultados obtenidos con respecto al cálculo de la distribución de las formas cuadráticas entre el método *Farebrother* (implementado para la versión asintótica de *PERMANOVA* con *MANTA*) y el de *Saddlepoint*. Pese a haberse realizado algunas simulaciones preliminares, se ha aplazado a posteriores trabajos por falta de tiempo.
- ✗ Redacción de la memoria: inclusión de los resultados obtenidos para el tercer objetivo. Este punto no se ha llevado a cabo por haber sido aplazado a posteriores trabajos el objetivo correspondiente.
- ✗ Pruebas secundarias en , otros objetivos: extender el tercer objetivo, ampliando la comparativa *Farebrother* vs. *Saddlepoint* a otros métodos (*Davies*, *Imhof*, *Liu*, etc.). Este punto no se ha llevado a cabo por haber sido aplazado a posteriores trabajos el objetivo correspondiente.
- ✗ Redacción de la memoria: inclusión, si cabe, de los resultados secundarios obtenidos. Pospuesta a la parte final del proyecto, una vez acabadas las simulaciones y obtenidos los resultados.
- ✎ Redacción de la memoria: inicio de los capítulos *Discusión y Conclusiones y trabajos futuros*. Pospuesta a la parte final del proyecto, una vez acabadas las simulaciones y obtenidos los resultados.
- ✓ Entrega de la *PEC3*.

✎ **Cuarta entrega: PEC4 - Cierre de la memoria y de la presentación**

- ✎ Finalizar la redacción de la memoria: adaptación del contenido generado en las diferentes *PEC* a los correspondientes apartados, finalizando las secciones anexas (*Bibliografía*, *Glosario*, etc.).
- ✎ Creación, bajo los criterios establecidos, de la presentación basada en la memoria final.
- ✎ Grabación en vídeo de la presentación.
- ☐ Entrega de la memoria, la presentación y el vídeo final obtenido (sendas copias en el *REC* y la aplicación *Present@*).
- ☐ **Defensa: Preparación para la defensa pública del TFM**
 - ☐ Preparación de la defensa a la espera de la asignación definitiva de fecha.
 - ☐ Defensa pública síncrona del TFM ante el tribunal asignado.

1.6. Desviaciones en la planificación y acciones de mitigación

Como ya se ha avanzado en los pormenores de la sección *Hitos*, han habido ciertas desviaciones en la temporización, en particular, con el tiempo necesario para completar correctamente el primer y segundo objetivo, lo que ha hecho necesario implementar las acciones de mitigación pertinentes:


- Cambiar el orden de prioridad, y el tiempo asignado, a las tareas relacionadas con el estudio del primer y segundo objetivo.
- Priorizar la obtención de resultados concretos en detrimento del estudio de escenarios, a priori, no tan relevantes.
- Transferir parte del tiempo previsto a la redacción en favor de la implementación correcta del código necesario, y de la posterior simulación de los escenarios considerados. Aplazando, así, la parte teórica a las dos últimas semanas programadas.
- Tras consensuarlo con los tutores, se creyó oportuno centrarse en el estudio de las dos metas principales, y posponer *sine die* el estudio de la tercera y sus subsiguientes derivadas. Se valoró que el tiempo de dedicación estimado pondría en peligro la compleción del proyecto en su globalidad dentro de las fechas establecidas.

1.7. Análisis de riesgos

En esta sección se añaden algunas de las contingencias que han ido surgiendo durante la realización del proyecto, indicando, a su vez, si alguna de ellas ha impedido su apropiado avance o, incluso, la no consecución de alguno de los objetivos planteados:

- **Tiempo limitado:** debido a que el TFM debe realizarse en un solo cuatrimestre, el tiempo disponible para desarrollar el proyecto suele ser muy ajustado. Cualquier contratiempo o retraso en la planificación, ya sea predecible o no, puede afectar gravemente a la consecución de los plazos y, eventualmente, impedir alcanzar alguno de los objetivos que se hayan planteado. Aunque se ha maximizado la dedicación al mismo con el tiempo disponible, se han identificado algunos escollos que podrían haber acabado atascando el proceso, lo cual ha sido clave para mitigar sus efectos.
- **Planificación incorrecta:** aunque en principio no parece que se haya realizado inicialmente una mala priorización o asignación de tiempo a las tareas pertinentes, si que nos hemos encontrado con más dificultades de las esperadas a la hora de la implementación de las simulaciones necesarias para uno de los objetivos, lo cual ha influido de manera negativa en la planificación. Como resultado, estos problemas han impedido disponer del tiempo necesario para obtener los resultados buscados en uno de los objetivos marcados.
- **Etapas de análisis y pruebas:** como ya se ha indicado, durante esta fase han aparecido algunos contratiempos, relacionados principalmente con la dificultad en la consecución del funcionamiento deseado de algunos *scripts* de código que implementan los escenarios de simulación buscados. Aunque no se produjeron problemas inesperados con el computador utilizado, la lentitud de algunos procesos de simulación, debido a la multitud de combinaciones de variables a simular, también ha hecho mella en los tiempos destinados a cada fase del trabajo.

1.8. Breve resumen de productos obtenidos

- **Plan de trabajo:** documento donde se incluye una distribución de tareas según los objetivos determinados, puntos clave y tiempos necesarios (disponible en la sección *Planificación del trabajo*). Este ha ido sufriendo cambios a lo largo del proyecto (especificados en la sección *Desviaciones y acciones de mitigación*), debido principalmente a la alteración en el tiempo dedicado a las simulaciones necesarias, en detrimento de la realización de uno de los objetivos iniciales, y de la redacción de la memoria.
- **Memoria:** producto derivado de todas las entregas parciales o *PEC* (basado en la estructura recomendada por la UOC), donde se detallará el contexto científico, los resultados obtenidos según el procedimiento seguido y, finalmente, las conclusiones extraídas tras su interpretación. Consta de los capítulos principales: *Introducción*, *Estado del arte*, *Resultados*, *Discusión* y *Conclusiones y trabajos futuros*.
- **Producto:** Todos los archivos, producto de la realización de este TFM, pueden encontrarse en el repositorio de GitHub [1]. Como se aprecia en el resumen de los lenguajes más usados, los diferentes *scripts* se han realizado principalmente en  y en L^AT_EX. Mientras que los del primer tipo implementan tanto las funciones necesarias, como las diversas simulaciones de los escenarios considerados en cada caso a estudio, los segundos han sido necesarios para producir la memoria final del presente trabajo final de máster.
- **Presentación virtual del TFM:** exposición oral y visual basada en la memoria producida. En ella se resaltan los aspectos más importantes del trabajo realizado, presentando las distintas fases del proyecto de forma resumida.
- **Autoevaluación del proyecto:** documento que, una vez finalizado el proyecto, debe redactarse para plasmar una evaluación crítica del trabajo realizado, determinando el grado de alcance de los objetivos, y valorando los aspectos potencialmente mejorables.

1.9. Comentarios de los directores del TFM

Durante todo el proceso se ha ido manteniendo un contacto periódico con los directores del proyecto con el fin de encauzar el trabajo iniciado, corregir algunos errores, y recibir recomendaciones tanto teóricas como prácticas. Esto ha permitido: clarificar diversos aspectos del proyecto, determinar más claramente algunos enfoques del mismo, redefinir las prioridades a la hora de obtener unos resultados en detrimento de otros, y a reestructurar debidamente el esquema temporal del proyecto.

1.10. Descripción de otros capítulos

En esta sección se realizará, en caso de ser necesario, una escueta descripción de los diversos capítulos de la memoria.

Capítulo 2

Estado del arte

2.1. Contexto biotecnológico

Aunque el presente trabajo se basa esencialmente en un estudio computacional comparativo entre dos métodos estadísticos multivariantes específicos, bajo la simulación de ciertos escenarios totalmente controlados, estos pueden desempeñar un papel importante en diversos campos de investigación. En este aspecto, cabe destacar su aplicabilidad en estudios biológicos, en concreto, en el campo de la genética avanzada.

Esta disciplina ha evolucionado enormemente en las últimas décadas gracias al desarrollo de nuevas tecnologías, permitiendo identificar variantes genéticas asociadas a una amplia variedad de fenotipos y rasgos. La secuenciación de nucleótidos que conforman la molécula de *ADN* permite el análisis detallado de su estructura, convirtiéndose en la herramienta idónea para identificar variantes en el material genético.

En los años 70, la *secuenciación Sanger* revolucionó la investigación originando la era genómica, la cual se ha asentado gracias a la evolución que ha sufrido este campo en los últimos años debido al desarrollo de las nuevas plataformas de secuenciación de alto rendimiento (*Next-Generation Sequencing* o *NGS*). Estas nuevas técnicas son capaces de generar paralelamente, y de forma masiva, millones de fragmentos de *ADN* en un único proceso de secuenciación, consiguiendo así el análisis de grandes cantidades de información genética en una escala inimaginable en los orígenes de esta disciplina [17, 18].

Recientemente, cada vez más investigaciones han empezado a decantarse por un tipo específico de secuenciación masiva: la *secuenciación de ARN* o *RNA-seq*. Dicha estrategia de análisis utiliza técnicas *NGS* para revelar la presencia y cantidad de *ARN* existente en una muestra biológica en un momento dado, lo cual le permite ser aplicada para analizar cambios en el transcriptoma. Algunos usos potenciales de esta técnica que cabe destacar son [19] [20]:

- Observación de transcritos resultantes del *empalme alternativo*, *modificación postranscripcional*, *fusiones génicas*, *mutaciones/polimorfismos de nucleótidos únicos* o *SNP* y *cambios de expresión de genes*.
- Caracterización de diferentes poblaciones de *ARN*: *miARN*, *tARN*, y *rARN*.
- Determinación de las *fronteras exón/intrón*
- Verificación y corrección de *regiones 5' y 3'*.

Gracias a estas técnicas de secuenciación masiva, los estudios de asociación del genoma completo (*GWAS*) han sido hasta ahora el enfoque prioritario en cuanto a los esfuerzos por identificar variantes genéticas asociadas con un fenotipo o rasgo particular. Esta estrategia analiza la variación genética en todo el genoma de un gran número de individuos para identificar regiones asociadas con el rasgo de interés, resultando especialmente útil en la identificación de variantes genéticas asociadas con enfermedades comunes y complejas, así como con rasgos específicos. En concreto, este tipo de estudios destacan por su eficiencia en los siguientes campos [21, 22, 16, 23, 24]:

- En la identificación de nuevas *asociaciones variante-rasgo*, estableciendo con éxito *loci* de riesgo para un gran número de enfermedades y rasgos. Aunque no pueden explicar toda la heredabilidad de los rasgos complejos, representan un medio práctico por el cual se pueden descubrir asociaciones genuinas. De forma que con el aumento del número de muestras *GWAS* se deberían seguir determinando nuevos *loci*.
- Los *GWAS* pueden conducir al descubrimiento de nuevos mecanismos biológicos. Los *loci GWAS* suelen implicar genes de función desconocida o que no se consideraban relevantes, cuyo seguimiento experimental puede conducir al descubrimiento de nuevos mecanismos biológicos subyacentes a las enfermedades.
- Tiene diversas aplicaciones clínicas, ya que las variantes genéticas descubiertas mediante *GWAS* pueden utilizarse para identificar a individuos con alto riesgo de padecer determinadas enfermedades, mejorando así la evolución de los pacientes mediante la detección precoz, la prevención o el tratamiento. Sus hallazgos pueden aplicarse a la clasificación y subtipificación de enfermedades.
- Pueden aportar información sobre la etnicidad de ciertos rasgos complejos, ya que se conoce que algunos *loci* de riesgo muestran considerables diferencias étnicas en frecuencia y/o tamaño del efecto.
- También son relevantes para el estudio de variantes raras y de baja frecuencia. En la actualidad, la mayor parte de este tipo de estudios se realizan utilizando datos obtenidos mediante *arrays de SNP* que, al incluir actualmente una mayor densidad de variantes y una gama más amplia de frecuencias alélicas, permiten genotipar directamente muchas variantes raras y de baja frecuencia. Las variantes raras y de baja frecuencia también pueden genotiparse utilizando matrices personalizadas centradas en exomas.
- Otras aplicaciones son: puede utilizarse para *identificar nuevos genes de enfermedades monogénicas y oligogénicas*, puede *estudiar variantes genéticas distintas de los SNV*, sus datos se utilizan para *múltiples aplicaciones más allá de la identificación de genes*, además, la *generación, gestión y análisis de datos GWAS son sencillos*, además, al ser fácilmente compartibles y de acceso público, *facilitan nuevos descubrimientos*.

Cabe destacar uno de estos puntos, ya que representa a la mayoría de los estudios tipo *GWAS* que se realizan hoy en día. Es el caso de las investigaciones que utilizan datos *GWAS basados en arrays de SNP*, que ofrecen unas ventajas particulares:

- **Utilizan una tecnología de genotipado muy precisa:** lo que es crucial para el éxito de cualquier estudio de asociación genética a gran escala, donde los sesgos sistemáticos inducidos por las fuentes de error (aunque sean pequeñas) pueden hacer crecer tanto los falsos positivos como negativos a la hora de determinar las *asociaciones variante-rasgo*. Concretamente, en la actualidad las *matrices SNP* de genoma completo contemporáneas alcanzan precisiones por encima del 99,7% y, además, se han desarrollado protocolos que determinan la utilización de solo aquellos datos que superen los umbrales de calidad para cada uno de los indicadores (*call rate*, concordancia de duplicados, consistencia mendeliana y prueba de equilibrio de Hardy-Weinberg), establecidos de forma independiente para cada estudio, garantizando así la confianza en los resultados.
- **Rentabilidad (coste y tiempo) a la hora de identificar loci de riesgo:** es un método rentable ya que el coste del análisis de *matrices SNP* seguido de la *imputación de variantes hasta un MAF del 0,1 %* se ha ido reduciendo durante los últimos años de forma constante. Este hecho permite, hoy en día, explorar gran parte de la variación genética del genoma a un coste razonable incluso en muestras de gran tamaño.

Por otro lado, también presentan algunas limitaciones, entre las cuales se destacan las siguientes [21, 23, 24]:

- Estos estudios se ven altamente afectados por la necesidad de adoptar un alto nivel de significación para tener en cuenta las pruebas múltiples realizadas, lo que comúnmente se realiza usando una *corrección de Bonferroni* para mantener la tasa de falsos positivos en todo el genoma en un 5 % (basado en la suposición de 1×10^6 pruebas independientes para la variación genética común). Todo esto afecta a la *potencia* de esta técnica para detectar toda la hereditariadad explicada por los *SNPs*, ya que las señales de asociación deberán alcanzar un umbral de $\mathbb{P} < 5 \times 10^{-8}$ para ser consideradas significativas. Una estrategia útil en algunas ocasiones para superar esta limitación es aumentar el tamaño de la muestra; en otras ocasiones, reducir el número de pruebas realizadas (lo cual se logra utilizando pruebas de asociación basadas en genes o limitando los análisis a regiones genómicas candidatas) es lo más adecuado.
- Pese a haber identificado un número sin precedentes de variantes genéticas asociadas con enfermedades y rasgos comunes, sólo es capaz de explicar una pequeña parte de la heredabilidad estimada de los rasgos más complejos. Una probable explicación es que los *SNPs* que afectan moderadamente se pierden porque no alcanzan el estricto umbral de significación establecido. El aumento constante en los tamaños de muestras, así como la adopción de nuevos métodos y diseños de estudio, pueden ayudar a solventar dicho escollo.
- La correlación local de múltiples variantes genéticas debido al desequilibrio de enlace facilita la identificación inicial de un *locus* pero dificulta el discernimiento de la variante o variantes causales. Una vez que se ha realizado un *GWAS*, a menudo se requieren pasos adicionales para identificar las variantes causales y sus genes de destino. Los avances en los métodos estadísticos como los enfoques bayesianos, han permitido avanzar en la restricción de las posibles variantes causales. Además, el aumento de bases de datos de elementos reguladores en una variedad de tipos de tejidos y células, disponibles públicamente (ENCODE, Epigenome RoadMap, FANTOM5 y GTEx), así como herramientas para la consulta de dichos bancos de datos, ha permitido integrar los hallazgos de *GWAS* con datos de genómica funcional en múltiples niveles, priorizando las variantes candidatas para el seguimiento funcional.
- Otras limitaciones también son: que no se pueden identificar todos los determinantes genéticos de rasgos complejos; su poca fiabilidad en la detección de epistasia en humanos; que las señales analizadas pueden deberse a la estratificación criptica de la población; usualmente, tienen una capacidad predictiva clínica limitada.
- En cuanto a los *GWAS basados en arrays de SNP*, existen algunas limitaciones particulares:
 - Dependen de la integridad de los estudios de secuenciación y los paneles de referencia resultantes que se utilizan para informar al diseño de matriz de genotipado e imputar variantes no tipadas en *GWAS*. Las primeras configuraciones de SNP para todo el genoma fueron diseñados seleccionando *SNPs* de los paneles de referencia de poblaciones predominantemente europeas, generando así un sesgo, y eludiendo la influencia que la variación de los patrones de desequilibrio de enlace entre grupos étnicos pueda tener. De un tiempo a esta parte, se ha intentado solventar desarrollando una nueva generación de matrices de alta densidad cuyos contenidos se basan en datos de secuenciación de poblaciones más diversas. Sin embargo, muchos grupos étnicos todavía no han sido secuenciados.
 - Aunque la evidencia empírica sugiere que gran parte de la heredabilidad de los rasgos complejos puede explicarse por variantes comunes, también se considera que las variantes raras y ultra raras han de contribuir de alguna manera. En este contexto, es destacable el hecho que hoy en día los *GWAS basados en arrays de SNP* son incapaces de detectar variantes ultrararas asociadas con la enfermedad.

Además, cabe tener en cuenta también que en este tipo de estudios una proporción de las variantes descritas son *QTLs* (*eQTLs*, *trQTLs*, etc.), siendo de particular interés los *locus de rasgo cuantitativo de empalme* (*sQTLs*), los cuales regulan el empalme alternativo del *pre-ARNm*, y pueden ser detectados usando datos *RNA-seq* [25, 11]. De esta manera, la correcta

integración del genoma secuenciado, los *QTLs* y el fenotipo celular, puede ayuda a comprender los genes causantes de ciertas enfermedades, las variantes genéticas causales que subyacen a los *GWAS* y los procesos biológicos que intervienen.

2.2. Estadística multivariante aplicada a estudios *GWAS* basados en datos *RNA-seq*

Precisamente, el presente trabajo trata de caracterizar dos de los métodos de estadística multivariante más novedosos, contextualizándose dentro de los estudios *GWAS* de ciertas asociaciones *SNPs-Rasgos característicos* que integran la influencia de los *sQTLs* mediante el análisis de datos *RNA-seq*. En particular, mediante un estudio cuantitativamente comparativo de la *potencia estadística* \mathbb{P} de sendos métodos: *MANOVA*, y la versión asintótica de *PERMANOVA*.

Para poder dar una visión suficientemente amplia de todos los métodos estadísticos aplicados a este tipo de estudios, a continuación se pretende elaborar una sucinta descripción, englobando tanto los métodos más comunmente utilizados como las estrategias más novedosas, entre las cuales se encuentran los métodos ya mencionados, protagonistas de los análisis comparativos que se llevarán a cabo en este proyecto.

2.2.1. Métodos univariantes y bivalentes

Inicialmente, las investigaciones basadas en *GWAS*, ya sea integrando la influencia de los diferentes *QTLs* o no, se realizaron con la finalidad de comprobar la asociación de los *SNPs* analizados con diferentes variantes genéticas mediante el estudio de como máximo un par de rasgos (variables o *traits*) de forma simultánea.

Para el caso univariante, el análisis es meramente descriptivo y trata de caracterizar el rasgo escogido, ya sea una variable cualitativa categórica o cuantitativa del conjunto de datos en cuestión, mediante métodos gráficos (*tablas de distribución de frecuencia*, *gráficos de barras*, *histogramas*, *gráficos circulares*), o bien a través de un *análisis de regresión* que trate de determinar cómo varía el atributo escogido con respecto al efecto individual de esa única variable. Siempre, sin buscar ningún tipo de relación entre esta y el resto del variables que conforman el conjunto de datos experimental [26, 27].

Aunque en parte vale la misma descripción para los métodos de análisis bivariado, estos se caracterizan por permitir evaluar hasta qué punto será viable predecir un valor para una posible variable dependiente, conociendo el valor de la otra (posiblemente independiente) mediante su correlación, y efectuando la consiguiente regresión lineal simple. Al igual que el análisis univariado, este puede ser descriptivo o inferencial, siendo el caso de análisis multivariado más sencillo pero que no resulta satisfactorio cuando el objetivo es, como en el tipo de estudios que contextualizan el presente trabajo, examinar de foma simultánea las posibles múltiples relaciones entre las múltiples variables del conjunto de datos considerado [26, 28].

Existen diversas formas para tratar de describir los patrones que se encuentran en los datos, habitualmente en formato de *sumario estadístico*, entre las cuales pueden incluirse los ya mencionados métodos gráficos y los siguientes estadísticos *medida de tendencia central*, *medida de variabilidad* y *estadísticos de dispersión* entre otros.

Tras todo lo anterior, resulta pues evidente que con este tipo de análisis solo se puede realizar estudios estadísticos meramente descriptivos de la variable o *trait* considerada en cada caso, y pese a que resultan útiles para determinar la calidad del elevado volumen de datos con el cual suele trabajarse en este tipo de estudios, tratar de inferir la influencia de la variable seleccionada con respecto al rasgo considerado dentro de un conjunto multivariable es una estrategia equivocada o, como mínimo, muy sesgada. Es en estas situaciones cuando probar la implementación de un análisis multivariante resulta más apropiado.

2.2.2. Métodos multivariantes: características y aplicaciones

Actualmente, debido a la gran cantidad de datos disponibles con perfiles genómicos complejos (alta diversidad de rasgos moleculares), la necesidad de encontrar correlaciones entre las diferentes variables analizables y los rasgos de interés ha desembocado en un crecimiento en la utilización de métodos multivariantes para su correcto análisis estadístico. Seguidamente se detallarán, de los más relevantes, sus características principales, los entornos de aplicación habituales y sus posibles desventajas:

- Los métodos que modelan el genotipo como variable dependiente comprobando a su vez la asociación con una suma ponderada de fenotipos (*MV-PLINK* ([2]) o análisis de correlación canónica, y *MultiPhen* [3] que utiliza la regresión ordinal) adolecen de la posibilidad de evaluar diseños complejos que presentan múltiples interacciones entre el genotipo y otras covariables.
- Tanto el análisis multivariante de la varianza (*MANOVA*), como el de los modelos multivariantes lineales mixtos (*mvLMMs*) [4], resultan ser más tolerantes a estos diseños complejos al tratar los fenotipos como variables dependientes, introduciendo de forma natural el posible parentesco genético entre los individuos analizados. Esta ventaja se torna inconveniente para grandes conjuntos de datos, sobre todo para el método *mvLMMs*, cuya continua mejora en su implementación computacional sigue requiriendo de tiempos excesivamente altos.
- La pluralidad de los métodos de regresión multivariante presuponen una normalidad en la distribución de los errores del modelo que puede no llegar a cumplirse. Todo y que pueden aplicarse transformaciones individuales a cada rasgo estudiado, no puede garantizarse la normalidad multivariante, lo que resulta en una reducción de la potencia estadística en comparación con el modelo aplicado a los rasgos no transformados.
- Hasta el momento, las diversas implementaciones de *métodos bayesianos* para el estudio de asociaciones multirasgo no han sido satisfactorias, requiriendo siempre un tiempo elevado de cálculo debido al coste computacional que implican.

- Para los métodos *MTAR* [5] o *MOSTest* [6] [7], existe la necesidad de garantizar la normalidad multivariante asintótica cuando se utilizan los sumarios estadísticos univariantes, lo que no es trivial. Sumado a que evitar la aparición de sesgos en la estimación de correlaciones de rasgos a partir de esta clase de estadísticos no es sencillo (afectaciones de heredabilidad de los rasgos, patrones de desequilibrio de ligamiento, etc.).

Considerando lo anteriormente expuesto, emerge de forma lógica la necesidad de implementar un método más adecuado a las características de los estudios estadísticos que deben llevarse a cabo en el marco de las investigaciones basadas en *GWAS* (con o sin influencia de los diferentes *QTLs*) que analizan las asociaciones *traits-SNPs*.


2.3. El modelo *PERMANOVA*

El modelo *PERMANOVA* ([8]) amplía el modelo lineal factorial univariante a múltiples dimensiones sin requerir una distribución de probabilidad conocida de las variables dependientes, introduciendo un enfoque basado en la distancia, poniendo a prueba la hipótesis de ausencia de efectos mediante un procedimiento de permutación basado en un estadístico *pseudo-F*, en el que las sumas de cuadrados del *ANOVA* se sustituyen por sumas de interdistancias entre observaciones.

=> Algo más de info + eq pseudo F

Pese a ser exitoso en muchos estudios, dando buenos resultados en un tiempo de cálculo reducido para diseños fijos unidireccionales, resulta inviable en los estudios actuales, donde el mayor tamaño y complejidad de los conjuntos de datos requiere una precisión para el cálculo del valor p que este procedimiento permutacional no puede alcanzar en las condiciones requeridas.

2.4. MANTA - Una implementación de la versión asintótica y no paramétrica de *PERMANOVA*

Gracias al programa *MANTA* ([12], desarrollado principalmente en ) se puede estudiar mediante la simulación de diversos escenarios de diseños complejos ([13]) la distribución asintótica de la estadística de pruebas *PERMANOVA* en el caso de la distancia euclídea (*valores p* de carácter no paramétrico y asintótico para modelos lineales multivariados), obteniendo resultados igualmente válidos tras cualquier transformación de los datos que preserve la independencia de las observaciones [9, 10, 11].

=> Algo más de info + eq pseudo F


Es mediante una implementación adaptada de *MANTA* a un estudio computacional comparativo a través de escenarios simulados con variables controladas, que se procederá a estudiar cómo varia la potencia estadística \mathbb{P} con respecto al método paramétrico *MANOVA*.

Capítulo 3

Metodología y Resultados

3.1. Objetivos finales del estudio

Tras los cambios en la planificación original, debidos principalmente a la alteración en el tiempo dedicado a las simulaciones necesarias, y detallados en las secciones *Planificación del trabajo* y *Desviaciones y acciones de mitigación*, se especificarán a continuación los objetivos de trabajo que finalmente han sido llevados a cabo de forma completa y satisfactoria:



- I. *Primer objetivo*: Estudiar la posible variación de la potencia estadística \mathbb{P} , es decir, de la probabilidad de no cometer un *error de tipo II*, entre el método basado en la versión asintótica de *PERMANOVA*, implementada en  mediante el paquete *MANTA*, y el del análisis multivariante de la varianza o *MANOVA* (aplicando la función *manova()* del paquete *Stats* [29]), bajo unas condiciones de simulación determinadas:
 - (a) Se comparará la \mathbb{P} de ambos métodos calculando la fracción de los *p-valores* de uno de los factores del conjunto de datos simulado (\mathbf{S} simulaciones del tipo $Y \sim A + B + AB$ bajo el modelo de distribución de datos *Multivariate normal* o *mvnorm*), en este caso el factor \mathbf{B} , que se encuentren por debajo del nivel de significación definido ($\alpha = 0.05$), con respecto a la variación controlada de la variable que determina la correlación de los datos (\mathbf{c} o \mathbf{Cor}), a la vez que estos son condicionados mediante cuatro tipos diferentes de *varianza* (*Var* o *v*). Dicho análisis se llevará a cabo sin aplicar ninguna transformación a los datos, y considerando una matriz de correlación homogénea con el mismo valor \mathbf{Cor} variable fuera de la diagonal.
 - (b) Repetición de las simulaciones del paso previo, pero aplicando las transformaciones *centered log-ratio* (*clr*), *log-ratio* y *raíz cuadrada*, con el fin de comprobar la idoneidad de estos métodos con dicho tipo de tratamiento de los datos. A su vez, se valorará la posible invarianza frente a la transformación de los datos de cada método por separado.
 - (c) Se contrastarán los resultados anteriores con la repetición del mismo tipo de simulaciones previamente realizadas pero bajo el condicionante de un nivel de significación menor al considerado, en particular para: $\alpha \in [0.01, 0.001]$.
 - (d) Similarmente al primer punto, se vuelve a comparar la \mathbb{P} de ambos métodos pero, ahora, imponiendo una matriz de correlación inhomogénea de valores aleatorios fuera de la diagonal a la vez que *Var* toma los cuatro valores impuestos. Para este caso se confrontarán gráficamente los resultados obtenidos para los diferentes valores del nivel de significación: $\alpha \in [0.05, 0.01, 0.001]$.

- II. *Segundo objetivo*: Estudio de la posible invarianza frente a la transformación de los datos del método asintótico *PERMANOVA*, implementado en *MANTA*, con respecto a su potencia estadística (\mathbb{P}). Teniendo en cuenta diferentes situaciones de simulación del conjunto de datos, mediante el uso de un *algoritmo simplex* con $n = 3$ [30, 31] (*3-simplex* o *tetraedro*).
- (a) Se usará una función de simulación de datos original, *Sim.simplex*, que implementa el algoritmo *3-simplex* deseado. Esta, a su vez, depende de otras dos funciones igualmente originales: *sim.simplex*, y *step2h1*. La implementación de las cuales está disponible en el archivo *fx.R* ([13]), desarrolladas *ex profeso* durante la elaboración del artículo [9].
 - (b) Se simularán y confrontarán los resultados de \mathbb{P} para todas las combinaciones $\Delta - q - Loc$ que sean posibles, siempre y cuando no se generen errores durante la ejecución del algoritmo implementado, estableciendo el nivel de significación en $\alpha = 0.05$.
 - (c) Se contrastará el resultado anterior con los obtenidos al repetir la simulación anterior para dos niveles de significación menores: $\alpha \in [0.01, 0.001]$.

3.2. Metodología

Durante todo el desarrollo del presente trabajo, se ha hecho uso de un computador de la marca *Apple* con las siguientes características principales de hardware:

- *Nombre del modelo*: MacBook Air
- *Identificador del modelo*: Mac14,15
- *Número de modelo*: MQKQ3Y/A
- *Chip*: Apple M2
- *Cantidad total de núcleos*: 8 (4 de rendimiento y 4 de eficiencia)
- *Memoria*: 8 GB

Tanto la implementación del código en  necesario para las simulaciones consideradas, como la creación y manipulación de los diversos archivos \LaTeX indispensables para conformar la memoria de este trabajo final de máster, se ha llevado a cabo mediante el hardware arriba especificado a través de la aplicación  Studio. A continuación se especifica la información relevante de la sesión de trabajo, obtenida mediante la función *toLatex(sessionInfo())*:

- *R version*: 4.3.1 (2023-06-16), aarch64-apple-darwin20
- *Locale*: en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
- *Time zone*: Europe/Madrid
- *TZcode source*: internal
- *Running under*: macOS Sonoma 14.1
- *Matrix products*: default

- **BLAS**: `/System/Library/Frameworks/Accelerate.framework/Versions/A/Frameworks/vecLib.framework/Versions/A/libBLAS.dylib`
- **LAPACK**: `/Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/lib/libRlapack.dylib` ; LAPACK version 3.11.0
- **Base packages**: base, datasets, graphics, grDevices, methods, stats, utils
- **Other packages**: arm 1.13-1, car 3.1-2, carData 3.0-5, common 1.1.1, compositions 2.0-6, CompQuadForm 1.4.3, copula 1.1-3, data.table 1.14.10, devtools 2.4.5, dplyr 1.1.3, fmsr 1.6.3, ggjoy 0.4.1, ggplot2 3.4.4, ggridges 0.5.5, glue 1.6.2, Hmisc 5.1-1, hms 1.1.3, knitr 1.44, latex2exp 0.9.6, lattice 0.21-8, lme4 1.1-35.1, manta 1.0.1, MASS 7.3-60, Matrix 1.6-4, permute 0.9-7, pheatmap 1.0.12, plyr 1.8.9, progress 1.2.3, randomcoloR 1.1.0.1, remotes 2.4.2.1, reshape2 1.4.4, stargazer 5.2.3, svMisc 1.2.3, tibble 3.2.1, usethis 2.2.2, vegan 2.6-4, versions 0.3, viridis 0.6.4, viridisLite 0.4.2
- **Loaded via a namespace (and not attached)**: abind 1.4-5, ADGofTest 0.3, backports 1.4.1, base64enc 0.1-3, bayesm 3.1-6, bookdown 0.37, boot 1.3-28.1, cachem 1.0.8, callr 3.7.3, checkmate 2.3.1, cli 3.6.2, cluster 2.1.4, coda 0.19-4, colorspace 2.1-0, compiler 4.3.1, crayon 1.5.2, curl 5.2.0, DEoptimR 1.1-3, digest 0.6.33, ellipsis 0.3.2, evaluate 0.22, fansi 1.0.5, farver 2.1.1, fastmap 1.1.1, foreign 0.8-84, Formula 1.2-5, fs 1.6.3, ganttrify 0.0.0.9009, generics 0.1.3, ggpp 0.5.5, grid 4.3.1, gridExtra 2.3, gsl 2.1-8, gtable 0.3.4, htmlTable 2.4.2, htmltools 0.5.6.1, htmlwidgets 1.6.2, httpuv 1.6.11, jsonlite 1.8.7, labeling 0.4.3, later 1.3.1, lifecycle 1.0.3, magrittr 2.0.3, memoise 2.0.1, mgcv 1.8-42, mime 0.12, miniUI 0.1.1.1, minqa 1.2.6, munsell 0.5.0, mvtnorm 1.2-4, nlme 3.1-162, nloptr 2.0.3, nnet 7.3-19, numDeriv 2016.8-1.1, parallel 4.3.1, pcaPP 2.0-4, pillar 1.9.0, pkgbuild 1.4.2, pkgconfig 2.0.3, pkgload 1.3.3, polynom 1.4-1, prettyunits 1.2.0, processx 3.8.2, profvis 0.3.8, promises 1.2.1, ps 1.7.5, pspline 1.0-19, purrr 1.0.2, R6 2.5.1, RColorBrewer 1.1-3, Rcpp 1.0.11, renv 1.0.3, rlang 1.1.1, rmarkdown 2.25, robustbase 0.99-1, rpart 4.1.19, rstudioapi 0.15.0, Rtsne 0.17, scales 1.2.1, sessioninfo 1.2.2, shiny 1.7.5, splines 4.3.1, stabledist 0.7-1, stats4 4.3.1, stringi 1.7.12, stringr 1.5.0, tensorA 0.36.2.1, tidyselect 1.2.0, tools 4.3.1, urlchecker 1.0.1, utf8 1.2.3, V8 4.4.1, vctrs 0.6.3, withr 2.5.1, xfun 0.40, xtable 1.8-4, yaml 2.3.7

=>Esquema simulaciones + tablas con variables cada simulación.

Los distintos parámetros de los escenarios de simulación implementados para el estudio de los puntos establecidos en el *Objetivo I* pueden encontrarse en las tablas 3.1 y 3.2.

Tabla 3.1: Simulaciones comparativas MANTA-MANOVA bajo el modelo de distribución *mvnorm* (*Objetivo I*), calculando la potencia estadística \mathbb{P} bajo un nivel de significación $\alpha = 0.05$ y con: $S = 1000$; $n = 300$; $q = 3$

Variable de simulación	Nombre	Valores
Nivel de significación	alpha	0.05
Número de simulaciones	S	1000
Tamaño de la muestra	n	300
Número de respuestas	q	3
Tipo de varianza	Var	Equal o Unequal (Type I, Type II, Type III)
Parámetro de generación de H_1	delta	De 0 a 0.35 (21 valores) Con un paso de 0.0175
Correlación de las variables	Cor	0, 0.2, 0.4, 0.6, 0.8

(a) Combinaciones $\Delta - Var$: imponiendo una matriz de correlación *homogénea*, con valores idénticos fuera de la diagonal, determinados por el valor de la variable **Cor**.

Variable de simulación	Nombre	Valores
Nivel de significación	alpha	0.05
Número de simulaciones	S	1000
Tamaño de la muestra	n	300
Número de respuestas	q	3
Tipo de varianza	Var	Equal o Unequal (Type I, Type II, Type III)
Parámetro de generación de H_1	delta	De 0 a 0.35 (21 valores) Con un paso de 0.0175
Correlación de las variables	Cor	Valores aleatorios

(b) Combinaciones $\Delta - Var$: imponiendo una matriz de correlación *inhomogénea*, con valores aleatorios fuera de la diagonal (opción implementada en la función *sim.mvnorm()*).

En cuanto al *Objetivo II*, en la tabla 3.3 se muestra la parametrización final que, tras las diversas pruebas previas, ha permitido la simulación de los escenarios deseados de forma satisfactoria.

Tabla 3.2: Simulaciones comparativas **MANTA-MANOVA** bajo el modelo de distribución *mvnorm* (*Objetivo I*), calculando la potencia estadística \mathbb{P} bajo unos niveles de significación estadística menores ($\alpha \in [0.01, 0.001]$) y con: $S = 1000$; $n = 300$; $q = 3$

Variable de simulación	Nombre	Valores	Variable de simulación	Nombre	Valores
Nivel de significación	alpha	0.01, 0.001	Nivel de significación	alpha	0.01, 0.001
Número de simulaciones	S	1000	Número de simulaciones	S	1000
Tamaño de la muestra	n	300	Tamaño de la muestra	n	300
Número de respuestas	q	3	Número de respuestas	q	3
Tipo de varianza	Var	Equal o Unequal (Type I, Type II, Type III)	Tipo de varianza	Var	Equal o Unequal (Type I, Type II, Type III)
Parámetro de generación de H_1	delta	De 0 a 0.35 (21 valores) Con un paso de 0.0175	Parámetro de generación de H_1	delta	De 0 a 0.35 (21 valores) Con un paso de 0.0175
Correlación de las variables	Cor	0, 0.2, 0.4, 0.6, 0.8	Correlación de las variables	Cor	Valores aleatorios

(a) Combinaciones $\Delta - Var$: imponiendo una matriz de correlación *homogénea*, con valores idénticos fuera de la diagonal, determinados por el valor de la variable **Cor**.

(b) Combinaciones $\Delta - Var$: imponiendo una matriz de correlación *inhomogénea*, con valores aleatorios fuera de la diagonal (opción implementada en la función *sim.mvnorm()*).

Tabla 3.3: Simulaciones para el estudio de la posible invarianza frente a la transformación de los datos del método asintótico *PERMANOVA*, implementado en *MANTA*, con respecto a su potencia estadística (\mathbb{P}). Teniendo en cuenta diferentes situaciones de simulación del conjunto de datos, mediante el uso de un *algoritmo simplex* con $n = 3$.

Variable de simulación	Nombre	Valores	Variable de simulación	Nombre	Valores
Nivel de significación	alpha	0.05	Nivel de significación	alpha	0.01, 0.001
Número de simulaciones	S	1000	Número de simulaciones	S	1000
Tamaño de la muestra	n	300	Tamaño de la muestra	n	300
Número de respuestas	q	3, 5	Número de respuestas	q	3, 5
Localización del modelo que genera el 3-simplex	loc	1, 2, 3, 5 ($\forall q = 3$) 1, 2, 3 ($\forall q = 5$)	Localización del modelo que genera el 3-simplex	loc	1, 2, 3, 5 ($\forall q = 3$) 1, 2, 3 ($\forall q = 5$)
Parámetro de generación de H_1	delta	De 0 a 0.025 (26 valores) Con un paso variable	Parámetro de generación de H_1	delta	De 0 a 0.025 (26 valores) Con un paso variable

(a) Combinaciones $\Delta - q - loc$: imponiendo un $\alpha \in 0.05$.

(b) Combinaciones $\Delta - q - loc$: imponiendo un $\alpha \in [0.01, 0.001]$.

3.3. Exposición de los resultados obtenidos

3.3.1. Resultados del primer objetivo

Tras las condiciones de simulación especificadas en la sección previa (*Metodología*), se obtuvieron un seguido de resultados que se mostrarán a continuación de la forma más detallada posible.

Como primer paso, y como alternativa a mostrar los primero o últimos valores, se detallará una tabla con valores extraídos aleatoriamente del conjunto de datos generado tras la computación de los escenarios expuestos en las tablas 3.1 y 3.2 (3.4, 3.5). En ella pueden encontrarse el valor de las variables involucradas en cada una de las simulaciones, así como el valor de la potencia estadística correspondiente, y el tiempo empleado en llevar a cabo cada iteración.

Tabla 3.4: Muestra aleatoria de 15 de los 5040 resultados obtenidos para los cálculos de la potencia estadística \mathbb{P} bajo las condiciones detalladas en las simulaciones 3.1a y 3.2a.

Datos	Modelo	alpha	S	n	Var	q	Cor	delta	Método	Potencia	t comp.(s)
Datos sin transformar	mvnorm	0.050	1,000	300	Unequal Type III	3	0.4	0.158	MANOVA	0.979	0.720
Transformación logarítmica	mvnorm	0.001	1,000	300	Equal	3	0.6	0.262	MANTA	0.241	1.390
Transformación Centered Log Ratio (clr)	mvnorm	0.050	1,000	300	Unequal Type II	3	0.4	0.315	MANOVA	0.092	1.530
Datos sin transformar	mvnorm	0.050	1,000	300	Unequal Type II	3	0.2	0.088	MANOVA	0.219	0.690
Datos sin transformar	mvnorm	0.010	1,000	300	Unequal Type III	3	0.2	0.210	MANOVA	0.500	0.660
Transformación Centered Log Ratio (clr)	mvnorm	0.050	1,000	300	Unequal Type II	3	0.8	0.140	MANTA	0.051	1.070
Datos sin transformar	mvnorm	0.050	1,000	300	Unequal Type I	3	0.0	0.018	MANOVA	0.085	0.690
Transformación logarítmica	mvnorm	0.001	1,000	300	Unequal Type I	3	0.8	0.018	MANOVA	0.008	0.650
Transformación Centered Log Ratio (clr)	mvnorm	0.010	1,000	300	Unequal Type I	3	0.2	0	MANOVA	0.948	0.730
Transformación Centered Log Ratio (clr)	mvnorm	0.010	1,000	300	Unequal Type II	3	0.8	0.315	MANTA	0.947	1.490
Datos sin transformar	mvnorm	0.010	1,000	300	Unequal Type III	3	0.8	0.262	MANTA	0.482	1.110
Transformación Centered Log Ratio (clr)	mvnorm	0.010	1,000	300	Equal	3	0.0	0.350	MANOVA	0.995	0.710
Transformación Centered Log Ratio (clr)	mvnorm	0.050	1,000	300	Equal	3	0.8	0.175	MANOVA	0.059	0.750
Transformación Centered Log Ratio (clr)	mvnorm	0.010	1,000	300	Unequal Type III	3	0.8	0.332	MANOVA	0.996	0.770
Datos sin transformar	mvnorm	0.010	1,000	300	Unequal Type II	3	0.8	0.350	MANTA	0.828	1.110

Tabla 3.5: Muestra aleatoria de 15 de los 1008 resultados obtenidos para los cálculos de la potencia estadística \mathbb{P} bajo las condiciones detalladas en las simulaciones 3.1b y 3.2b.

Datos	Modelo	alpha	S	n	Var	q	Cor	delta	Método	Potencia	t comp.(s)
Datos sin transformar	mvnorm	0.010	1,000	300	Unequal Type II	3	Aleat.	0.280	MANTA	0.560	1.280
Transformación Centered Log Ratio (clr)	mvnorm	0.010	1,000	300	Unequal Type I	3	Aleat.	0.018	MANOVA	0.997	0.810
Transformación Centered Log Ratio (clr)	mvnorm	0.050	1,000	300	Unequal Type II	3	Aleat.	0.280	MANOVA	0.999	0.770
Transformación raíz cuadrada	mvnorm	0.001	1,000	300	Unequal Type I	3	Aleat.	0.192	MANOVA	0.255	0.680
Transformación logarítmica	mvnorm	0.001	1,000	300	Unequal Type I	3	Aleat.	0.192	MANTA	0.086	1.140
Transformación logarítmica	mvnorm	0.010	1,000	300	Unequal Type II	3	Aleat.	0.070	MANOVA	0.104	0.690
Datos sin transformar	mvnorm	0.010	1,000	300	Unequal Type I	3	Aleat.	0.158	MANOVA	0.271	0.730
Transformación raíz cuadrada	mvnorm	0.001	1,000	300	Unequal Type I	3	Aleat.	0.088	MANTA	0.008	1.030
Transformación raíz cuadrada	mvnorm	0.001	1,000	300	Equal	3	Aleat.	0.245	MANTA	0.167	1.030
Datos sin transformar	mvnorm	0.001	1,000	300	Unequal Type III	3	Aleat.	0.158	MANTA	0.019	1.210
Transformación logarítmica	mvnorm	0.001	1,000	300	Unequal Type I	3	Aleat.	0.332	MANOVA	0.760	0.820
Transformación raíz cuadrada	mvnorm	0.001	1,000	300	Unequal Type II	3	Aleat.	0.175	MANOVA	0.191	0.770
Datos sin transformar	mvnorm	0.010	1,000	300	Equal	3	Aleat.	0.035	MANTA	0.022	1.030
Transformación logarítmica	mvnorm	0.010	1,000	300	Unequal Type III	3	Aleat.	0.332	MANOVA	0.892	0.730
Transformación Centered Log Ratio (clr)	mvnorm	0.001	1,000	300	Unequal Type III	3	Aleat.	0.018	MANOVA	0.986	0.770

Seguidamente, se mostrarán las características del estimador estadístico a estudio, la potencia \mathbb{P} de los métodos comparados (MANTA y MANOVA), teniendo en cuenta las diferentes

consideraciones del escenario de simulación: determinado por la computación de un conjunto de datos no transformados bajo una distribución *mvnorm*, aplicando una matriz de correlación *homogénea*, y considerando diferentes niveles de significación para el cálculo de \mathbb{P} .

Tabla 3.6: Descripción de los estadísticos potencia (\mathbb{P}) y tiempo de computación (*t comp.*), para los métodos **MANTA** y **MANOVA**, bajo una distribución *mvnorm*, con una matriz de correlación *homogénea*, y considerando diferentes niveles de significación.

Tipo de Datos	α	Statistic	N	Mean	St. Dev.	Min	Max
Datos sin transformar	0.05	Potencia	210	0.6686	0.3754	0.0410	1.0000
		tcomp	210	1.0975	0.0715	1.0157	1.5871
	0.01	Potencia	210	0.2722	0.2795	0.0140	0.8280
		tcomp	210	1.3319	1.5585	1.0310	17.5338
	0.001	Potencia	210	0.1373	0.1845	0.0030	0.5840
		tcomp	210	1.1096	0.0567	1.0255	1.3353

(a) Método **MANTA**.

Tipo de Datos	α	Statistic	N	Mean	St. Dev.	Min	Max
Datos sin transformar	0.05	Potencia	210	0.6823	0.3762	0.0520	1.0000
		tcomp	210	0.6850	0.0557	0.6396	1.3357
	0.01	Potencia	210	0.3979	0.3234	0.0200	0.9000
		tcomp	210	0.8146	1.3082	0.6232	17.4757
	0.001	Potencia	210	0.2820	0.2806	0.0020	0.7890
		tcomp	210	0.6706	0.0342	0.6323	0.7993

(b) Método **MANOVA**.

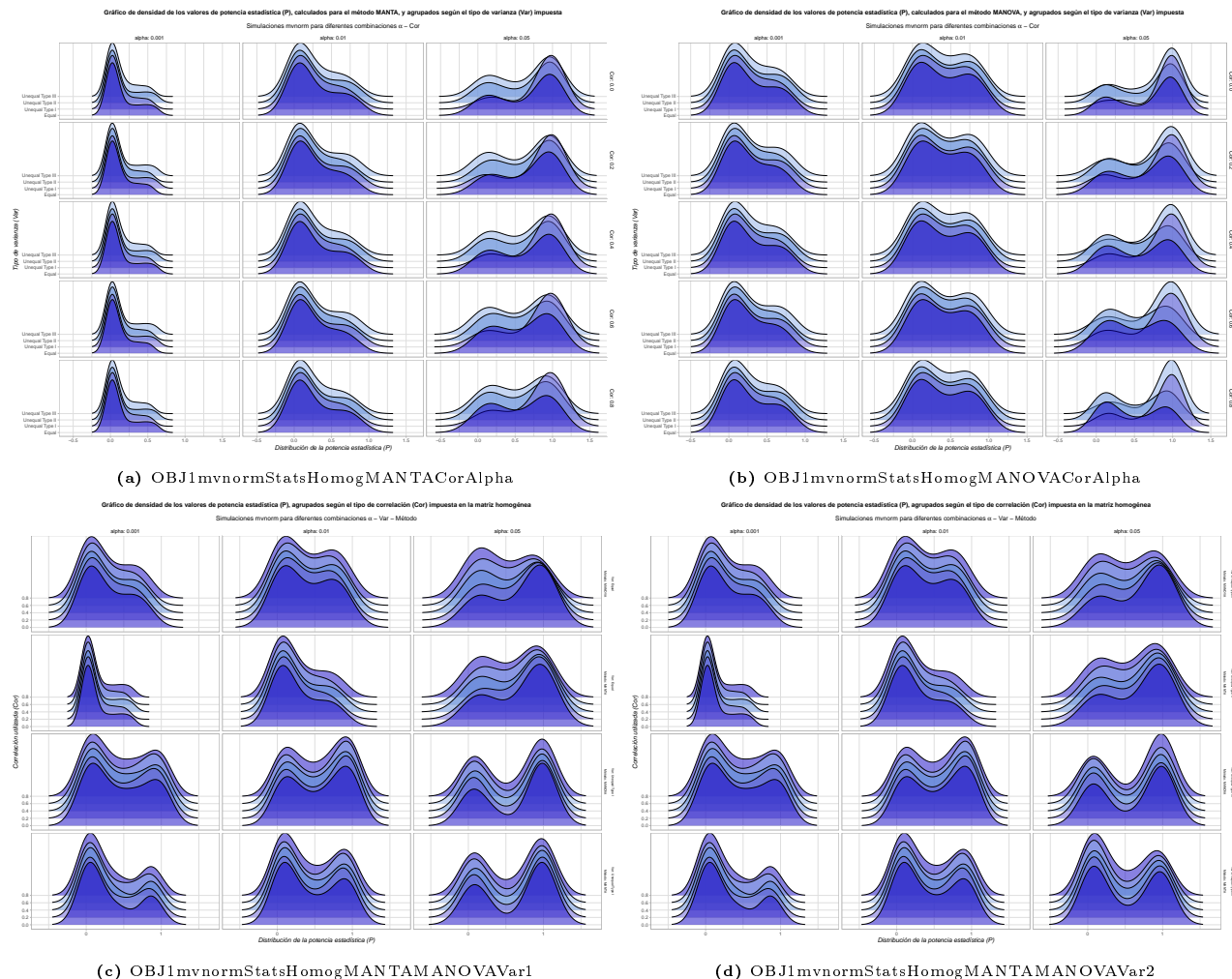


Figura 3.1: OBJ1mvnormStatsHomog

Complementariamente, se muestra también para el supuesto alternativo: determinado por el mismo escenario anterior pero con la salvedad de haber forzado la aplicación de una matriz de correlación *inhomogénea*, con valores aleatorios de la variable *Cor* fuera de la diagonal.

Aún a sabiendas de que algunas de las transformaciones que más típicamente se aplican al tratamiento de un conjunto de datos (*logarítmica* o *log-ratio*, *centered log-ratio* o *clr* y *raíz cuadrada*) no garantizan los supuestos de aplicación de los métodos a estudio, en particular de *MANOVA*, se llevaron a cabo las simulaciones pertinentes con el fin de verificar y, si cabe, cuantificar, dicho efecto. Las tablas correspondientes pueden encontrarse en el apéndice de tablas del presente documento (A.5, A.6, A.7 y A.8).

Para tratar de interpretar mejor los datos tabulados anteriormente expuestos, se realizarán representaciones gráficas de estos estadísticos en forma de comparativas de distribuciones de densidad, teniendo en cuenta diferentes combinaciones de las variables de simulación involucradas. Estas son complementarias a las tablas presentadas con anterioridad (??).

Tras esta exposición, se procederá a mostrar los resultados del estudio comparativo detallado previamente en forma de representaciones de mallas gráficas para las consideraciones del primer punto (a) del *Objetivo I*.

Tabla 3.7: Descripción de los estadísticos potencia (\mathbb{P}) y tiempo de computación ($t_{comp.}$), para los métodos **MANTA** y **MANOVA**, bajo una distribución *mvnorm*, con una matriz de correlación *inhomogénea*, y considerando diferentes niveles de significación.

Tipo de Datos	α	Statistic	N	Mean	St. Dev.	Min	Max
Datos sin transformar	0.05	Potencia	42	0.4257	0.3292	0.0620	0.9550
		tcomp	42	1.1242	0.1114	1.0503	1.6741
	0.01	Potencia	42	0.2722	0.2822	0.0140	0.8280
		tcomp	42	1.1125	0.0759	1.0149	1.2862
	0.001	Potencia	42	0.1373	0.1862	0.0030	0.5840
		tcomp	42	1.0848	0.0625	1.0200	1.3337

(a) Método **MANTA**.

Tipo de Datos	α	Statistic	N	Mean	St. Dev.	Min	Max
Datos sin transformar	0.05	Potencia	42	0.5154	0.3346	0.0620	0.9600
		tcomp	42	0.7004	0.0560	0.6452	0.9949
	0.01	Potencia	42	0.3979	0.3265	0.0200	0.9000
		tcomp	42	0.6956	0.0416	0.6473	0.8252
	0.001	Potencia	42	0.2820	0.2833	0.0020	0.7890
		tcomp	42	0.6897	0.0344	0.6532	0.8221

(b) Método **MANOVA**.

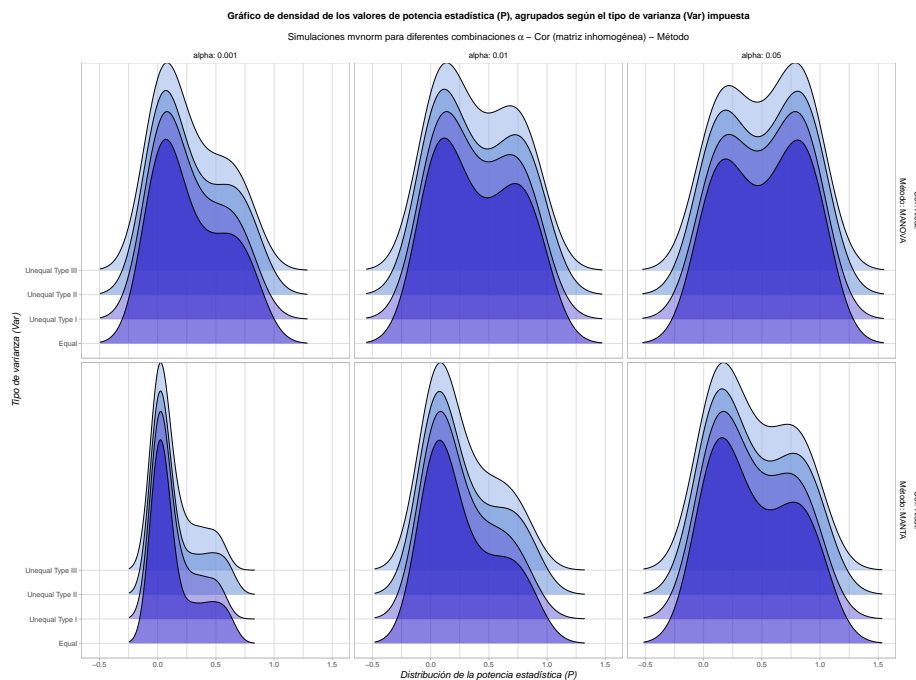


Figura 3.2: A.

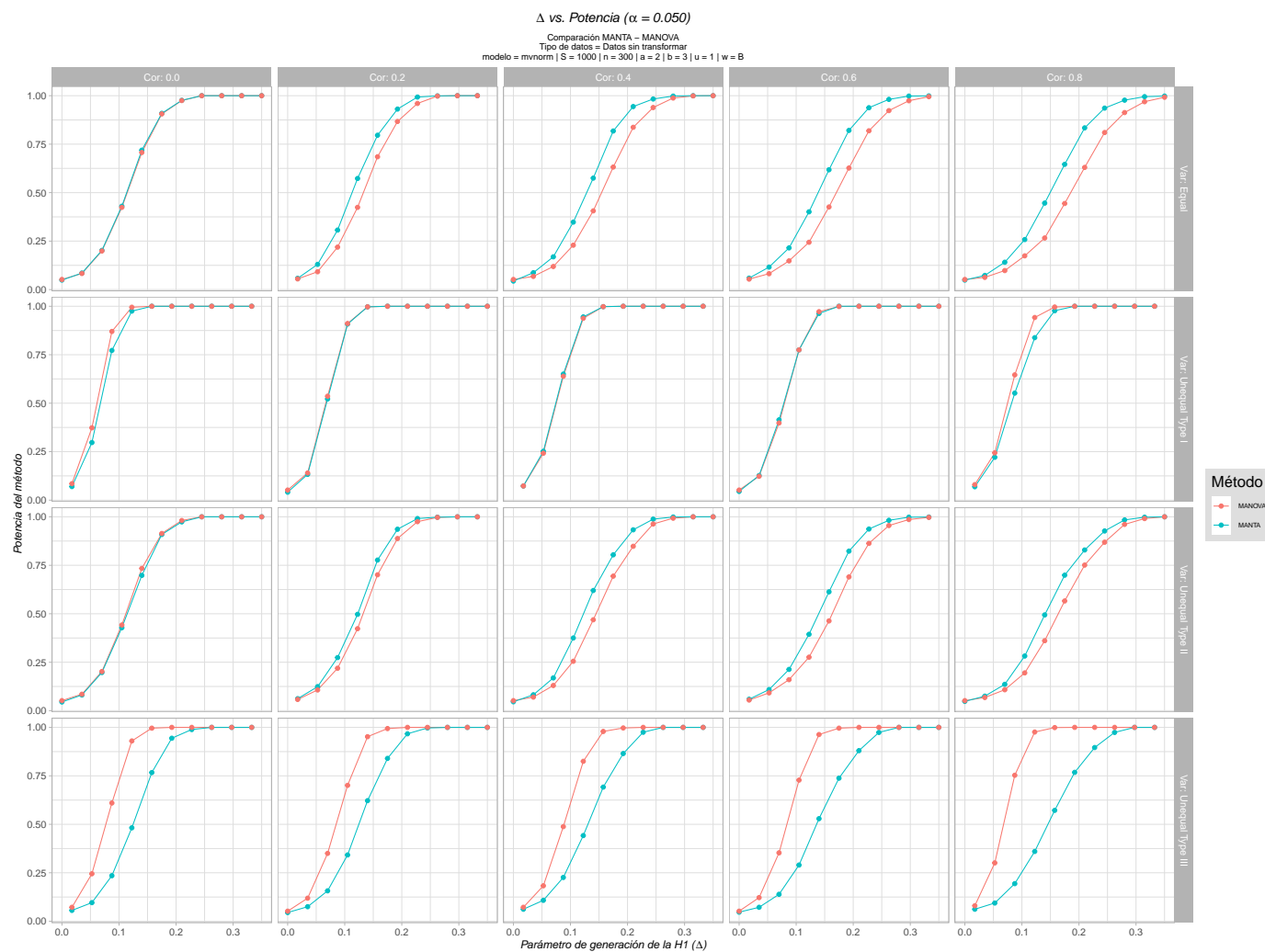


Figura 3.3: A.

Seguidamente, se presentan los productos resultantes de las simulaciones pertenecientes al punto *b*.

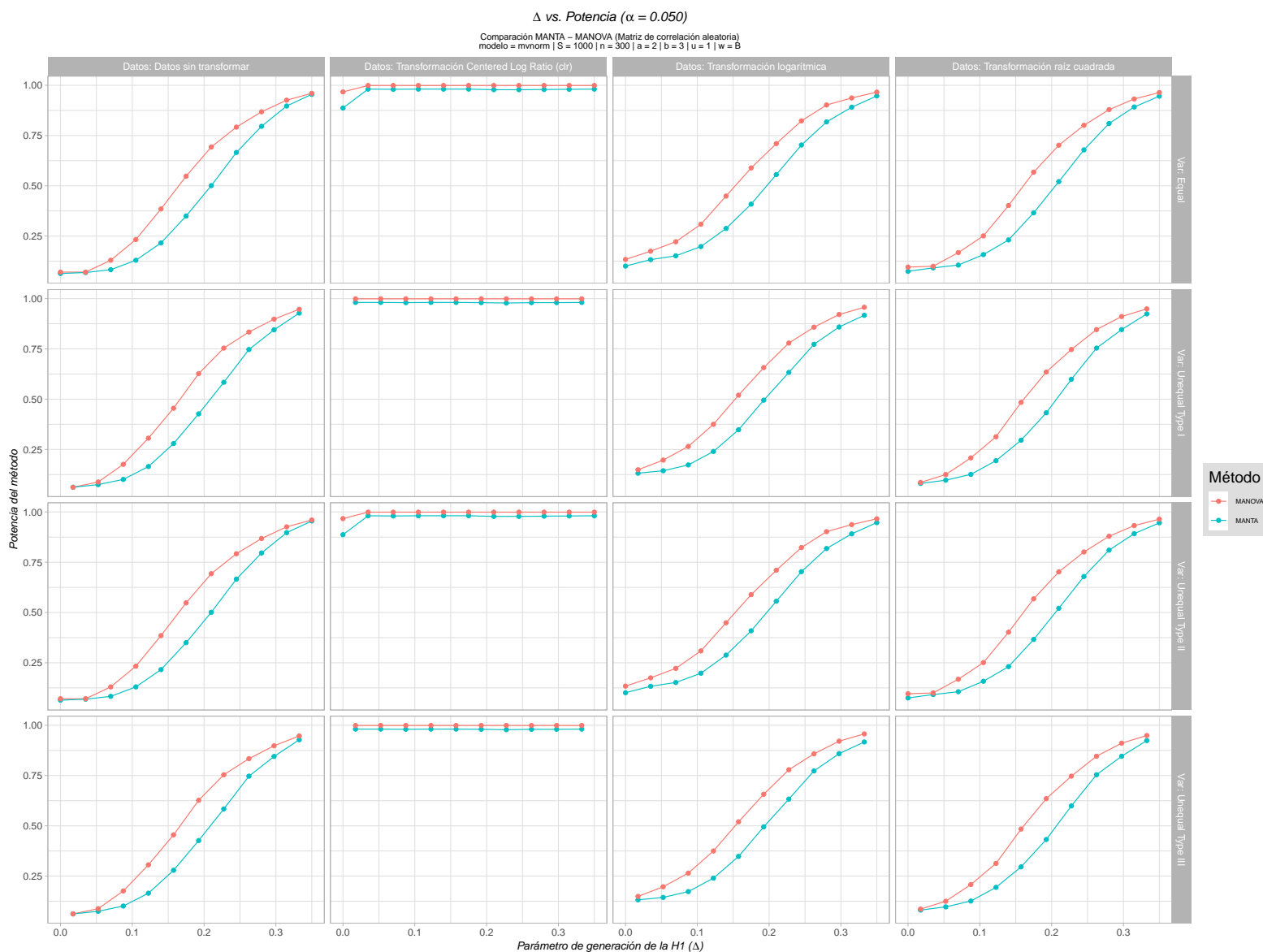
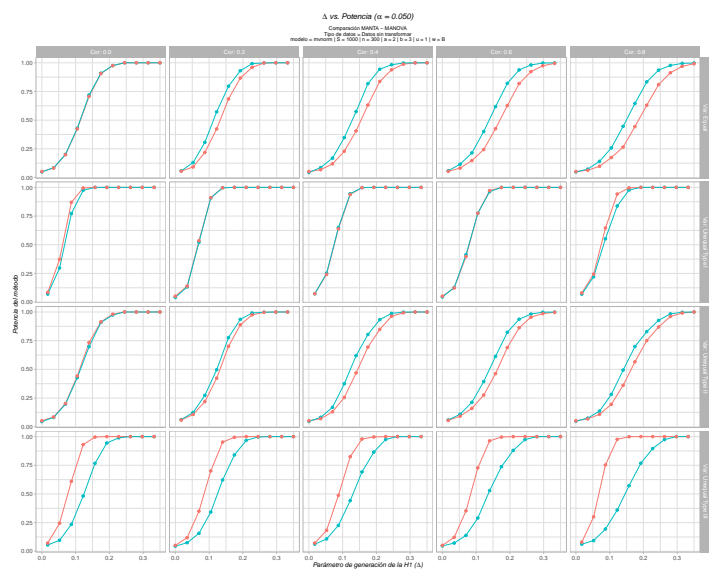


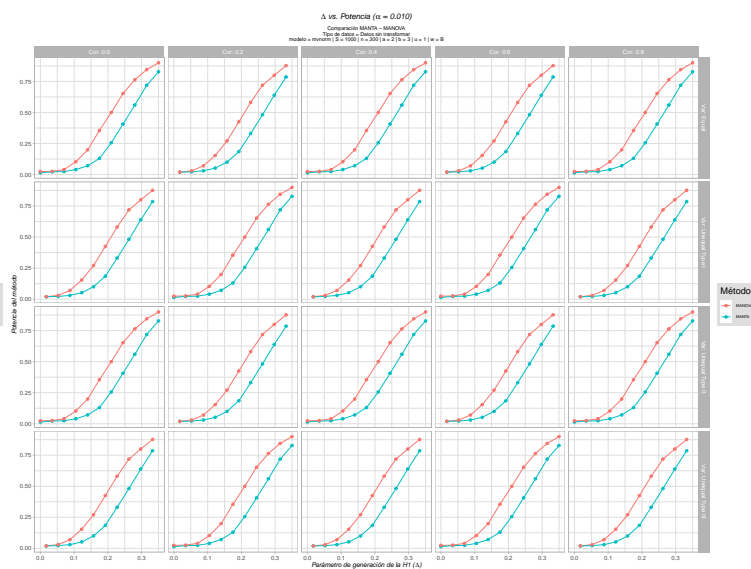
Figura 3.4: A.

De la misma forma, se añden los resultados gráficos de la computación de los escenarios determinados en *c*.

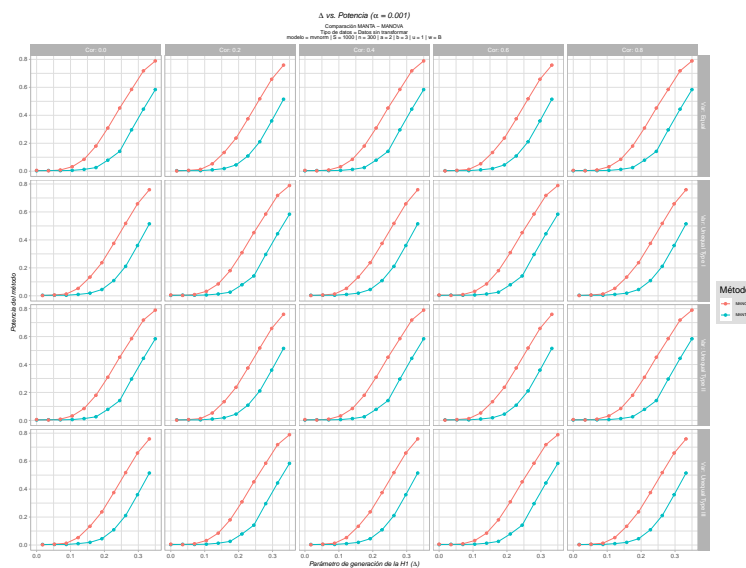
Finalmente, para concluir la exposición de los resultados obtenidos para los escenarios de simulación expuestos en el detallado del primer objetivo del presente trabajo, se añadirá del punto *b* del primer objetivo:



(a) A.



(b) A.



(c) A.

Figura 3.5: A.

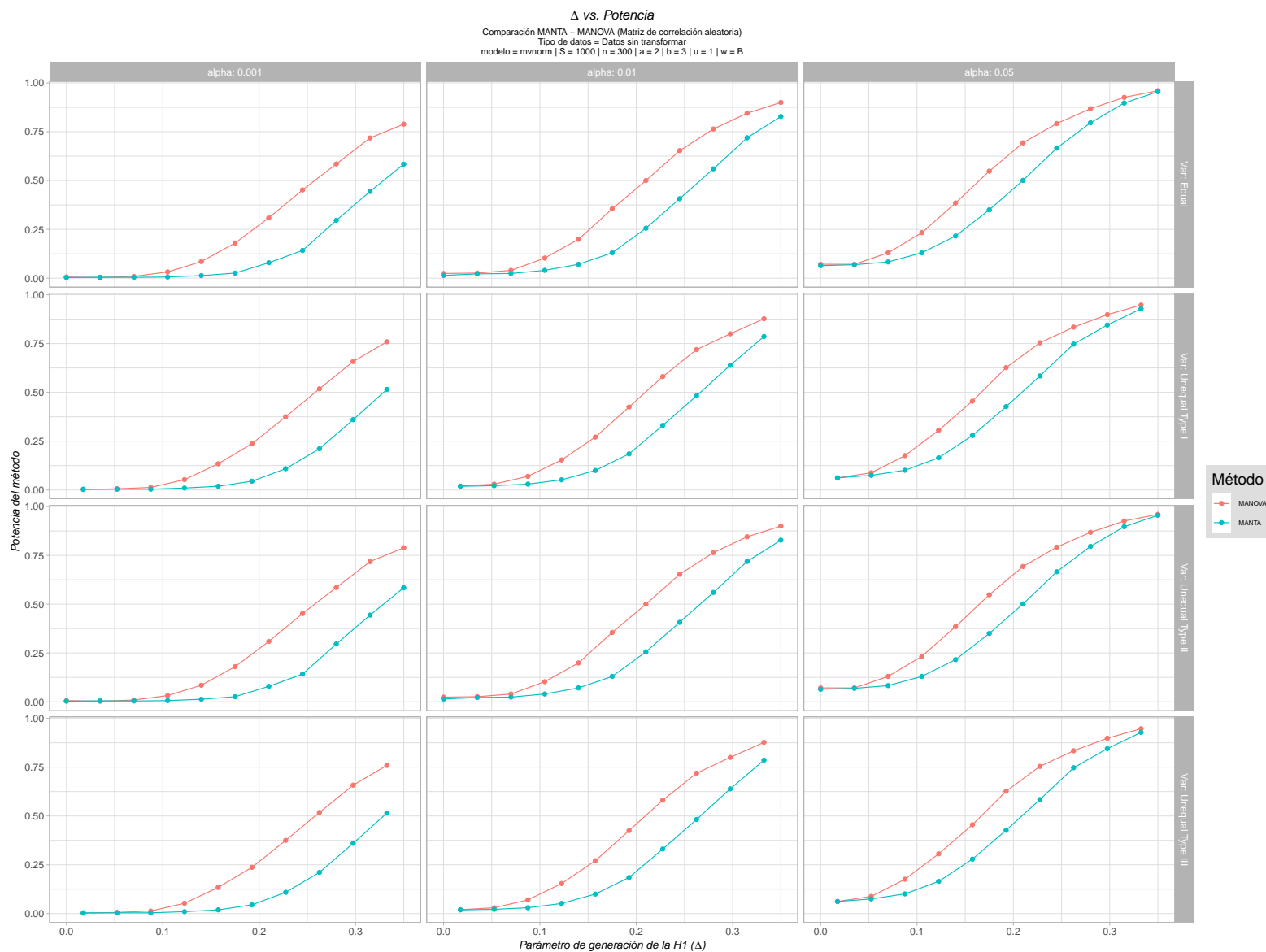


Figura 3.6: A.

3.3.2. Resultados del segundo objetivo

En esta subsección se detallarán los resultados del segundo objetivo procediendo, como en el caso anterior, a mostrar en forma de malla gráfica comparativa los diversos supuestos de simulación.

Previamente, se mostrará a modo de ejemplo una tabla con valores aleatorios del conjunto de datos generado tras la computación del escenario expuesto en 3.3a y 3.3b, el cual contiene información sobre el valor de las variables involucradas en cada una de las simulaciones, así como el valor del cálculo de la potencia estadística correspondiente, y el tiempo empleado:

Tabla 3.8: Muestra aleatoria de 15 de los 2144 resultados obtenidos para los cálculos de la potencia estadística \mathbb{P} bajo las condiciones detalladas en las simulaciones 3.3a y 3.3b.

Datos	Modelo	alpha	S	n	Var	q	Cor	delta	Método	Potencia	t comp.(s)
Transformación raíz cuadrada	simplex	0.0500	1,000	300	5	3	0.0260	0.0250	MANTA	1	13.1700
Transformación raíz cuadrada	simplex	0.0100	1,000	300	5	1	0.0270	0.0235	MANTA	1	10.4600
Datos sin transformar	simplex	0.0500	1,000	300	3	5	0.0230	0.0005	MANTA	0.0480	1.1400
Transformación raíz cuadrada	simplex	0.0500	1,000	300	3	3	0.0240	0.0240	MANTA	1	17.3500
Datos sin transformar	simplex	0.0500	1,000	300	5	3	0.0260	0.0090	MANTA	0.7960	1.1000
Transformación logarítmica	simplex	0.0500	1,000	300	5	3	0.0260	0.0170	MANTA	1	1
Transformación Centered Log Ratio (clr)	simplex	0.0010	1,000	300	3	5	0.0230	0.0085	MANTA	0.0900	1.0500
Transformación raíz cuadrada	simplex	0.0100	1,000	300	3	5	0.0230	0.0080	MANTA	0.5310	1.4400
Transformación Centered Log Ratio (clr)	simplex	0.0500	1,000	300	5	2	0.0270	0.0175	MANTA	0.9970	1.0500
Datos sin transformar	simplex	0.0010	1,000	300	3	3	0.0240	0.0140	MANTA	0.9710	1.0900
Transformación logarítmica	simplex	0.0010	1,000	300	3	5	0.0230	0.0045	MANTA	0.0260	0.9900
Datos sin transformar	simplex	0.0010	1,000	300	3	5	0.0230	0.0005	MANTA	0	1.0700
Transformación logarítmica	simplex	0.0100	1,000	300	3	2	0.0240	0.0220	MANTA	1	1
Transformación raíz cuadrada	simplex	0.0500	1,000	300	3	5	0.0230	0.0165	MANTA	1	14.2000
Datos sin transformar	simplex	0.0010	1,000	300	5	1	0.0270	0.0020	MANTA	0.0010	1.1300

Además, se han realizado un seguido de sumarios del estimador estadístico principal (\mathbb{P}), a la vez que del tiempo de computación ($t_{comp.}$), para los diferentes tipos de subgrupos de simulación definidos por las transformaciones de datos aplicadas, con el fin de poder aportar más información a la interpretación de los resultados que se mostrarán más adelante.

Para una mejor interpretación, se han realizado representaciones gráficas de estos estadísticos en forma de distribuciones de densidad, forzando la comparativa entre diversas variables para garantizar una correcta visualización de las posibles dependencias entre ellas (3.7a y 3.7b).

Tabla 3.9: Sumario estadístico (para diferentes valores del nivel de significación α) de la potencia estadística (\mathbb{P}) calculada, y del tiempo de computación empleado en las simulaciones *3-simplex*, sin aplicar al conjunto de datos ninguna transformación.

Statistic	N	Mean	St. Dev.	Min	Max
Potencia	179	0.7593	0.3475	0.0470	1.0000
tcomp	179	1.1578	0.0798	1.0764	1.4821

(a) Combinaciones $\Delta - q - loc$: imponiendo un $\alpha \in 0.05$.

Statistic	N	Mean	St. Dev.	Min	Max
Potencia	178	0.6971	0.3906	0.0050	1.0000
tcomp	178	1.1214	0.0620	1.0489	1.3689

(b) Combinaciones $\Delta - q - loc$: imponiendo un $\alpha \in 0.01$.

Statistic	N	Mean	St. Dev.	Min	Max
Potencia	179	0.6277	0.4212	0.0000	1.0000
tcomp	179	1.1132	0.0546	1.0469	1.3755

(c) Combinaciones $\Delta - q - loc$: imponiendo un $\alpha \in 0.001$.

Tabla 3.10: Sumario estadístico (para diferentes valores del nivel de significación α) de la potencia estadística (\mathbb{P}) calculada, y el tiempo de computación empleado en las simulaciones *3-simplex*, aplicando al conjunto de datos una transformación logarítmica.

Statistic	N	Mean	St. Dev.	Min	Max
Potencia	179	0.7516	0.3514	0.0470	1.0000
tcomp	179	1.0767	0.0917	0.9947	1.6458

(a) Combinaciones $\Delta - q - loc$: imponiendo un $\alpha \in 0.05$.

Statistic	N	Mean	St. Dev.	Min	Max
Potencia	178	0.6878	0.3942	0.0060	1.0000
tcomp	178	1.0395	0.0462	0.9804	1.2416

(b) Combinaciones $\Delta - q - loc$: imponiendo un $\alpha \in 0.01$.

Statistic	N	Mean	St. Dev.	Min	Max
Potencia	179	0.6176	0.4236	0.0000	1.0000
tcomp	179	1.0469	0.0558	0.9882	1.3023

(c) Combinaciones $\Delta - q - loc$: imponiendo un $\alpha \in 0.001$.

Tabla 3.11: Sumario estadístico (para diferentes valores del nivel de significación α) de la potencia estadística (\mathbb{P}) calculada, y el tiempo de computación empleado en las simulaciones *3-simplex*, aplicando al conjunto de datos una transformación *Centered Log Ratio* (*clr*).

Statistic	N	Mean	St. Dev.	Min	Max
Potencia	179	0.6964	0.3745	0.0510	1.0000
tcomp	179	1.1206	0.0843	1.0358	1.5834

(a) Combinaciones $\Delta - q - loc$: imponiendo un $\alpha \in 0.05$.

Statistic	N	Mean	St. Dev.	Min	Max
Potencia	178	0.6112	0.4158	0.0050	1.0000
tcomp	178	1.0879	0.0470	1.0276	1.2818

(b) Combinaciones $\Delta - q - loc$: imponiendo un $\alpha \in 0.01$.

Statistic	N	Mean	St. Dev.	Min	Max
Potencia	179	0.5194	0.4351	0.0000	1.0000
tcomp	179	1.0762	0.0375	1.0300	1.2588

(c) Combinaciones $\Delta - q - loc$: imponiendo un $\alpha \in 0.001$.

Tabla 3.12: Sumario estadístico (para diferentes valores del nivel de significación α) de la potencia estadística (\mathbb{P}) calculada, y el tiempo de computación empleado en las simulaciones *3-simplex*, aplicando al conjunto de datos una transformación de raíz cuadrada (*sqrt*).

Statistic	N	Mean	St. Dev.	Min	Max
Potencia	179	0.7389	0.3574	0.0470	1.0000
tcomp	179	5.2245	12.1286	0.9998	95.8778

(a) Combinaciones $\Delta - q - loc$: imponiendo un $\alpha \in 0.05$.

Statistic	N	Mean	St. Dev.	Min	Max
Potencia	178	0.6708	0.4002	0.0060	1.0000
tcomp	178	5.4130	13.1872	0.9952	101.6103

(b) Combinaciones $\Delta - q - loc$: imponiendo un $\alpha \in 0.01$.

Statistic	N	Mean	St. Dev.	Min	Max
Potencia	179	0.5969	0.4278	0.0000	1.0000
tcomp	179	5.2086	12.1741	0.9893	96.1683

(c) Combinaciones $\Delta - q - loc$: imponiendo un $\alpha \in 0.001$.

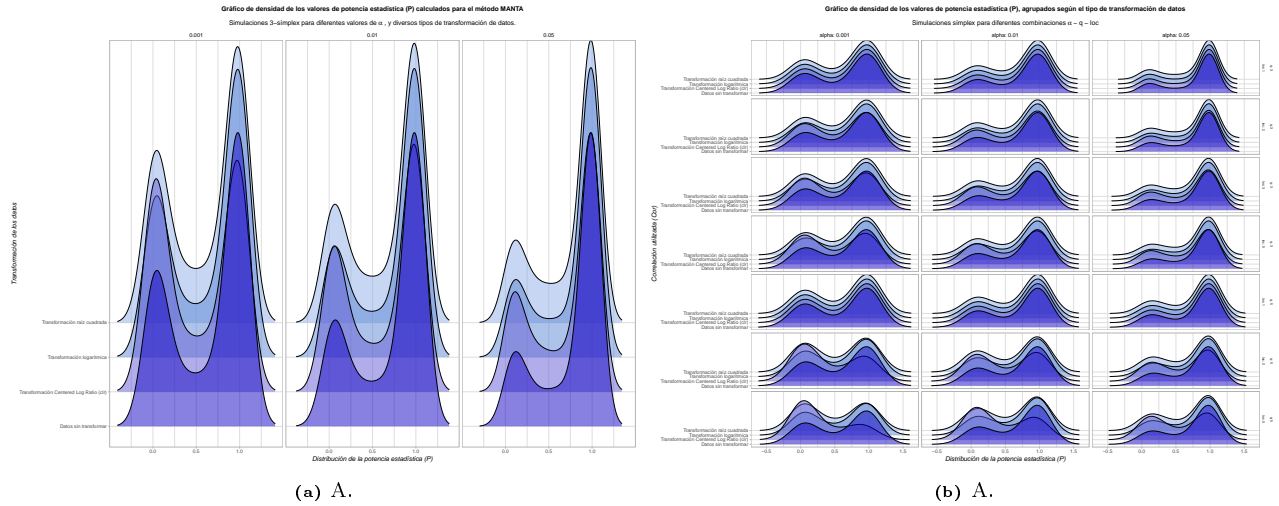


Figura 3.7: OBJ1mvnormStatsHomog

Tras lo anterior, y como primer resultado gráfico, se podrá encontrar el estudio de la posible invarianza frente a la transformación de los datos del método asintótico *PERMANOVA* con respecto a su potencia estadística \mathbb{P} (calculada bajo un nivel de significación $\alpha = 0.05$), teniendo en cuenta las siguientes variaciones $\Delta - q - Loc$ en la simulación del conjunto de datos mediante el uso de un *3-símplex*.

Para poder apreciar con más detalle la separación de los puntos de datos presentados en la figura 3.8, se han realizado ampliaciones de las curvas agrandando las zonas pertenecientes a sendas colas: *izquierda* (*izqda.*) con valores bajos de Δ (3.9a), y la que muestra sus valores más altos (*dcha.*) disponible en 3.9b.

Como ya se avanzó en el detallado del segundo objetivo en la sección 3.1, se repitieron las simulaciones anteriores pero realizando el cálculo de \mathbb{P} bajo niveles de significación más bajos, en particular: $\alpha \in [0.01, 0.001]$

Similarmente al caso previo, para $\alpha = 0.01$ se obtuvieron las figuras 3.10 y 3.11, y para $\alpha = 0.001$ los gráficos que se encuentran disponibles en 3.12, 3.13.

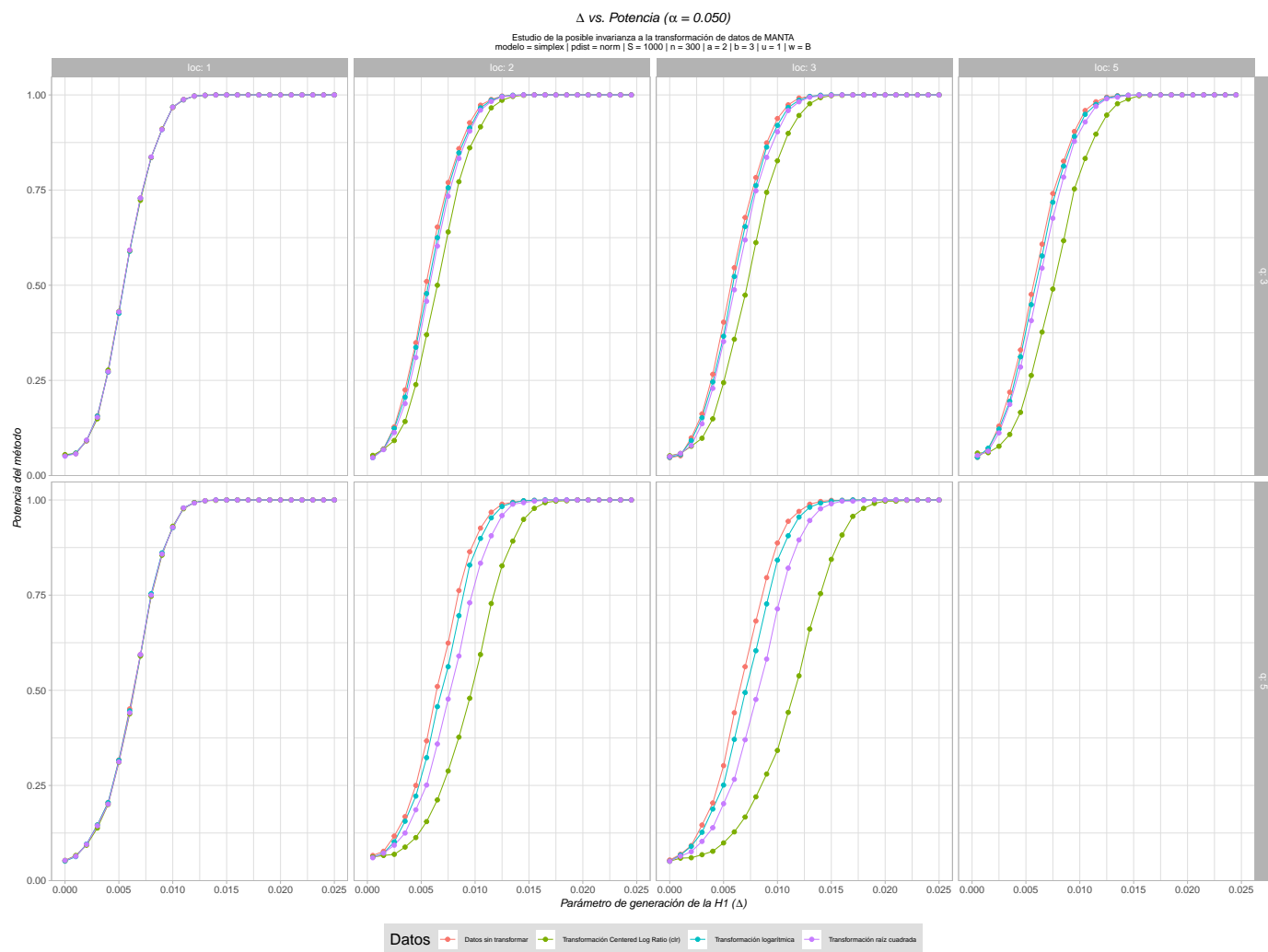
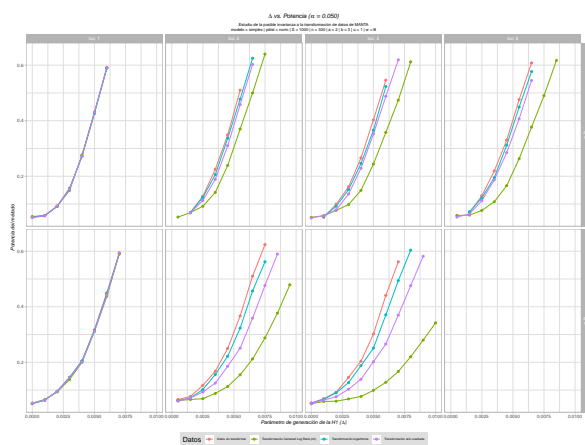
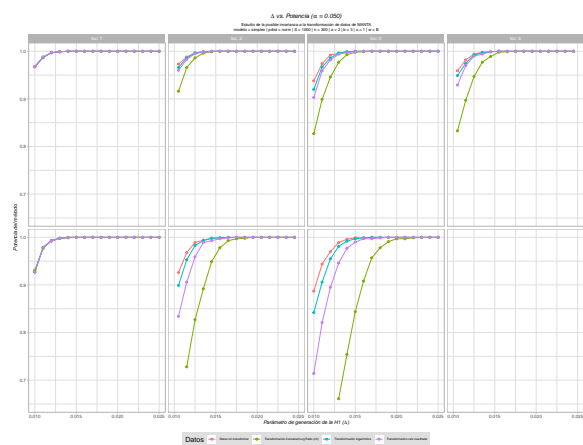


Figura 3.8: A.



(a) A.



(b) A.

Figura 3.9: A.

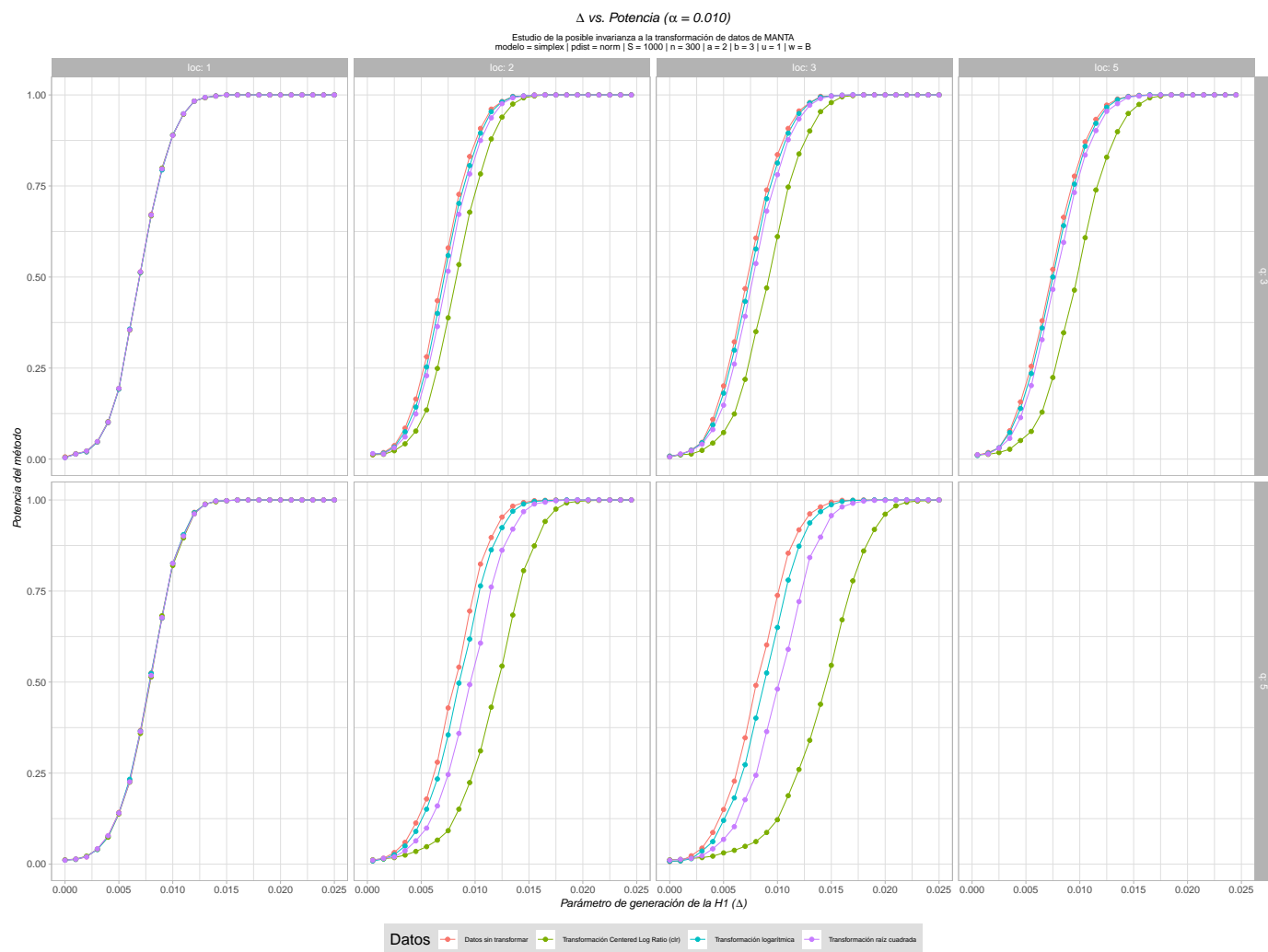
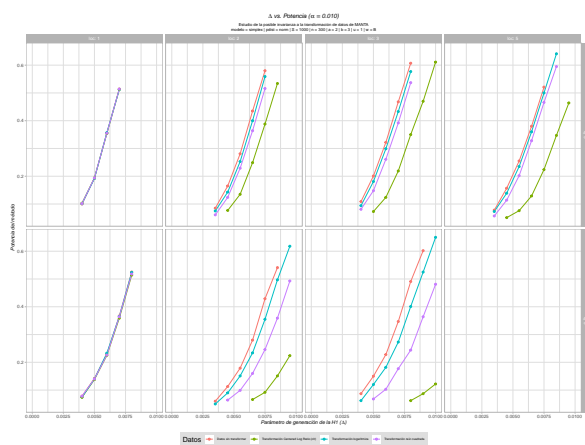
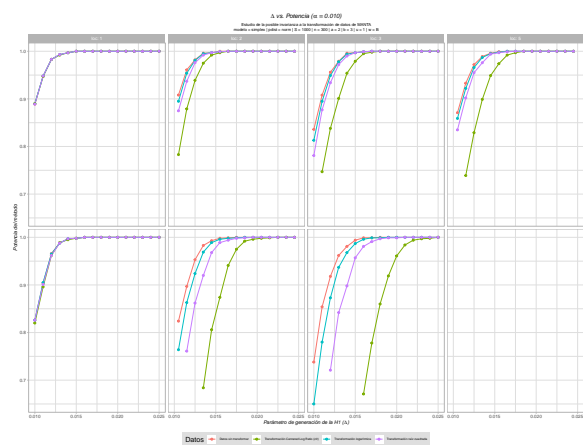


Figura 3.10: A.



(a) A.



(b) A.

Figura 3.11: A.

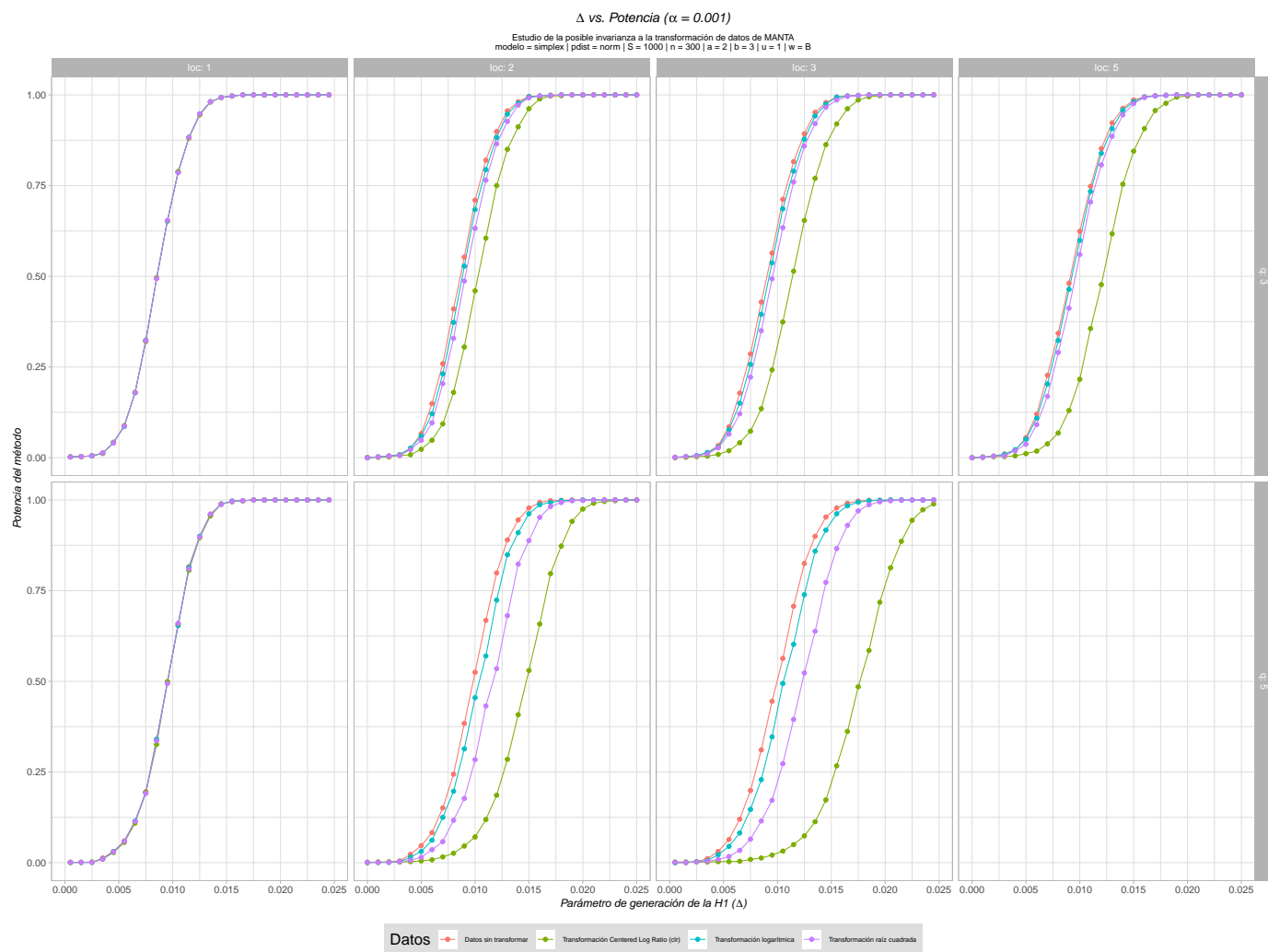
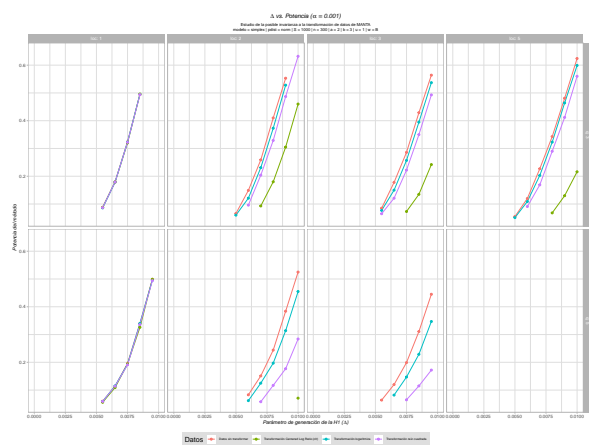
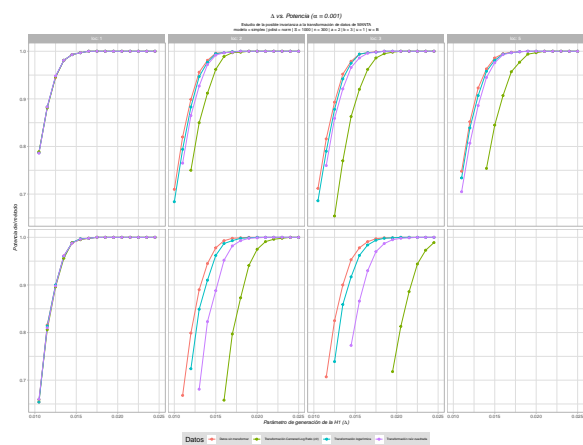


Figura 3.12: A.



(a) A.



(b) A.

Figura 3.13: A.

Capítulo 4

Discusión

...

Capítulo 5

Conclusiones y trabajos futuros

5.1. Conclusiones

De los resultados obtenidos para los diferentes escenarios planteados en el presente estudio, disponibles en las distintas secciones del capítulo *Resultados*, se puede resolver lo siguiente:

(I) Primera conclusión...

5.2. Líneas de futuro

Aunque solo sea por completitud del presente proyecto, preveemos como mínimo una línea de trabajo futuro: llevar a cabo el tercer objetivo planteado originalmente y que, por los motivos ya comentados anteriormente (*Desviaciones y acciones de mitigación* y *Análisis de riesgos*), no pudo llevarse a cabo.

Teniendo este punto en cuenta, y tras valorar los resultados obtenidos en el estudio cuantitativamente comparativo de la *potencia estadística* (\mathbb{P}) de sendos métodos estadísticos multivariantes, *MANOVA* y la versión asintótica de *PERMANOVA*, se determina que:

(I) Primera línea de trabajo futuro...

5.3. Seguimiento de la planificación

A modo de conclusión, se detallará el seguimiento real de la planificación original del trabajo, resaltando los escollos que fueron surgiendo, y las acciones de mitigación que se llevaron a cabo para subsanar los desvíos que estos causaban en los tiempos estimados para cada tarea programada:

- A

Finalmente, resulta conveniente hacer una valoración global y personal del proyecto llevado a cabo. ...

Glosario

algoritmo simplex Un algoritmo *simplex* es un conjunto de métodos en los cuales de alguna manera se busca el máximo de una función lineal sobre un conjunto de variables que satisfaga un conjunto de inecuaciones lineales. El algoritmo simplex primal o de George Dantzig (1947), procede examinando vértices adyacentes del poliedro de soluciones. Un algoritmo simplex es de alguna manera un algoritmo de pivote. Otro método alternativo es el de Nelder-Mead (1965) o método de descenso (o ascenso) simplex, método numérico que busca un mínimo (o máximo) local de una función cualquiera examinando en cada paso los vértices de un simplex. [4](#), [5](#), [31](#), [33](#)

análisis de regresión Proceso estadístico para entender cómo depende el atributo a estudiar de la variable escogida, y que gráficamente describe una línea que muestra la relación entre ambas. [27](#)

error de tipo II En un estudio de investigación, el error de tipo II, también llamado error de tipo beta (donde β es la probabilidad de que exista este error) o falso negativo, se comete cuando el investigador no rechaza la hipótesis nula (H_0 : el supuesto inicial) siendo esta falsa en la población. Es equivalente a la probabilidad de un resultado falso negativo, ya que el investigador llega a la conclusión de que ha sido incapaz de encontrar una diferencia que existe en la realidad. De forma general y dependiendo de cada caso, se suele aceptar en un estudio que el valor del error beta esté entre el 5 y el 20 %. [30](#)

estadística multivariante La estadística multivariante o multivariada es una rama de las estadísticas que abarca la observación y el análisis simultáneos de más de una variable respuesta. La aplicación de la estadística multivariante es llamada análisis estadístico multivariante. [14](#)

estadísticos de dispersión Las medidas de dispersión (variabilidad o propagación) son un número real no negativo que es cero si todos los datos son iguales y aumenta a medida que los datos se vuelven más diversos, y tratan de cuantificar el grado con el que una distribución se estira o se comprime. Las más comunes son: la varianza, la desviación estándar y el rango intercuartil. Suelen contrastarse con la ubicación o la tendencia central, y juntas son las propiedades más utilizadas de las distribuciones. [28](#)

gráficos circulares Gráfico que representa en un círculo dividido en porciones las frecuencias o porcentajes relativos de una población o una muestra que pertenece a diferentes categorías. [27](#)

gráficos de barras El gráfico de barras es un gráfico que consiste en barras rectangulares, las cuales representan un número o porcentaje de observaciones de categorías existentes en una variable. La longitud o la altura de las barras da una representación visual de las diferencias proporcionales entre las categorías. [27](#)

GWAS En genética, un estudio de asociación del genoma completo (Genome-wide association study) o GWAS (Whole genome association study) es un análisis de una variación genética a lo largo de todo el genoma humano con el objetivo de identificar su asociación a un rasgo observable. Los GWAS suelen centrarse en asociaciones entre los polimorfismos de un solo nucleótido (SNPs) y rasgos como las principales enfermedades. [14](#)

histogramas Los histogramas se utilizan para estimar la distribución de los datos, con la frecuencia de los valores asignados a un rango de valores llamado contenedor. [27](#)

MANOVA En estadística el análisis multivariante de la varianza (Multivariate analysis of variance) es una extensión del análisis de la varianza o ANOVA para cubrir los casos donde hay más de una variable dependiente que no pueden ser combinadas de manera simple. Además de identificar si los cambios en las variables independientes tienen efectos significativos en las variables dependientes, la técnica también intenta identificar las interacciones entre las variables independientes y su grado de asociación con las dependientes. [15](#), [28](#)

MANTA Multivariate Asymptotic Non-parametric Test of Association. Este paquete, programado en lenguaje R, permite el cálculo no paramétrico y asimptótico del p-valor para modelos lineales multivariados. [16](#), [29](#)

medida de tendencia central Medidas descriptivas numéricas comunmente utilizadas para estimar la ubicación central de los datos univariados mediante el cálculo de la media, la mediana y el modo. La media tiene la ventaja de que su cálculo incluye cada valor del conjunto de datos, pero es particularmente susceptible a la influencia de los valores atípicos. La mediana es una mejor medida cuando el conjunto de datos contiene valores atípicos. El modo es fácil de localizar, y es la única medida de tendencia central que se puede utilizar si los datos que se analizan son categóricos. Sin embargo, si los datos son de naturaleza numérica, podrán utilizarse los tres tipos para describir los datos. [28](#)

medida de variabilidad La medida de variabilidad o de la dispersión (desviación de la media) de un conjunto de datos univariable puede revelar la forma de una distribución de datos univariable de manera más suficiente, proporcionando algo más de información sobre la variación entre los valores de los datos. Junto con las medidas de tendencia central dan una mejor imagen de los datos que las medidas de tendencia central por sí solas. Las más utilizadas son: el rango, la varianza y la desviación estándar. Dependiendo su idoneidad del tipo de datos, la forma de la distribución de los datos y de la medida de tendencia central que se esté utilizando. Para datos son categóricos, entonces no hay ninguna medida de variabilidad que informar. Para los numéricos, las tres son posibles, teniendo en cuenta que si la distribución de los datos es simétrica, entonces las medidas de variabilidad más oportunas son la varianza y la desviación estándar, mientras que si los datos están sesgados, la más apropiada será el rango. [28](#)

MOSTest Es una herramienta para unir el análisis genético de múltiples rasgos, que utiliza el análisis multivariado para aumentar la potencia, y así poder descubrir los loci asociados. [15](#), [29](#)

MTAR Marco desarrollado para el análisis multi-trait de RVAS. Se basa en un meta-modelo analítico de efectos aleatorios que utiliza diferentes estructuras de correlación de los efectos genéticos para representar un amplio espectro de patrones de asociación a través de rasgos y variantes. [15](#), [29](#)

MultiPhen Paquete de R que permite testear la asociación de múltiples rasgos. Realiza pruebas de asociación genética entre SNPs y múltiples fenotipos (por separado o en conjunto). [15](#), [28](#)

mvLMMs Los modelos lineales mixtos multivariados son poderosas herramientas para probar asociaciones entre polimorfismos de núcleo único y múltiples fenotipos correlacionados mientras controlan la estratificación de la población en estudios de asociación de todo el genoma. [15](#), [28](#)

PERMANOVA El análisis multivariante de la varianza con permutaciones (Permutational multivariate analysis of variance, PERMANOVA) es una prueba de permutación estadística multivariada no paramétrica. Se utiliza para comparar grupos de objetos y probar la hipótesis nula de que los centroides y la dispersión de los grupos definidos por el espacio de medida son equivalentes para todos los grupos. Un rechazo de la hipótesis nula significa que el centro y/o la dispersión de los objetos es diferente entre los grupos. De esta manera, la prueba se basa en el cálculo previo de la distancia entre cualesquier dos objetos incluidos en el experimento. [14](#)

pleiotropía En biología, la pleiotropía o polifenia es el fenómeno por el cual un solo gen o alelo es responsable de efectos fenotípicos o caracteres distintos y no relacionados (e.g. la fenilcetonuria, la talasemia o anemia de células falciformes, o el albinismo de los animales que tiene un efecto pleitrópico en sus emociones haciéndolos más reactivos a su entorno). [14](#)

potencia estadística La potencia o poder de una prueba estadística es la probabilidad de que la hipótesis alternativa sea aceptada cuando la hipótesis alternativa es verdadera, es decir, la probabilidad de no cometer un error del tipo II. En general, es una función de las distribuciones posibles, a menudo determinada por un parámetro, bajo la hipótesis alternativa. A medida que aumenta la potencia, las posibilidades de que se presente un error del tipo II se reducen (disminución de la tasa de falsos negativos β), de esta manera, la potencia se representa como $1 - \beta$ (sensibilidad). El análisis de la potencia se puede utilizar para calcular el tamaño mínimo de la muestra necesario para que uno pueda detectar razonablemente un efecto de un determinado tamaño, o también para calcular el tamaño del efecto mínimo que es probable que se detecte en un estudio usando un tamaño de muestra dado. Además, el concepto de *alimentación* se utiliza para hacer comparaciones entre diferentes procedimientos de análisis estadísticos (e.g. entre una prueba paramétrica y una no paramétrica de la misma hipótesis. [14](#)

pseudo-F En el análisis multivariante de la varianza con permutaciones (*PERMANOVA*), el estadístico de prueba es una pseudo-ratio F, similar a la relación F en ANOVA. Compara la suma total de diferencias cuadradas (o diferencias de orden) entre objetos pertenecientes a diferentes grupos con la de objetos que pertenecen al mismo grupo. Las F-ratios más grandes indican una separación de grupo más pronunciada, sin embargo, la significación estadística de esta relación suele ser más interesante que su magnitud. [15](#), [29](#)

RNA-seq La secuenciación de ARN, también llamada *Secuenciación del Transcriptoma Entero para Clonación al Azar*, utiliza la secuenciación masiva (NGS) para revelar la presencia y cantidad de ARN, en una muestra biológica en un momento dado. De esta manera, la RNA-seq se usa para analizar cambios en el transcriptoma, concretamente, facilita la observación de transcritos resultantes del empalme alternativo, modificación postranscripcional, fusiones génicas, mutaciones/polimorfismos de nucleótidos únicos y cambios de expresión de genes. Puede ayudar a caracterizar poblaciones diferentes de RNA como miRNA, tRNA, y rRNA, o para determinar las fronteras exón/intrón y verificar o enmendar regiones 5' y 3'. [14](#)

- SNPs** Un polimorfismo puntual, también denominado de un solo nucleótido o SNP (Single Nucleotide Polymorphism, pronunciado snip), es una variación en la secuencia de ADN que afecta a una sola base (adenina (A), timina (T), citosina (C) o guanina (G)) de una secuencia del genoma. Sin embargo, generalmente se considera que cambios de unos pocos nucleótidos, como también pequeñas inserciones y deleciones (indels) pueden ser consideradas como SNP. Una de estas variaciones debe darse al menos en un 1% de la población para ser considerada como un SNP. Si no se llega al 1% no se considera SNP y sí una mutación puntual. En ocasiones estas variaciones de nucleótido único se asocian a otro término conocido como SNV (Single Nucleotide Variant), que a diferencia de los SNPs carece de limitaciones de frecuencia. [14](#)
- sQTLs** Los *Splicing quantitative trait loci* (sQTLs o splicing QTLs) son los loci que regulan el splicing alternativo del ARNm. Se pueden detectar utilizando datos de RNA-seq. Se han desarrollado diversos métodos para descubrir sQTLs, entre los que se incluyen: LeafCutter, Altrans, Cufflinks, y MISO. [14](#)
- sumario estadístico** En estadística descriptivas, el sumario estadístico se utiliza para resumir un conjunto de observaciones, con el fin de comunicar la mayor cantidad de información lo más sencillamente posible, tratando de describir las observaciones según diferentes estadísticos. Comunmente, suelen ofrecerse sumarios de 5 o 7 estadísticos representativos, acompañados de su pertinente diagrama de cajas. Si se está midiendo solo una variable, resultarán de utilidad: la tendencia central mediante la media aritmética, la dispersión estadística gracias a la media de la desviación estándar absoluta, una medida de la forma de la distribución a través de la desviación o curtosis. Si por el contrario se mide más de una variable, el coeficiente de correlación permitirá medir de dependencia estadística entre las variables seleccionadas. [28](#)
- simplex** En geometría, un *simplex* o *n-simplex* es el análogo en n-dimensiones de un triángulo. Más exactamente, es la envoltura convexa de un conjunto de $(n + 1)$ puntos independientes afines en un espacio euclídeo de dimensión n o mayor, es decir, el conjunto de puntos tal que ningún m -plano contiene más que $(m + 1)$ de ellos (*eg.* un 0-simplex es un punto; un 1-simplex un segmento de una línea; un 2-simplex un triángulo; un 3-simplex es un tetraedro; y un 4-simplex es un pentácoron). [50](#)
- tablas de distribución de frecuencia** La frecuencia es la cantidad de veces que se produce un número. La frecuencia de una observación en estadística nos indica el número de veces que la observación ocurre en los datos. [27](#)
- trait** En el ámbito de la genética, un *trait* o *rasgo* es una característica específica de un individuo, los cuales pueden ser determinados por genes, factores ambientales o por una combinación de ambos. Se clasifican como cualitativos (e.g. el color de los ojos) o cuantitativos (e.g. la altura o la presión sanguínea). Cada uno de ellos forma parte del fenotipo general de un individuo. [14](#)
- valores p** En estadística general y contrastes de hipótesis, los valores p (p, p-valor, valor de p consignado, o p-value) se define como la probabilidad de que un valor estadístico calculado sea posible dada una hipótesis nula cierta. Ayuda a diferenciar resultados que son producto del azar del muestreo, de resultados que son estadísticamente significativos. Alternativamente, se define como la probabilidad de observar los resultados del estudio, u otros más alejados de la hipótesis nula, si la hipótesis nula fuera cierta, de manera que si este cumple con la condición de ser menor que un nivel de significancia impuesto arbitrariamente, este se considera como un resultado estadísticamente significativo y, por lo tanto, permite rechazar la hipótesis nula. [16](#), [29](#)

Referencias

- [1] Aitor Invernón de Campos, “Código TFM,” Jan. 2024, original-date: 2024-01-06T10:59:06Z. [Online]. Available: <https://github.com/aitorinvernondecampos/TFMgit>
- [2] M. A. R. Ferreira and S. M. Purcell, “A multivariate test of association,” *Bioinformatics (Oxford, England)*, vol. 25, no. 1, pp. 132–133, Jan. 2009.
- [3] L. Coin, P. O’Reilly, Y. Pompyen, and C. H. a. F. Calboli, “MultiPhen: A Package to Test for Multi-Trait Association,” Feb. 2020. [Online]. Available: <https://cran.r-project.org/web/packages/MultiPhen/index.html>
- [4] X. Zhou and M. Stephens, “Efficient multivariate linear mixed model algorithms for genome-wide association studies,” *Nature Methods*, vol. 11, no. 4, pp. 407–409, Apr. 2014, number: 4 Publisher: Nature Publishing Group. [Online]. Available: <https://www.nature.com/articles/nmeth.2848>
- [5] L. Luo, J. Shen, H. Zhang, A. Chhibber, D. V. Mehrotra, and Z.-Z. Tang, “Multi-trait analysis of rare-variant association summary statistics using MTAR,” *Nature Communications*, vol. 11, no. 1, p. 2850, Jun. 2020, number: 1 Publisher: Nature Publishing Group. [Online]. Available: <https://www.nature.com/articles/s41467-020-16591-0>
- [6] “precimed/mostest,” Jan. 2023, original-date: 2019-06-27T12:33:49Z. [Online]. Available: <https://github.com/precimed/mostest>
- [7] D. van der Meer, O. Frei, T. Kaufmann, A. A. Shadrin, A. Devor, O. B. Smeland, W. K. Thompson, C. C. Fan, D. Holland, L. T. Westlye, O. A. Andreassen, and A. M. Dale, “Understanding the genetic determinants of the brain with MOSTest,” *Nature Communications*, vol. 11, no. 1, p. 3512, Jul. 2020, number: 1 Publisher: Nature Publishing Group. [Online]. Available: <https://www.nature.com/articles/s41467-020-17368-1>
- [8] M. J. Anderson, “A new method for non-parametric multivariate analysis of variance,” *Austral Ecology*, vol. 26, no. 1, pp. 32–46, 2001, eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1442-9993.2001.01070.pp.x>. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1442-9993.2001.01070.pp.x>
- [9] D. Garrido-Martín, M. Calvo, F. Reverter, and R. Guigó, “A fast non-parametric test of association for multiple traits,” *Bioinformatics*, preprint, Jun. 2022. [Online]. Available: <http://biorxiv.org/lookup/doi/10.1101/2022.06.06.493041>
- [10] D. Garrido-Martín, B. Borsari, M. Calvo, F. Reverter, and R. Guigó, “Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome,” *Nature Communications*, vol. 12, no. 1, p. 727, Feb. 2021, number: 1 Publisher: Nature Publishing Group. [Online]. Available: <https://www.nature.com/articles/s41467-020-20578-2>
- [11] J. Monlong, M. Calvo, P. G. Ferreira, and R. Guigó, “Identification of genetic variants associated with alternative splicing using sQTLseeker,” *Nature Communications*, vol. 5, no. 1, p. 4698, Aug. 2014, number: 1 Publisher: Nature Publishing Group. [Online]. Available: <https://www.nature.com/articles/ncomms5698>
- [12] D. Garrido-Martín, “MANTA,” Feb. 2023, original-date: 2019-02-27T12:09:53Z. [Online]. Available: <https://github.com/dgarrimar/manta>
- [13] —, “manta-sim (sim),” Dec. 2022, original-date: 2022-02-16T12:40:23Z. [Online]. Available: <https://github.com/dgarrimar/manta-sim>
- [14] R. B. DAVIES, “Numerical inversion of a characteristic function,” *Biometrika*, vol. 60, no. 2, pp. 415–417, Aug. 1973. [Online]. Available: <https://doi.org/10.1093/biomet/60.2.415>
- [15] R. B. Davies, “Algorithm AS 155: The Distribution of a Linear Combination of Chi2 Random Variables,” *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 29, no. 3, pp. 323–333, 1980, publisher: [Wiley, Royal Statistical Society]. [Online]. Available: <https://www.jstor.org/stable/2346911>

- [16] T. Qi, Y. Wu, H. Fang, F. Zhang, S. Liu, J. Zeng, and J. Yang, “Genetic control of RNA splicing and its distinct role in complex trait variation,” *Nature Genetics*, vol. 54, no. 9, pp. 1355–1363, Sep. 2022, number: 9 Publisher: Nature Publishing Group. [Online]. Available: <https://www.nature.com/articles/s41588-022-01154-4>
- [17] S. Behjati and P. S. Tarpey, “What is next generation sequencing?” *Archives of Disease in Childhood. Education and Practice Edition*, vol. 98, no. 6, pp. 236–238, Dec. 2013. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3841808/>
- [18] “Next-Generation Sequencing (NGS) | Explore the technology.” [Online]. Available: <https://www.illumina.com/science/technology/next-generation-sequencing.html#>
- [19] Z. Wang, M. Gerstein, and M. Snyder, “RNA-Seq: a revolutionary tool for transcriptomics,” *Nature reviews. Genetics*, vol. 10, no. 1, pp. 57–63, Jan. 2009. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2949280/>
- [20] “RNA-Seq,” Feb. 2023, page Version ID: 149381500. [Online]. Available: <https://es.wikipedia.org/w/index.php?title=RNA-Seq&oldid=149381500>
- [21] V. Tam, N. Patel, M. Turcotte, Y. Bossé, G. Paré, and D. Meyre, “Benefits and limitations of genome-wide association studies,” *Nature Reviews Genetics*, vol. 20, no. 8, pp. 467–484, Aug. 2019, number: 8 Publisher: Nature Publishing Group. [Online]. Available: <https://www.nature.com/articles/s41576-019-0127-1>
- [22] D. Garrido Martín, “A Multivariate approach to study the genetic determinants of phenotypic traits,” Ph.D. Thesis, Universitat Pompeu Fabra, Jan. 2020, accepted: 2020-02-05T17:05:52Z Publication Title: TDX (Tesis Doctorals en Xarxa). [Online]. Available: <https://www.tdx.cat/handle/10803/668497>
- [23] S. Turner, L. L. Armstrong, Y. Bradford, C. S. Carlson, D. C. Crawford, A. T. Crenshaw, M. de Andrade, K. F. Doheny, J. L. Haines, G. Hayes, G. Jarvik, L. Jiang, I. J. Kullo, R. Li, H. Ling, T. A. Manolio, M. Matsumoto, C. A. McCarty, A. N. McDavid, D. B. Mirel, J. E. Paschall, E. W. Pugh, L. V. Rasmussen, R. A. Wilke, R. L. Zuvich, and M. D. Ritchie, “Quality Control Procedures for Genome Wide Association Studies,” *Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.]*, vol. CHAPTER, p. Unit1.19, Jan. 2011. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3066182/>
- [24] “Genome-wide association study,” Apr. 2023, page Version ID: 1151122410. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Genome-wide_association_study&oldid=1151122410
- [25] “Splicing quantitative trait loci,” Mar. 2021, page Version ID: 1015148167. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Splicing_quantitative_trait_loci&oldid=1015148167
- [26] B. Everitt, *The Cambridge dictionary of statistics*. Cambridge, UK ; New York: Cambridge University Press, 1998.
- [27] “Univariate (statistics) - Wikipedia.” [Online]. Available: [https://en.wikipedia.org/wiki/Univariate_\(statistics\)#Analysis](https://en.wikipedia.org/wiki/Univariate_(statistics)#Analysis)
- [28] “Bivariate analysis - Wikipedia.” [Online]. Available: https://en.wikipedia.org/wiki/Bivariate_analysis
- [29] “R: The R Stats Package.” [Online]. Available: <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/stats-package.html>
- [30] “Simplex - Wikipedia, la enciclopedia libre.” [Online]. Available: <https://es.wikipedia.org/wiki/S%C3%ADmplex>
- [31] “Algoritmo simplex - Wikipedia, la enciclopedia libre.” [Online]. Available: https://es.wikipedia.org/wiki/Algoritmo_s%C3%ADmplex

Apéndice A

Anexo de tablas

Tabla A.1: Simulaciones comparativas **MANTA-MANOVA** bajo el modelo de distribución *mvnorm* (*Objetivo I*), calculando la potencia estadística \mathbb{P} bajo un nivel de significación $\alpha = 0.05$ y con: $S = 1000$; $n = 300$; $q = 3$

Variable de simulación	Nombre	Valores
Nivel de significación	alpha	0.05
Número de simulaciones	S	1000
Tamaño de la muestra	n	300
Número de respuestas	q	3
Tipo de varianza	Var	Equal o Unequal (Type I, Type II, Type III)
Parámetro de generación de H_1	delta	De 0 a 0.35 (21 valores) Con un paso de 0.0175
Correlación de las variables	Cor	0, 0.2, 0.4, 0.6, 0.8

(a) Combinaciones $\Delta - Var$: imponiendo una matriz de correlación *homogénea*, con valores idénticos fuera de la diagonal, determinados por el valor de la variable **Cor**.

Variable de simulación	Nombre	Valores
Nivel de significación	alpha	0.05
Número de simulaciones	S	1000
Tamaño de la muestra	n	300
Número de respuestas	q	3
Tipo de varianza	Var	Equal o Unequal (Type I, Type II, Type III)
Parámetro de generación de H_1	delta	De 0 a 0.35 (21 valores) Con un paso de 0.0175
Correlación de las variables	Cor	Valores aleatorios

(b) Combinaciones $\Delta - Var$: imponiendo una matriz de correlación *inhomogénea*, con valores aleatorios fuera de la diagonal (opción implementada en la función *sim.mvnorm()*).

Tabla A.2: Simulaciones comparativas **MANTA-MANOVA** bajo el modelo de distribución *mvnorm* (*Objetivo I*), calculando la potencia estadística \mathbb{P} bajo unos niveles de significación estadísticos menores ($\alpha \in [0.01, 0.001]$) y con: $S = 1000$; $n = 300$; $q = 3$

Variable de simulación	Nombre	Valores
Nivel de significación	alpha	0.01, 0.001
Número de simulaciones	S	1000
Tamaño de la muestra	n	300
Número de respuestas	q	3
Tipo de varianza	Var	Equal o Unequal (Type I, Type II, Type III)
Parámetro de generación de H_1	delta	De 0 a 0.35 (21 valores) Con un paso de 0.0175
Correlación de las variables	Cor	0, 0.2, 0.4, 0.6, 0.8

(a) Combinaciones $\Delta - Var$: imponiendo una matriz de correlación *homogénea*, con valores idénticos fuera de la diagonal, determinados por el valor de la variable **Cor**.

Variable de simulación	Nombre	Valores
Nivel de significación	alpha	0.01, 0.001
Número de simulaciones	S	1000
Tamaño de la muestra	n	300
Número de respuestas	q	3
Tipo de varianza	Var	Equal o Unequal (Type I, Type II, Type III)
Parámetro de generación de H_1	delta	De 0 a 0.35 (21 valores) Con un paso de 0.0175
Correlación de las variables	Cor	Valores aleatorios

(b) Combinaciones $\Delta - Var$: imponiendo una matriz de correlación *inhomogénea*, con valores aleatorios fuera de la diagonal (opción implementada en la función *sim.mvnorm()*).

Tabla A.3: Muestra aleatoria de 15 de los 5040 resultados obtenidos para los cálculos de la potencia estadística \mathbb{P} bajo las condiciones detalladas en las simulaciones A.1a y A.2a.

Datos	Modelo	alpha	S	n	Var	q	Cor	delta	Método	Potencia	t comp.(s)
Datos sin transformar	mvnorm	0.050	1,000	300	Unequal Type III	3	0.4	0.158	MANOVA	0.979	0.720
Transformación logarítmica	mvnorm	0.001	1,000	300	Equal	3	0.6	0.262	MANTA	0.241	1.390
Transformación Centered Log Ratio (clr)	mvnorm	0.050	1,000	300	Unequal Type II	3	0.4	0.315	MANTA	0.092	1.530
Datos sin transformar	mvnorm	0.050	1,000	300	Unequal Type II	3	0.2	0.088	MANOVA	0.219	0.690
Datos sin transformar	mvnorm	0.010	1,000	300	Unequal Type III	3	0.2	0.210	MANOVA	0.500	0.660
Transformación Centered Log Ratio (clr)	mvnorm	0.050	1,000	300	Unequal Type II	3	0.8	0.140	MANTA	0.051	1.070
Datos sin transformar	mvnorm	0.050	1,000	300	Unequal Type I	3	0.0	0.018	MANOVA	0.085	0.690
Transformación logarítmica	mvnorm	0.001	1,000	300	Unequal Type I	3	0.8	0.018	MANOVA	0.008	0.650
Transformación Centered Log Ratio (clr)	mvnorm	0.010	1,000	300	Unequal Type I	3	0.2	0	MANOVA	0.948	0.730
Transformación Centered Log Ratio (clr)	mvnorm	0.010	1,000	300	Unequal Type II	3	0.8	0.315	MANTA	0.947	1.490
Datos sin transformar	mvnorm	0.010	1,000	300	Unequal Type III	3	0.8	0.262	MANTA	0.482	1.110
Transformación Centered Log Ratio (clr)	mvnorm	0.010	1,000	300	Equal	3	0.0	0.350	MANOVA	0.995	0.710
Transformación Centered Log Ratio (clr)	mvnorm	0.050	1,000	300	Equal	3	0.8	0.175	MANOVA	0.059	0.750
Transformación Centered Log Ratio (clr)	mvnorm	0.010	1,000	300	Unequal Type III	3	0.8	0.332	MANOVA	0.996	0.770
Datos sin transformar	mvnorm	0.010	1,000	300	Unequal Type II	3	0.8	0.350	MANTA	0.828	1.110

Tabla A.4: Muestra aleatoria de 15 de los 1008 resultados obtenidos para los cálculos de la potencia estadística \mathbb{P} bajo las condiciones detalladas en las simulaciones A.1b y A.2b.

Datos	Modelo	alpha	S	n	Var	q	Cor	delta	Método	Potencia	t comp.(s)
Datos sin transformar	mvnorm	0.010	1,000	300	Unequal Type II	3	Aleat.	0.280	MANTA	0.560	1.280
Transformación Centered Log Ratio (clr)	mvnorm	0.010	1,000	300	Unequal Type I	3	Aleat.	0.018	MANOVA	0.997	0.810
Transformación Centered Log Ratio (clr)	mvnorm	0.050	1,000	300	Unequal Type II	3	Aleat.	0.280	MANOVA	0.999	0.770
Transformación raíz cuadrada	mvnorm	0.001	1,000	300	Unequal Type I	3	Aleat.	0.192	MANOVA	0.255	0.680
Transformación logarítmica	mvnorm	0.001	1,000	300	Unequal Type I	3	Aleat.	0.192	MANTA	0.086	1.140
Transformación logarítmica	mvnorm	0.010	1,000	300	Unequal Type II	3	Aleat.	0.070	MANOVA	0.104	0.690
Datos sin transformar	mvnorm	0.010	1,000	300	Unequal Type I	3	Aleat.	0.158	MANOVA	0.271	0.730
Transformación raíz cuadrada	mvnorm	0.001	1,000	300	Unequal Type I	3	Aleat.	0.088	MANTA	0.008	1.030
Transformación raíz cuadrada	mvnorm	0.001	1,000	300	Equal	3	Aleat.	0.245	MANTA	0.167	1.030
Datos sin transformar	mvnorm	0.001	1,000	300	Unequal Type III	3	Aleat.	0.158	MANTA	0.019	1.210
Transformación logarítmica	mvnorm	0.001	1,000	300	Unequal Type I	3	Aleat.	0.332	MANOVA	0.760	0.820
Transformación raíz cuadrada	mvnorm	0.001	1,000	300	Unequal Type II	3	Aleat.	0.175	MANOVA	0.191	0.770
Datos sin transformar	mvnorm	0.010	1,000	300	Equal	3	Aleat.	0.035	MANTA	0.022	1.030
Transformación logarítmica	mvnorm	0.010	1,000	300	Unequal Type III	3	Aleat.	0.332	MANOVA	0.892	0.730
Transformación Centered Log Ratio (clr)	mvnorm	0.001	1,000	300	Unequal Type III	3	Aleat.	0.018	MANOVA	0.986	0.770

Tabla A.5: Descripción de los estadísticos potencia (\mathbb{P}) y tiempo de computación ($t_{comp.}$), para el modelo **MANTA**, bajo una distribución *mvnorm*, con una matriz de correlación *homogénea*, y considerando diferentes niveles de significación.

Tipo de Datos	Statistic	N	Mean	St. Dev.	Min	Max
Datos sin transformar	Potencia	210	0.6686	0.3754	0.0410	1.0000
	tcomp	210	1.0975	0.0715	1.0157	1.5871
Transformación log-ratio	Potencia	210	0.6672	0.3758	0.0420	1.0000
	tcomp	210	1.1054	0.1084	1.0082	1.7344
Transformación centered log-ratio	Potencia	210	0.0609	0.0197	0.0240	0.1210
	tcomp	210	1.1252	0.0824	1.0432	1.6845
Transformación raíz cuadrada	Potencia	210	0.6681	0.3756	0.0420	1.0000
	tcomp	210	1.0901	0.0985	1.0057	1.8174

(a) Nivel de significación $\alpha = 0.05$.

Tipo de Datos	Statistic	N	Mean	St. Dev.	Min	Max
Datos sin transformar	Potencia	210	0.2722	0.2795	0.0140	0.8280
	tcomp	210	1.3319	1.5585	1.0310	17.5338
Transformación log-ratio	Potencia	210	0.3068	0.2723	0.0300	0.8390
	tcomp	210	1.4757	3.8801	1.0055	55.2734
Transformación centered log-ratio	Potencia	210	0.9411	0.0317	0.8000	0.9520
	tcomp	210	1.6501	1.0700	1.4024	16.8122
Transformación raíz cuadrada	Potencia	210	0.2830	0.2786	0.0190	0.8380
	tcomp	210	1.1436	0.2601	1.0215	4.2436

(b) Nivel de significación $\alpha = 0.01$.

Tipo de Datos	Statistic	N	Mean	St. Dev.	Min	Max
Datos sin transformar	Potencia	210	0.1373	0.1845	0.0030	0.5840
	tcomp	210	1.1096	0.0567	1.0255	1.3353
Transformación log-ratio	Potencia	210	0.1556	0.1818	0.0080	0.5980
	tcomp	210	1.1080	0.0662	1.0232	1.4161
Transformación centered log-ratio	Potencia	210	0.8598	0.0411	0.6770	0.8730
	tcomp	210	1.5201	0.1165	1.4057	2.1641
Transformación raíz cuadrada	Potencia	210	0.1417	0.1821	0.0050	0.5980
	tcomp	210	1.0994	0.0552	1.0105	1.2962

(c) Nivel de significación $\alpha = 0.001$.

Tabla A.6: Descripción de los estadísticos potencia (\mathbb{P}) y tiempo de computación ($t_{comp.}$), para el modelo **MANOVA**, bajo una distribución *mvnorm*, con una matriz de correlación *homogénea*, y considerando diferentes niveles de significación.

Tipo de Datos	Statistic	N	Mean	St. Dev.	Min	Max
Datos sin transformar	Potencia	210	0.6823	0.3762	0.0520	1.0000
	tcomp	210	0.6850	0.0557	0.6396	1.3357
Transformación log-ratio	Potencia	210	0.6816	0.3764	0.0450	1.0000
	tcomp	210	0.7040	0.0880	0.6375	1.3116
Transformación centered log-ratio	Potencia	210	0.0539	0.0123	0.0340	0.0880
	tcomp	210	0.7609	0.0729	0.6939	1.3823
Transformación raíz cuadrada	Potencia	210	0.6822	0.3762	0.0460	1.0000
	tcomp	210	0.6811	0.0413	0.6280	0.8964

(a) Nivel de significación $\alpha = 0.05$.

Tipo de Datos	Statistic	N	Mean	St. Dev.	Min	Max
Datos sin transformar	Potencia	210	0.3979	0.3234	0.0200	0.9000
	tcomp	210	0.8146	1.3082	0.6232	17.4757
Transformación log-ratio	Potencia	210	0.9944	0.0104	0.9480	0.9970
	tcomp	210	0.8063	0.5809	0.6884	9.0555
Transformación centered log-ratio	Potencia	210	0.9411	0.0317	0.8000	0.9520
	tcomp	210	1.6501	1.0700	1.4024	16.8122
Transformación raíz cuadrada	Potencia	210	0.4068	0.3192	0.0300	0.9120
	tcomp	210	0.9175	2.4927	0.6267	34.4371

(b) Nivel de significación $\alpha = 0.01$.

Tipo de Datos	Statistic	N	Mean	St. Dev.	Min	Max
Datos sin transformar	Potencia	210	0.2820	0.2806	0.0020	0.7890
	tcomp	210	0.6706	0.0342	0.6323	0.7993
Transformación log-ratio	Potencia	210	0.9811	0.0182	0.9000	0.9870
	tcomp	210	0.7444	0.0501	0.6965	0.9458
Transformación centered log-ratio	Potencia	210	0.8598	0.0411	0.6770	0.8730
	tcomp	210	1.5201	0.1165	1.4057	2.1641
Transformación raíz cuadrada	Potencia	210	0.2864	0.2763	0.0050	0.7920
	tcomp	210	0.6728	0.0418	0.6316	0.8797

(c) Nivel de significación $\alpha = 0.001$.

Tabla A.7: Descripción de los estadísticos potencia (\mathbb{P}) y tiempo de computación ($t_{comp.}$), para el modelo **MANTA**, bajo una distribución *mvnorm*, con una matriz de correlación *inhomogénea*, y considerando diferentes niveles de significación.

Tipo de Datos	Statistic	N	Mean	St. Dev.	Min	Max
Datos sin transformar	Potencia	42	0.4257	0.3292	0.0620	0.9550
	tcomp	42	1.1242	0.1114	1.0503	1.6741
Transformación log-ratio	Potencia	42	0.4719	0.3041	0.1010	0.9470
	tcomp	42	1.1114	0.0686	1.0414	1.3262
Transformación centered log-ratio	Potencia	42	0.9757	0.0201	0.8870	0.9810
	tcomp	42	1.5270	0.1091	1.4347	1.9019
Transformación raíz cuadrada	Potencia	42	0.4393	0.3204	0.0750	0.9460
	tcomp	42	1.0911	0.0686	1.0050	1.3543

(a) Nivel de significación $\alpha = 0.05$.

Tipo de Datos	Statistic	N	Mean	St. Dev.	Min	Max
Datos sin transformar	Potencia	42	0.2722	0.2822	0.0140	0.8280
	tcomp	42	1.1125	0.0759	1.0149	1.2862
Transformación log-ratio	Potencia	42	0.3068	0.2749	0.0300	0.8390
	tcomp	42	1.0599	0.0425	0.9808	1.2123
Transformación centered log-ratio	Potencia	42	0.9411	0.0320	0.8000	0.9520
	tcomp	42	1.4969	0.1111	1.3790	1.8822
Transformación raíz cuadrada	Potencia	42	0.2830	0.2813	0.0190	0.8380
	tcomp	42	1.0752	0.0614	0.9945	1.2817

(b) Nivel de significación $\alpha = 0.01$.

Tipo de Datos	Statistic	N	Mean	St. Dev.	Min	Max
Datos sin transformar	Potencia	42	0.1373	0.1862	0.0030	0.5840
	tcomp	42	1.0848	0.0625	1.0200	1.3337
Transformación log-ratio	Potencia	42	0.1556	0.1836	0.0080	0.5980
	tcomp	42	1.0903	0.0629	1.0085	1.2468
Transformación centered log-ratio	Potencia	42	0.8598	0.0415	0.6770	0.8730
	tcomp	42	1.5001	0.1081	1.4033	1.8978
Transformación raíz cuadrada	Potencia	42	0.1417	0.1838	0.0050	0.5980
	tcomp	42	1.0763	0.0616	1.0124	1.2847

(c) Nivel de significación $\alpha = 0.001$.

Tabla A.8: Descripción de los estadísticos potencia (\mathbb{P}) y tiempo de computación (t_{comp}), para el modelo **MANOVA**, bajo una distribución *mvnorm*, con una matriz de correlación *inhomogénea*, y considerando diferentes niveles de significación.

Tipo de Datos	Statistic	N	Mean	St. Dev.	Min	Max
Datos sin transformar	Potencia	42	0.5154	0.3346	0.0620	0.9600
	tcomp	42	0.7004	0.0560	0.6452	0.9949
Transformación log-ratio	Potencia	42	0.5664	0.3038	0.1340	0.9660
	tcomp	42	0.6911	0.0434	0.6504	0.9253
Transformación centered log-ratio	Potencia	42	0.9975	0.0069	0.9670	0.9990
	tcomp	42	0.7868	0.0633	0.7242	1.0152
Transformación raíz cuadrada	Potencia	42	0.5318	0.3252	0.0860	0.9640
	tcomp	42	0.6880	0.0217	0.6524	0.7515

(a) Nivel de significación $\alpha = 0.05$.

Tipo de Datos	Statistic	N	Mean	St. Dev.	Min	Max
Datos sin transformar	Potencia	42	0.3979	0.3265	0.0200	0.9000
	tcomp	42	0.6956	0.0416	0.6473	0.8252
Transformación log-ratio	Potencia	42	0.4318	0.3126	0.0540	0.9160
	tcomp	42	0.6869	0.0395	0.6418	0.8519
Transformación centered log-ratio	Potencia	42	0.9944	0.0105	0.9480	0.9970
	tcomp	42	0.7710	0.0564	0.7009	0.9673
Transformación raíz cuadrada	Potencia	42	0.4068	0.3223	0.0300	0.9120
	tcomp	42	0.6879	0.0468	0.6465	0.8527

(b) Nivel de significación $\alpha = 0.01$.

Tipo de Datos	Statistic	N	Mean	St. Dev.	Min	Max
Datos sin transformar	Potencia	42	0.2820	0.2833	0.0020	0.7890
	tcomp	42	0.6897	0.0344	0.6532	0.8221
Transformación log-ratio	Potencia	42	0.3010	0.2769	0.0080	0.8090
	tcomp	42	0.6916	0.0373	0.6473	0.8217
Transformación centered log-ratio	Potencia	42	0.9811	0.0184	0.9000	0.9870
	tcomp	42	0.7694	0.0441	0.7214	0.9227
Transformación raíz cuadrada	Potencia	42	0.2864	0.2790	0.0050	0.7920
	tcomp	42	0.6990	0.0525	0.6434	0.8634

(c) Nivel de significación $\alpha = 0.001$.

Tabla A.9: Descripción de los estadísticos potencia (\mathbb{P}) y tiempo de computación (t_{comp}), para los métodos **MANTA** y **MANOVA**, bajo una distribución *mvnorm*, con una matriz de correlación *homogénea*, y considerando diferentes niveles de significación.

Tipo de Datos	α	Statistic	N	Mean	St. Dev.	Min	Max
Datos sin transformar	0.05	Potencia	210	0.6686	0.3754	0.0410	1.0000
		tcomp	210	1.0975	0.0715	1.0157	1.5871
	0.01	Potencia	210	0.2722	0.2795	0.0140	0.8280
		tcomp	210	1.3319	1.5585	1.0310	17.5338
	0.001	Potencia	210	0.1373	0.1845	0.0030	0.5840
		tcomp	210	1.1096	0.0567	1.0255	1.3353

(a) Método **MANTA**.

Tipo de Datos	α	Statistic	N	Mean	St. Dev.	Min	Max
Datos sin transformar	0.05	Potencia	210	0.6823	0.3762	0.0520	1.0000
		tcomp	210	0.6850	0.0557	0.6396	1.3357
	0.01	Potencia	210	0.3979	0.3234	0.0200	0.9000
		tcomp	210	0.8146	1.3082	0.6232	17.4757
	0.001	Potencia	210	0.2820	0.2806	0.0020	0.7890
		tcomp	210	0.6706	0.0342	0.6323	0.7993

(b) Método **MANOVA**.

Tabla A.10: Descripción de los estadísticos potencia (\mathbb{P}) y tiempo de computación (t_{comp}), para los métodos **MANTA** y **MANOVA**, bajo una distribución *mvnorm*, con una matriz de correlación *inhomogénea*, y considerando diferentes niveles de significación.

Tipo de Datos	α	Statistic	N	Mean	St. Dev.	Min	Max
Datos sin transformar	0.05	Potencia	42	0.4257	0.3292	0.0620	0.9550
		tcomp	42	1.1242	0.1114	1.0503	1.6741
	0.01	Potencia	42	0.2722	0.2822	0.0140	0.8280
		tcomp	42	1.1125	0.0759	1.0149	1.2862
	0.001	Potencia	42	0.1373	0.1862	0.0030	0.5840
		tcomp	42	1.0848	0.0625	1.0200	1.3337

(a) Método **MANTA**.

Tipo de Datos	α	Statistic	N	Mean	St. Dev.	Min	Max
Datos sin transformar	0.05	Potencia	42	0.5154	0.3346	0.0620	0.9600
		tcomp	42	0.7004	0.0560	0.6452	0.9949
	0.01	Potencia	42	0.3979	0.3265	0.0200	0.9000
		tcomp	42	0.6956	0.0416	0.6473	0.8252
	0.001	Potencia	42	0.2820	0.2833	0.0020	0.7890
		tcomp	42	0.6897	0.0344	0.6532	0.8221

(b) Método **MANOVA**.

Tabla A.11: Simulaciones para el estudio de la posible invarianza frente a la transformación de los datos del método asintótico *PERMANOVA*, implementado en *MANTA*, con respecto a su potencia estadística (\mathbb{P}). Teniendo en cuenta diferentes situaciones de simulación del conjunto de datos, mediante el uso de un *algoritmo simplex* con $n = 3$.

Variable de simulación	Nombre	Valores	Variable de simulación	Nombre	Valores
Nivel de significación	alpha	0.05	Nivel de significación	alpha	0.01, 0.001
Número de simulaciones	S	1000	Número de simulaciones	S	1000
Tamaño de la muestra	n	300	Tamaño de la muestra	n	300
Número de respuestas	q	3, 5	Número de respuestas	q	3, 5
Localización del modelo que genera el 3-símplex	loc	1, 2, 3, 5 ($\forall q = 3$) 1, 2, 3 ($\forall q = 5$)	Localización del modelo que genera el 3-símplex	loc	1, 2, 3, 5 ($\forall q = 3$) 1, 2, 3 ($\forall q = 5$)
Parámetro de generación de H_1	delta	De 0 a 0.025 (26 valores) Con un paso variable	Parámetro de generación de H_1	delta	De 0 a 0.025 (26 valores) Con un paso variable
(a) Combinaciones $\Delta - q - loc$: imponiendo un $\alpha \in 0.05$.			(b) Combinaciones $\Delta - q - loc$: imponiendo un $\alpha \in [0.01, 0.001]$.		

Tabla A.12: Muestra aleatoria de 15 de los 2144 resultados obtenidos para los cálculos de la potencia estadística \mathbb{P} bajo las condiciones detalladas en las simulaciones A.11a y A.11b.

Datos	Modelo	alpha	S	n	Var	q	Cor	delta	Método	Potencia	t comp.(s)
Transformación raíz cuadrada	simplex	0.0500	1,000	300	5	3	0.0260	0.0250	MANTA	1	13.1700
Transformación raíz cuadrada	simplex	0.0100	1,000	300	5	1	0.0270	0.0235	MANTA	1	10.4600
Datos sin transformar	simplex	0.0500	1,000	300	3	5	0.0230	0.0005	MANTA	0.0480	1.1400
Transformación raíz cuadrada	simplex	0.0500	1,000	300	3	3	0.0240	0.0240	MANTA	1	17.3500
Datos sin transformar	simplex	0.0500	1,000	300	5	3	0.0260	0.0090	MANTA	0.7960	1.1000
Transformación logarítmica	simplex	0.0500	1,000	300	5	3	0.0260	0.0170	MANTA	1	1
Transformación Centered Log Ratio (clr)	simplex	0.0010	1,000	300	3	5	0.0230	0.0085	MANTA	0.0900	1.0500
Transformación raíz cuadrada	simplex	0.0100	1,000	300	3	5	0.0230	0.0080	MANTA	0.5310	1.4400
Transformación Centered Log Ratio (clr)	simplex	0.0500	1,000	300	5	2	0.0270	0.0175	MANTA	0.9970	1.0500
Datos sin transformar	simplex	0.0010	1,000	300	3	3	0.0240	0.0140	MANTA	0.9710	1.0900
Transformación logarítmica	simplex	0.0010	1,000	300	3	5	0.0230	0.0045	MANTA	0.0260	0.9900
Datos sin transformar	simplex	0.0010	1,000	300	3	5	0.0230	0.0005	MANTA	0	1.0700
Transformación logarítmica	simplex	0.0100	1,000	300	3	2	0.0240	0.0220	MANTA	1	1
Transformación raíz cuadrada	simplex	0.0500	1,000	300	3	5	0.0230	0.0165	MANTA	1	14.2000
Datos sin transformar	simplex	0.0010	1,000	300	5	1	0.0270	0.0020	MANTA	0.0010	1.1300

Tabla A.13: Sumario estadístico (para diferentes valores del nivel de significación α) de la potencia estadística (\mathbb{P}) calculada, y del tiempo de computación empleado en las simulaciones *3-simplex*, sin aplicar al conjunto de datos ninguna transformación.

Statistic	N	Mean	St. Dev.	Min	Max
Potencia	179	0.7593	0.3475	0.0470	1.0000
tcomp	179	1.1578	0.0798	1.0764	1.4821
(a) Combinaciones $\Delta - q - loc$: imponiendo un $\alpha \in 0.05$.					
Statistic	N	Mean	St. Dev.	Min	Max
Potencia	178	0.6971	0.3906	0.0050	1.0000
tcomp	178	1.1214	0.0620	1.0489	1.3689
(b) Combinaciones $\Delta - q - loc$: imponiendo un $\alpha \in 0.01$.					
Statistic	N	Mean	St. Dev.	Min	Max
Potencia	179	0.6277	0.4212	0.0000	1.0000
tcomp	179	1.1132	0.0546	1.0469	1.3755
(c) Combinaciones $\Delta - q - loc$: imponiendo un $\alpha \in 0.001$.					

Tabla A.14: Sumario estadístico (para diferentes valores del nivel de significación α) de la potencia estadística (\mathbb{P}) calculada, y el tiempo de computación empleado en las simulaciones *3-simplex*, aplicando al conjunto de datos una transformación logarítmica.

Statistic	N	Mean	St. Dev.	Min	Max
Potencia	179	0.7516	0.3514	0.0470	1.0000
tcomp	179	1.0767	0.0917	0.9947	1.6458
(a) Combinaciones $\Delta - q - loc$: imponiendo un $\alpha \in 0.05$.					
Statistic	N	Mean	St. Dev.	Min	Max
Potencia	178	0.6878	0.3942	0.0060	1.0000
tcomp	178	1.0395	0.0462	0.9804	1.2416
(b) Combinaciones $\Delta - q - loc$: imponiendo un $\alpha \in 0.01$.					
Statistic	N	Mean	St. Dev.	Min	Max
Potencia	179	0.6176	0.4236	0.0000	1.0000
tcomp	179	1.0469	0.0558	0.9882	1.3023
(c) Combinaciones $\Delta - q - loc$: imponiendo un $\alpha \in 0.001$.					

Tabla A.15: Sumario estadístico (para diferentes valores del nivel de significación α) de la potencia estadística (\mathbb{P}) calculada, y el tiempo de computación empleado en las simulaciones *3-simplex*, aplicando al conjunto de datos una transformación *Centered Log Ratio* (*clr*).

Statistic	N	Mean	St. Dev.	Min	Max
Potencia	179	0.6964	0.3745	0.0510	1.0000
tcomp	179	1.1206	0.0843	1.0358	1.5834

(a) Combinaciones $\Delta - q - loc$: imponiendo un $\alpha \in 0.05$.

Statistic	N	Mean	St. Dev.	Min	Max
Potencia	178	0.6112	0.4158	0.0050	1.0000
tcomp	178	1.0879	0.0470	1.0276	1.2818

(b) Combinaciones $\Delta - q - loc$: imponiendo un $\alpha \in 0.01$.

Statistic	N	Mean	St. Dev.	Min	Max
Potencia	179	0.5194	0.4351	0.0000	1.0000
tcomp	179	1.0762	0.0375	1.0300	1.2588

(c) Combinaciones $\Delta - q - loc$: imponiendo un $\alpha \in 0.001$.

Tabla A.16: Sumario estadístico (para diferentes valores del nivel de significación α) de la potencia estadística (\mathbb{P}) calculada, y el tiempo de computación empleado en las simulaciones *3-simplex*, aplicando al conjunto de datos una transformación de *raíz cuadrada* (*sqrt*).

Statistic	N	Mean	St. Dev.	Min	Max
Potencia	179	0.7389	0.3574	0.0470	1.0000
tcomp	179	5.2245	12.1286	0.9998	95.8778

(a) Combinaciones $\Delta - q - loc$: imponiendo un $\alpha \in 0.05$.

Statistic	N	Mean	St. Dev.	Min	Max
Potencia	178	0.6708	0.4002	0.0060	1.0000
tcomp	178	5.4130	13.1872	0.9952	101.6103

(b) Combinaciones $\Delta - q - loc$: imponiendo un $\alpha \in 0.01$.

Statistic	N	Mean	St. Dev.	Min	Max
Potencia	179	0.5969	0.4278	0.0000	1.0000
tcomp	179	5.2086	12.1741	0.9893	96.1683

(c) Combinaciones $\Delta - q - loc$: imponiendo un $\alpha \in 0.001$.

Apéndice B

Anexo de figuras

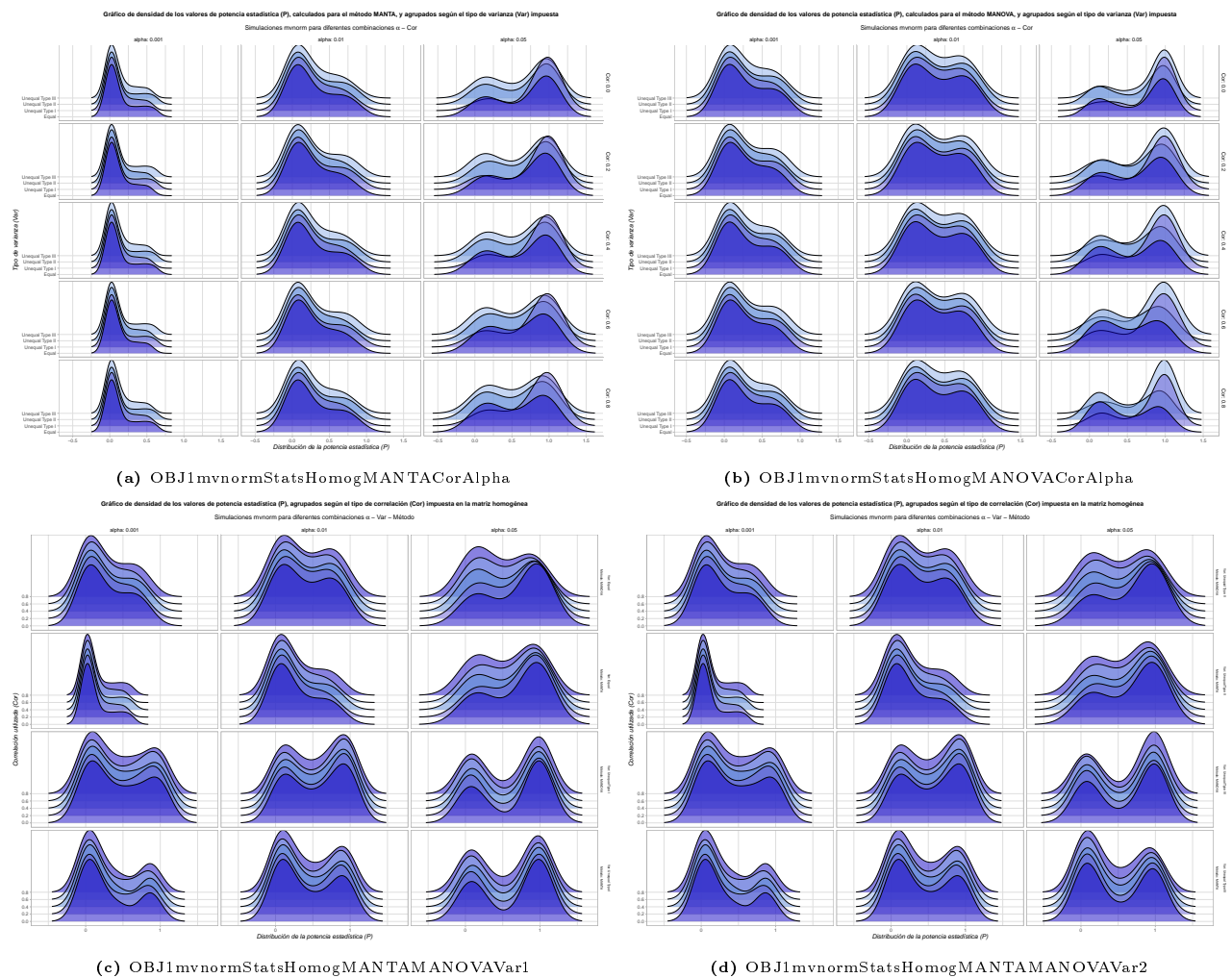


Figura B.1: OBJ1mvnormStatsHomog

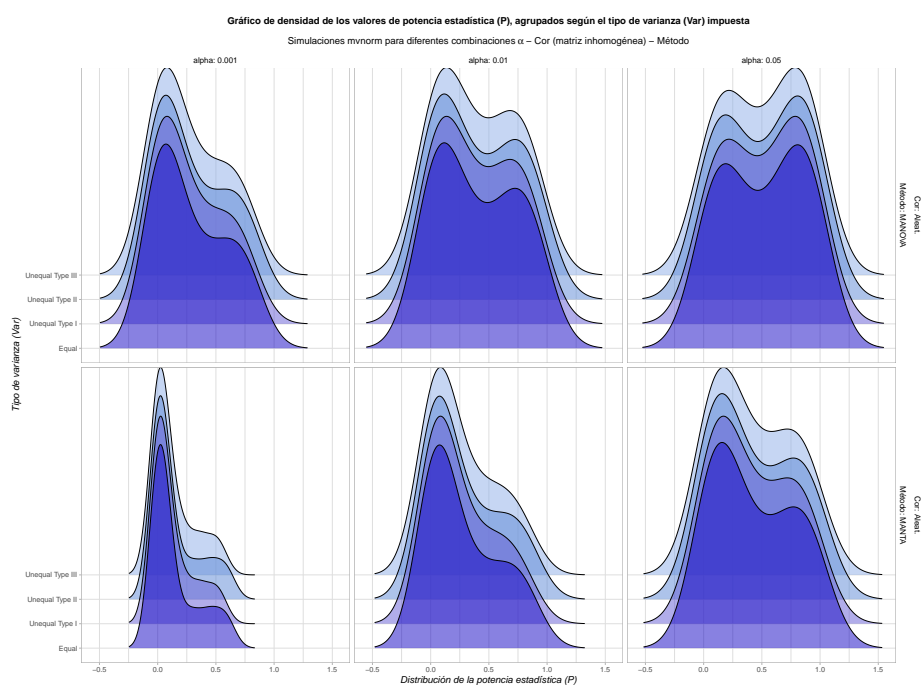


Figura B.2: A.

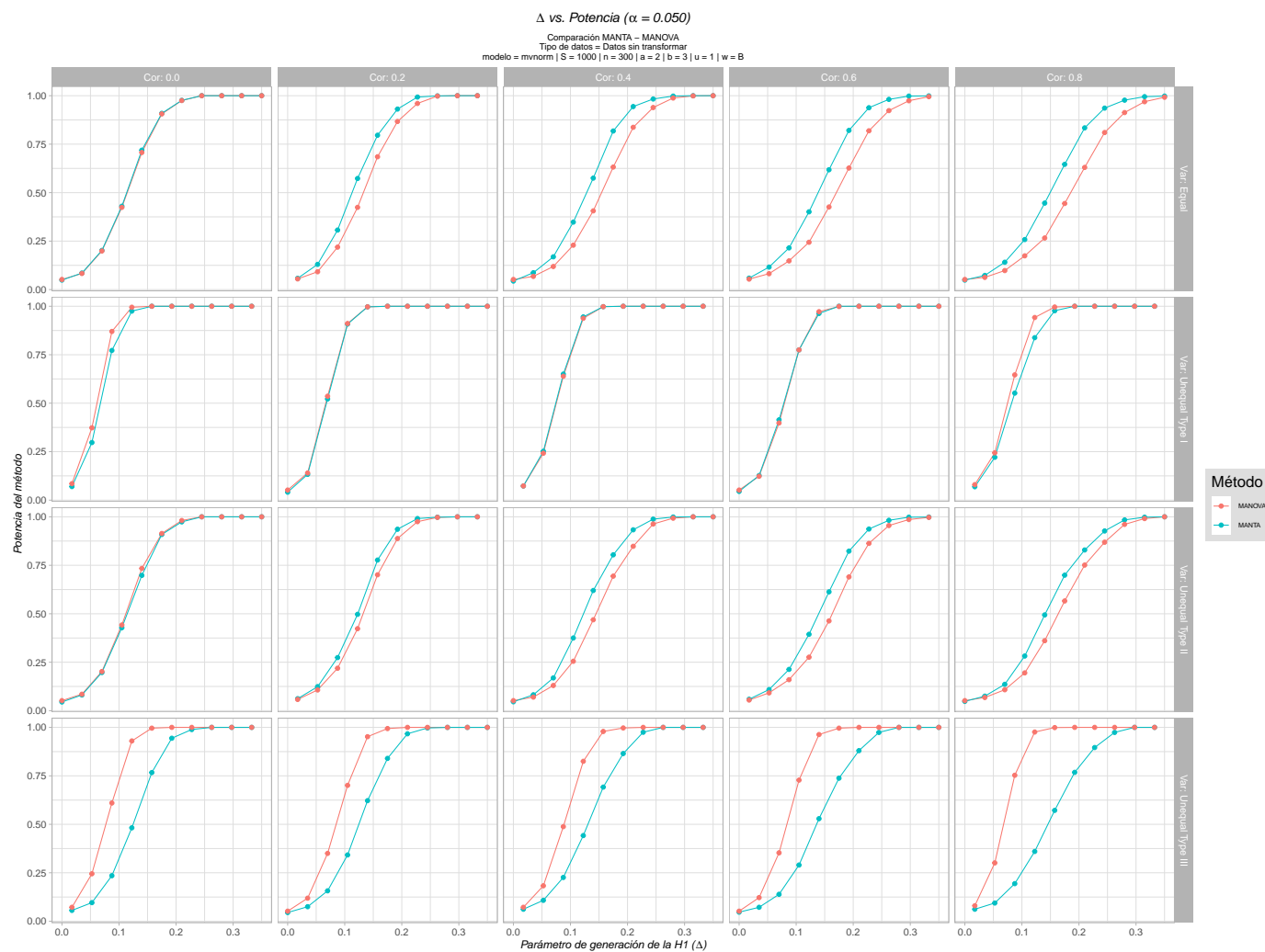


Figura B.3: A.

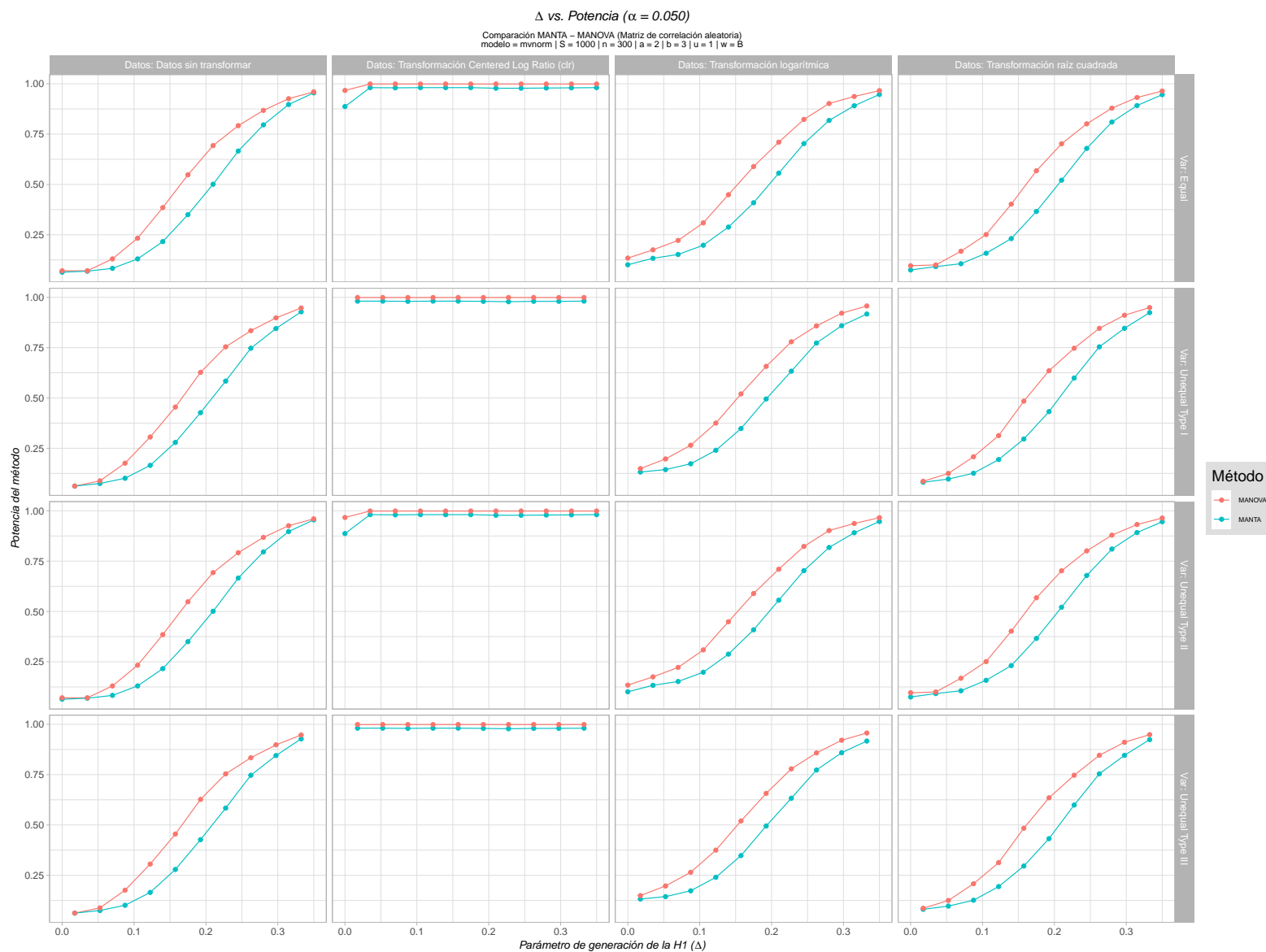
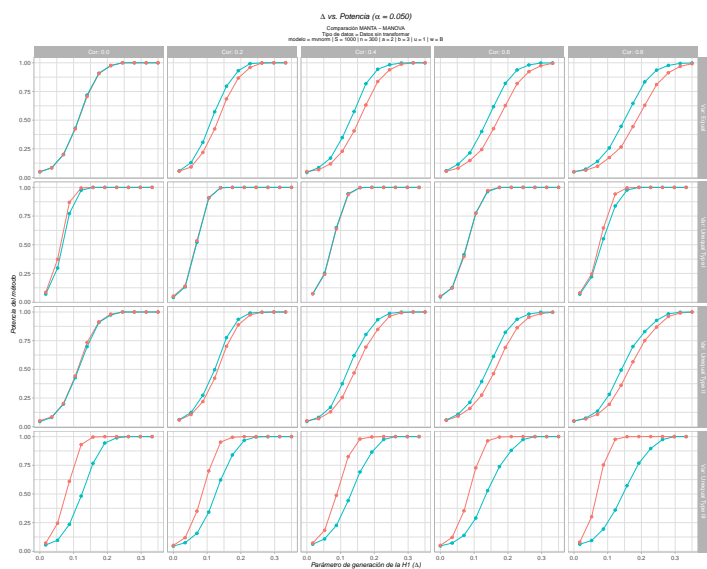
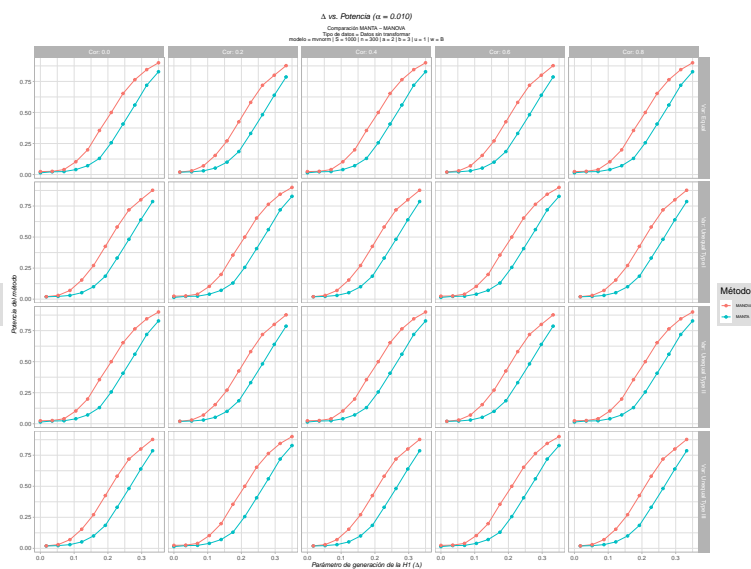


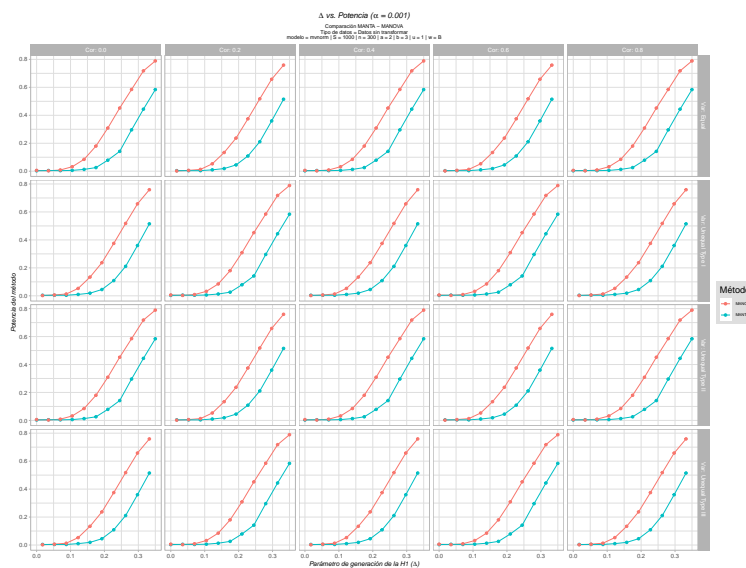
Figura B.4: A.



(a) A.



(b) A.



(c) A.

Figura B.5: A.

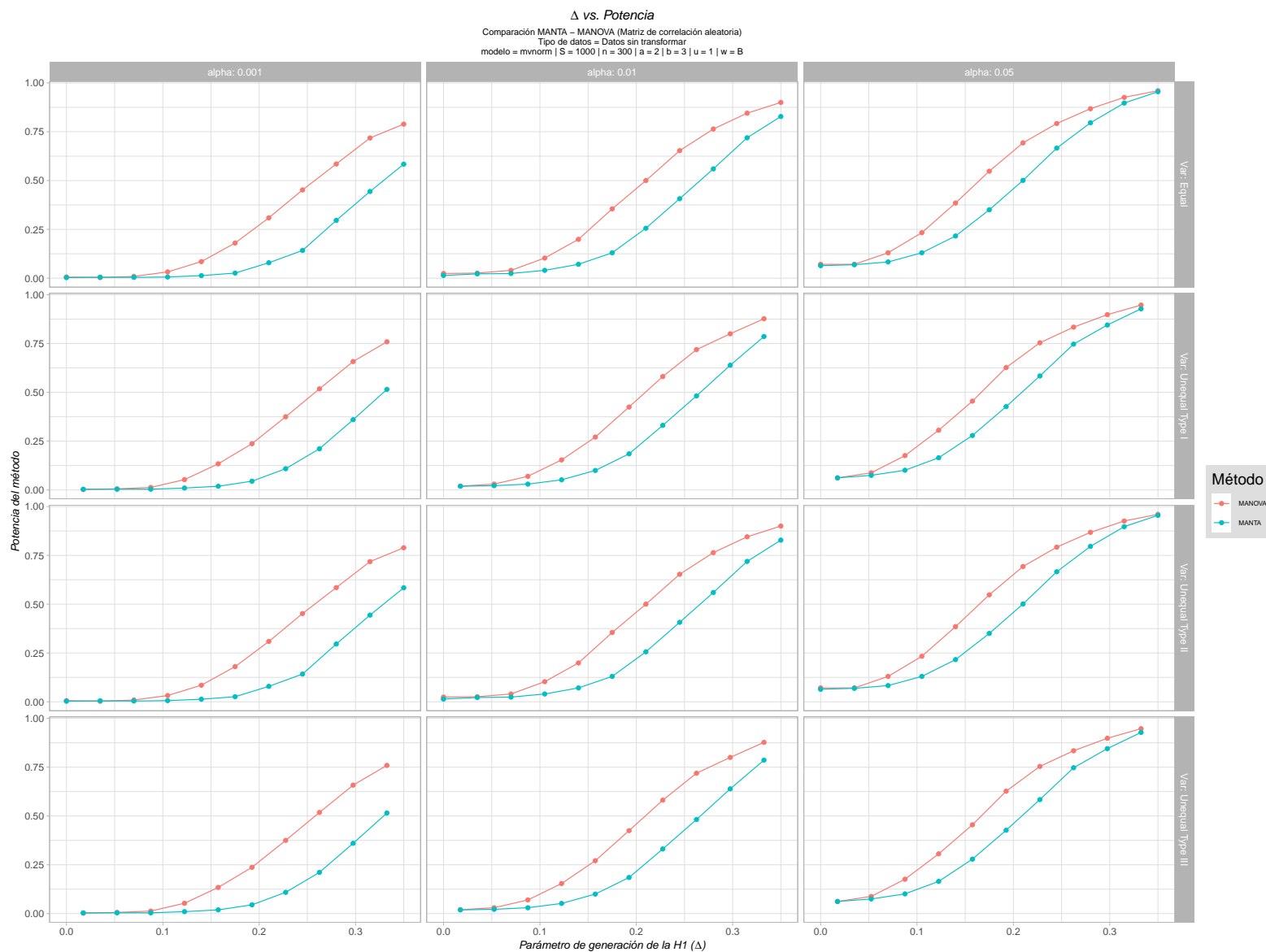


Figura B.6: A.

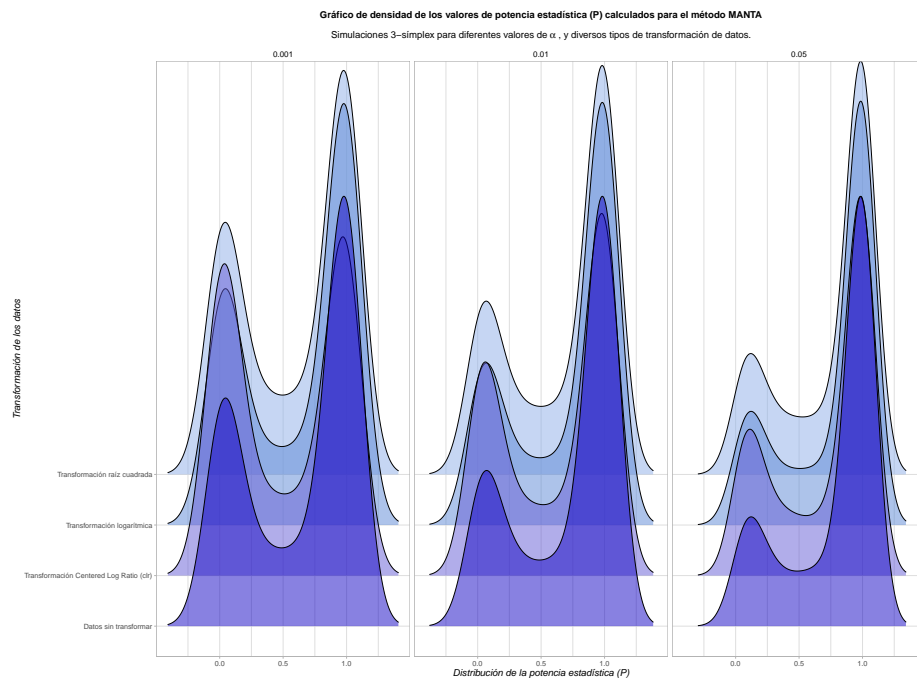


Figura B.7: A.

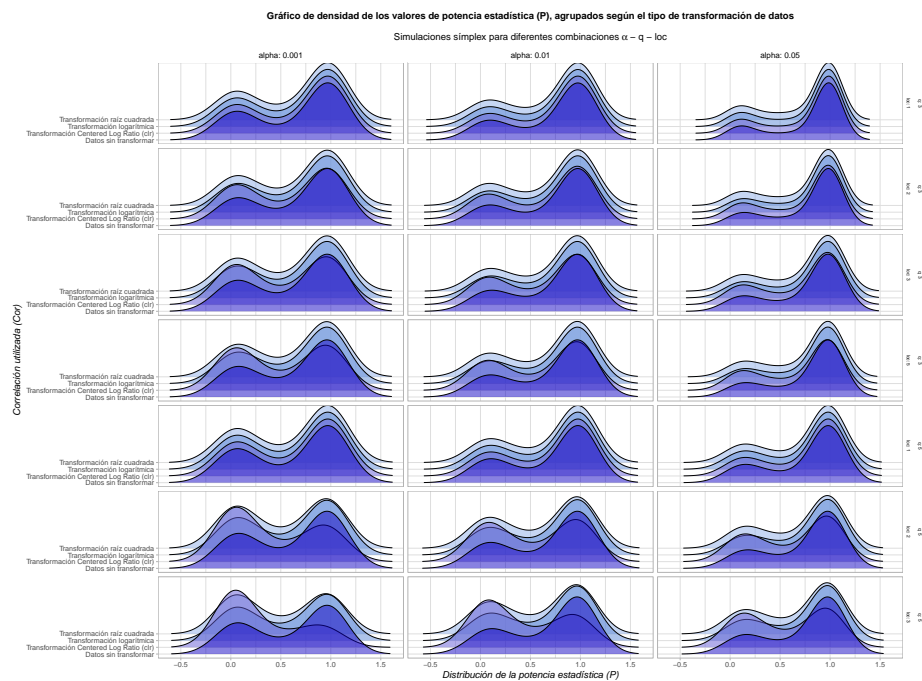


Figura B.8: A.

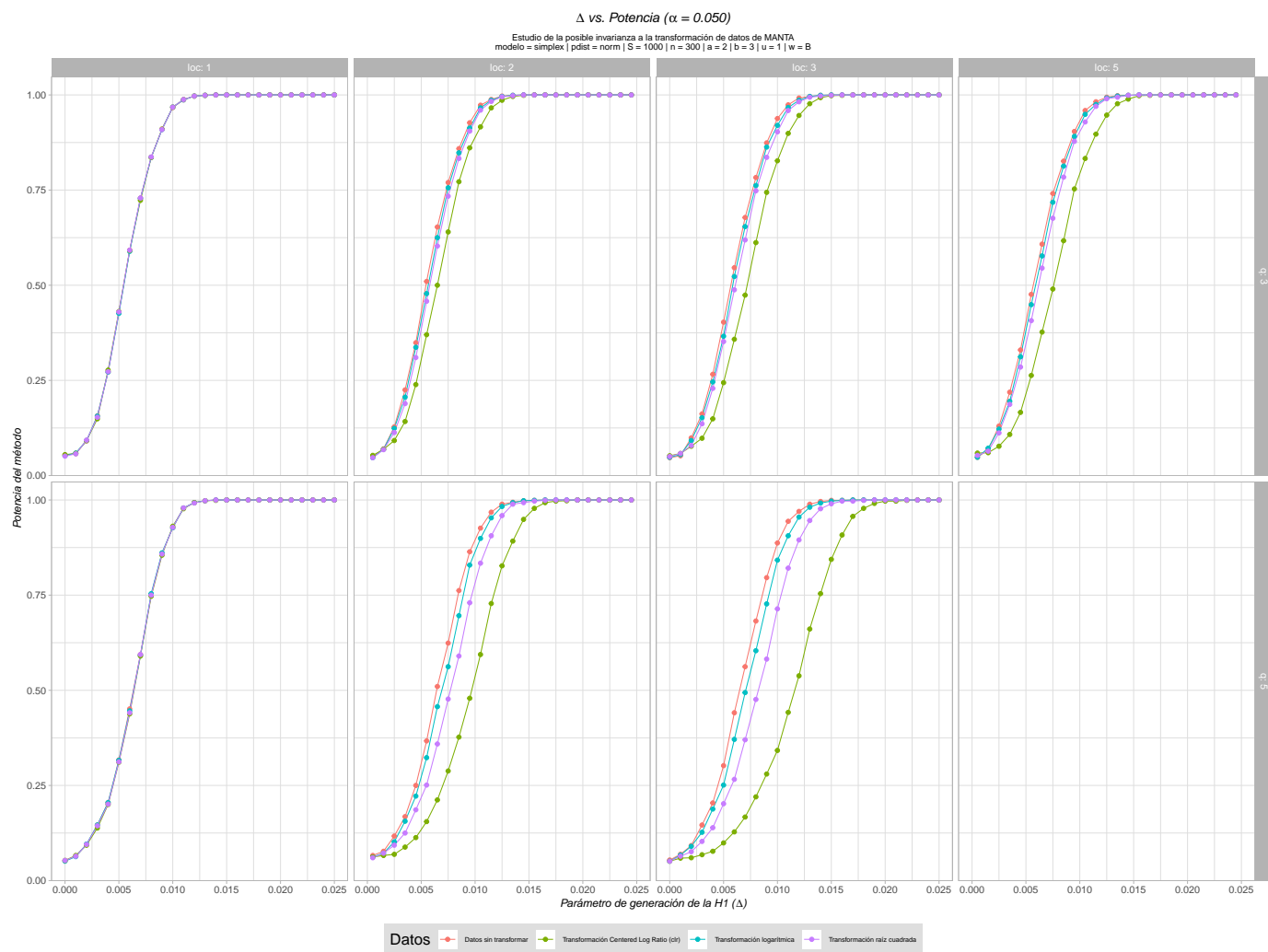
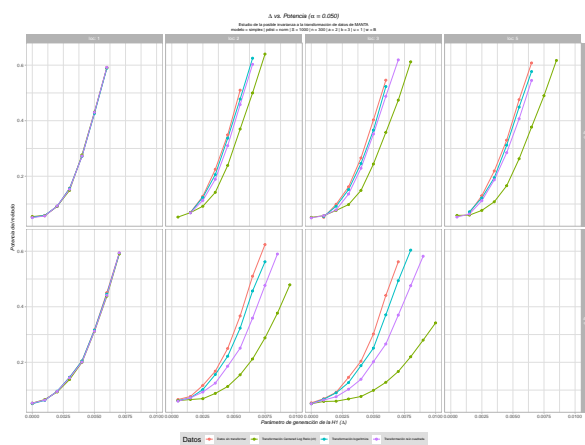
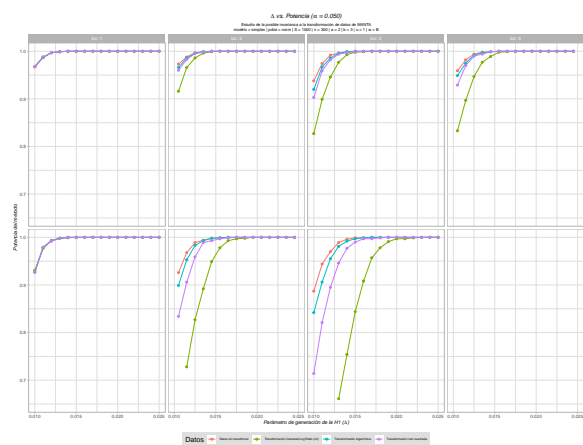


Figura B.9: A.



(a) A.



(b) A.

Figura B.10: A.

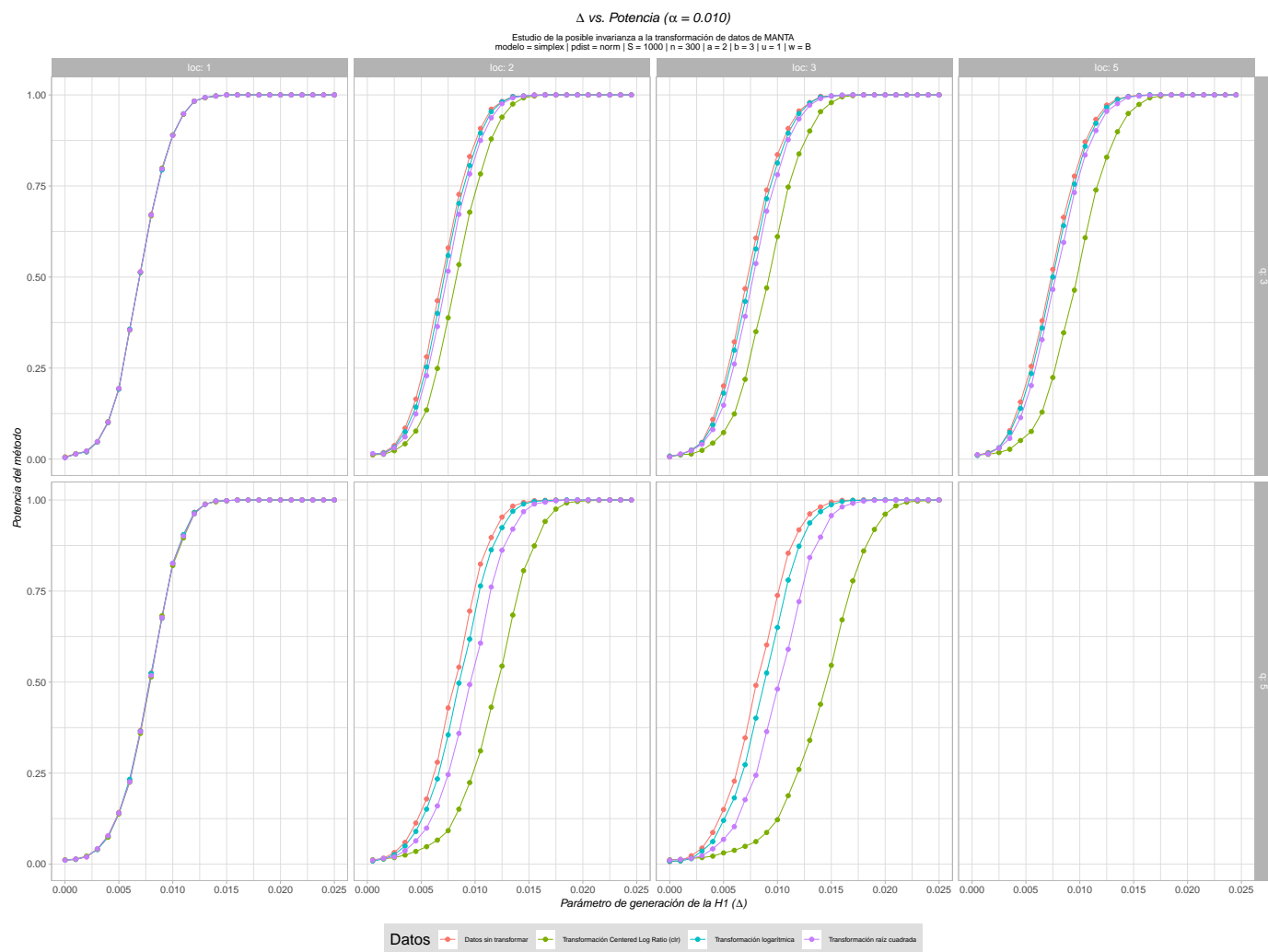
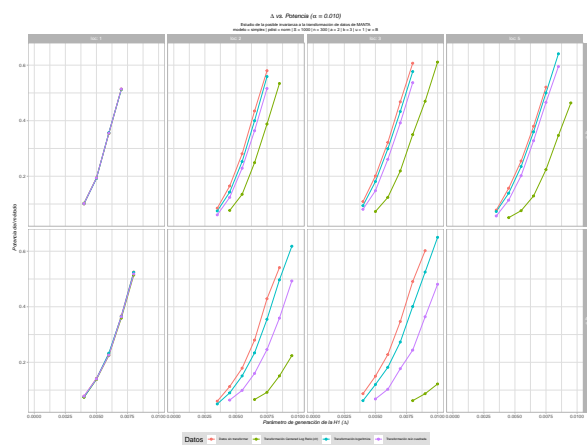
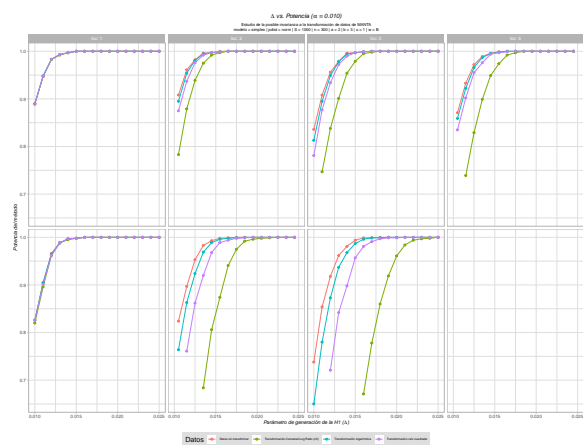


Figura B.11: A.



(a) A.



(b) A.

Figura B.12: A.

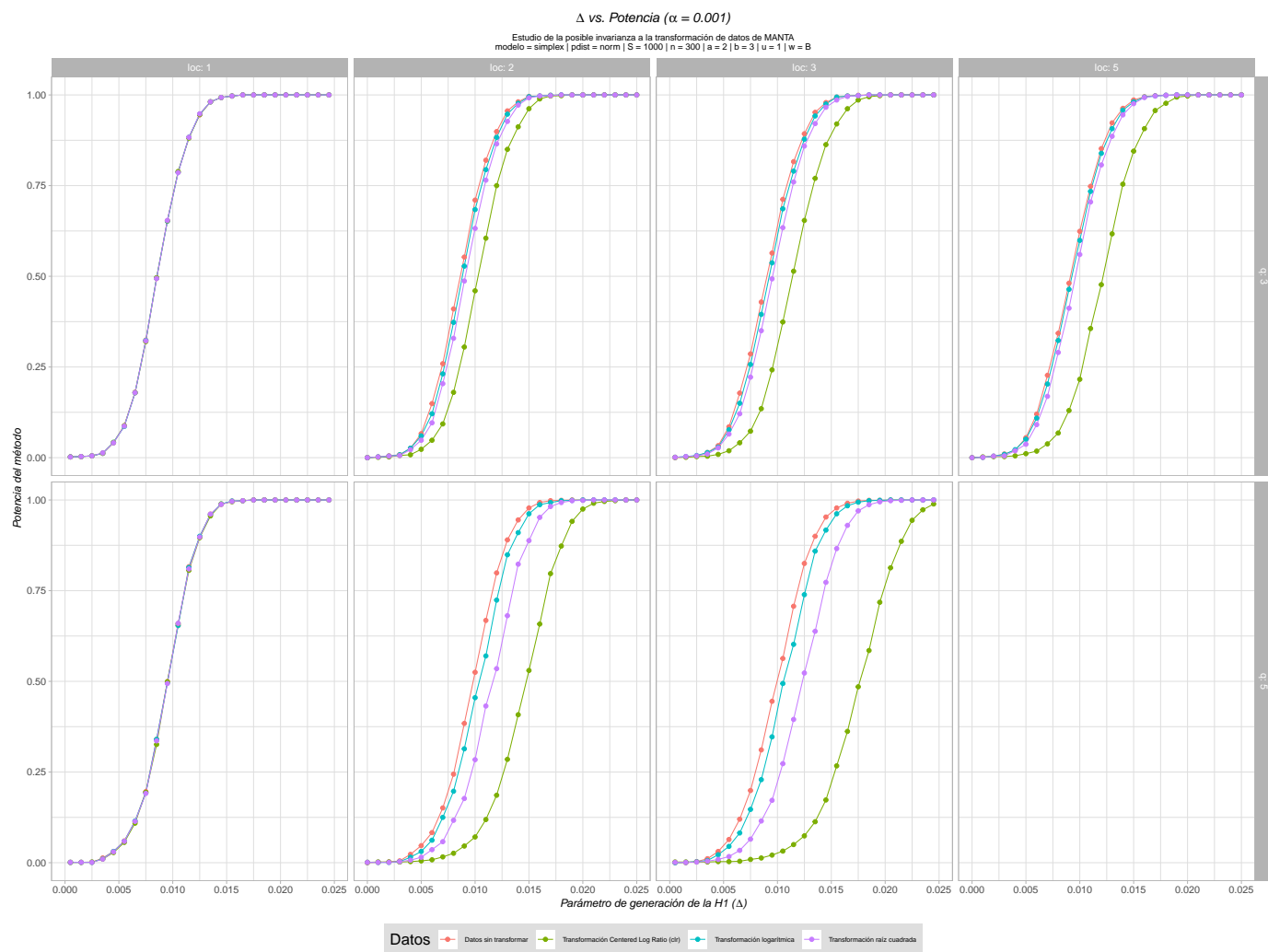
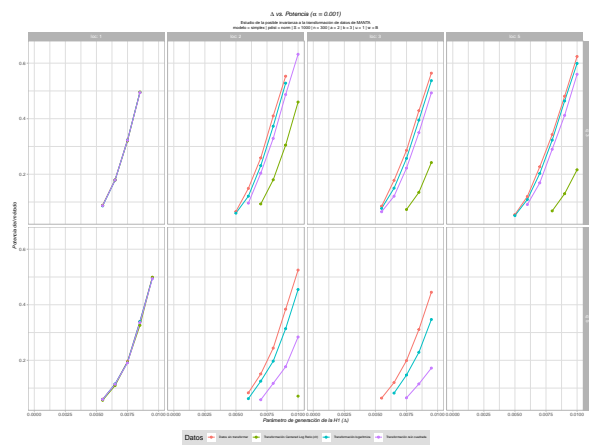
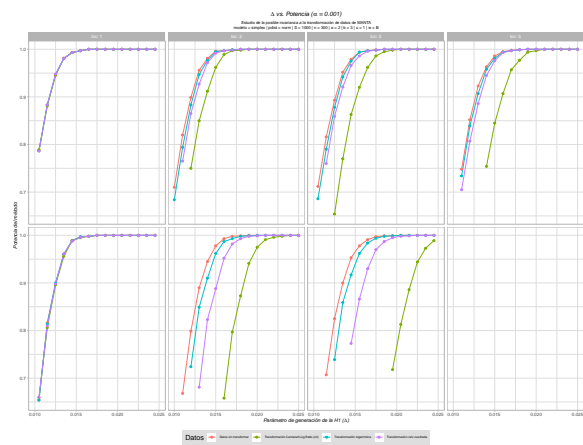


Figura B.13: A.



(a) A.



(b) A.

Figura B.14: A.