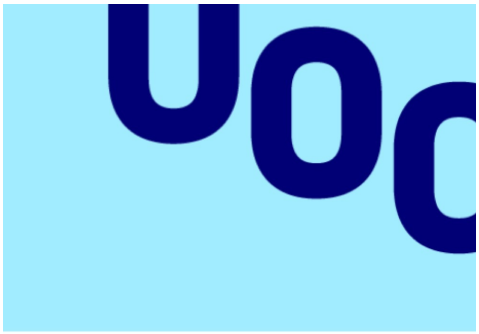
Assessing the properties of asymptotic PERMANOVA test through comprehensive simulations in the context of genetic studies



Universitat Oberta de Catalunya

Aitor Invernón de Campos

Analisis de datos Ómicos

Máster universitario de Bioinformática y Bioestadística (UOC, UB)

Miquel Calvo Llorca
Diego Garrido-Martín

David Merino Arranz

17/10/2023

UNIVERSITAT DE BARCELONA

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:	17/10/2023
Titulación o programa:	Máster universitario de Bioinformática y Bioestadística (UOC, UB)
Área del trabajo final:	Analisis 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

Este será mi resumen.

Abstract

This will be my abstract.

Índice general

Índice de figuras

Índice de tablas

Índice de Ecuaciones

PEC1

Capítulo 0

PEC1 - Plan de trabajo

0.1. 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 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* ([1]) o análisis de correlación canónica, y ***MultiPhen*** [2] que utiliza la regresión ordinal) adolecen de la posibilidad de evaluar diseños complejos que presentan mutiples interacciones entre el genotipo y otras covariables.
- Tanto el análisis multivariante de la varianza (***MANOVA***), como el de los modelos multivariantes lineales mixtos (***mvLMMs***) [3], resultan ser más tolerantes a estos diseños complejos al tratar los fenotipos como variables dependientes, introduciendo de froma 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*** [4] o ***MOSTest*** [5] [6] 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* ([7]) 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.* ([8] y [9]), y el trabajo de Monlong, J. *et al.* [10]. Donde, gracias al programa ***MANTA*** ([11], desarrollado principalmente en R), se estudia mediante diversas simulaciones ([12]) 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. ([13], [14]), o aproximaciones numéricas como la de Liu–Tang–Zhang ([15]), 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.

0.2. Descripción general

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 ([8], [9]), [10]), 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.

0.3. Objetivos del trabajo

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. ([13], [14]), o aproximaciones numéricas como la de Liu–Tang–Zhang ([15]), 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.

0.4. Enfoque y método seguido

En cuanto al tiempo de dedicación, se ha enfocado el trabajo siguiendo las pautas marcadas por las diferentes entregas programadas por la UOC (*Pruebas de Evaluación Continua* o *PEC*), estableciendo los siguientes bloques de trabajo:

- **Primera entrega:** *PEC1 - Definición y plan de trabajo* (10 % de dedicación).
- **Segunda entrega:** *PEC2 - Desarrollo del trabajo - Fase 1* (35 % de dedicación).
- **Tercera entrega:** *PEC3 - Desarrollo del trabajo - Fase 2* (35 % de dedicación).
- **Cuarta entrega:** *PEC4 - Cierre de la memoria y de la presentación* (15 % de dedicación).
- **Defensa pública** (5 % de dedicación).

Se puede encontrar una planificación más detallada en la sección *Planificación del trabajo*.

0.5. Planificación del trabajo

Una planificación orientativa de las tareas que conforman cada bloque de trabajo específico ideado, basados en la estructura de las diferentes PEC a entregar y de las necesidades del tema escogido, puede encontrarse en 1.

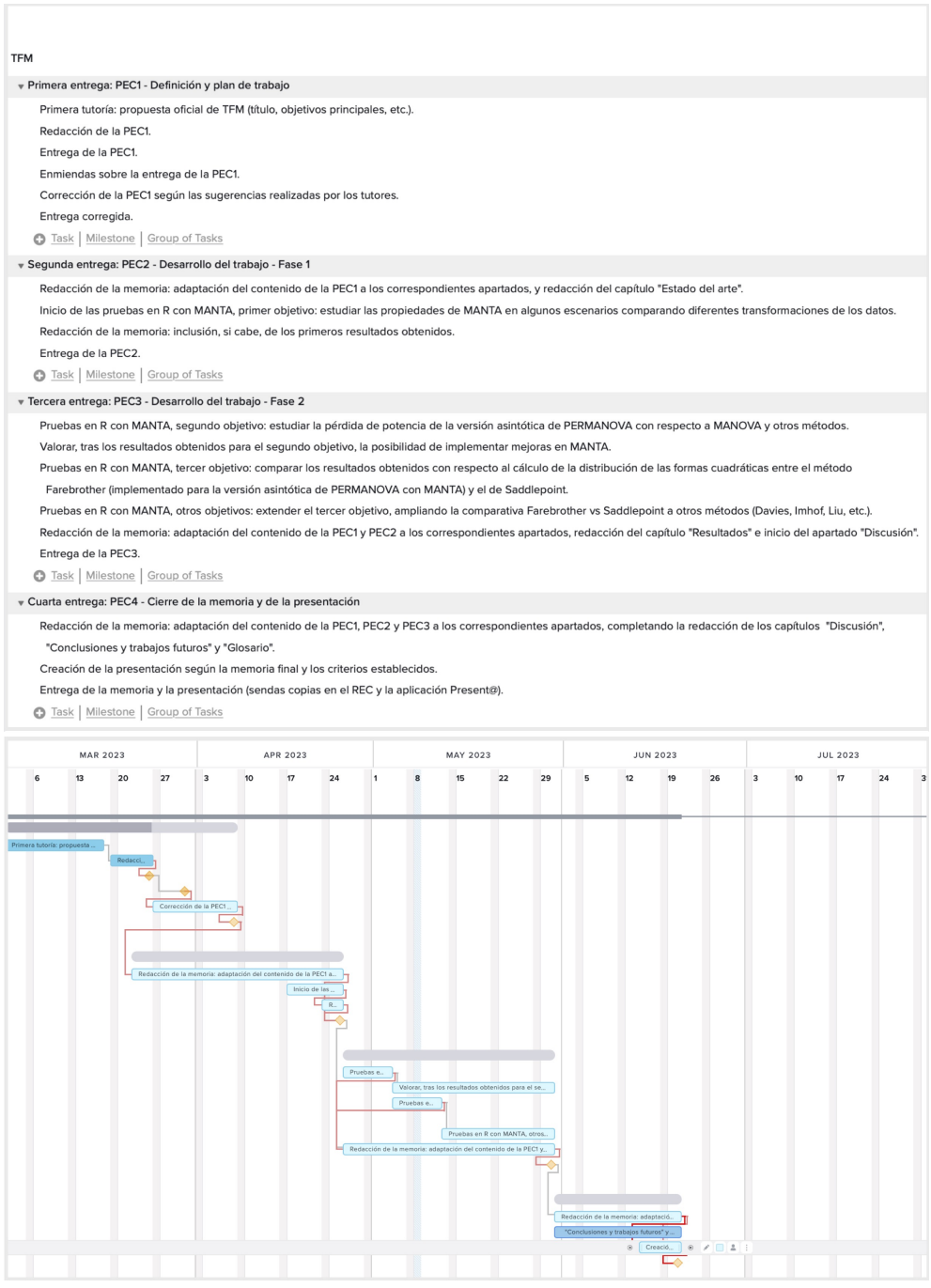


Figura 1: Planificación orientativa de las tareas a realizar para la consecución de la memoria y presentación del presente TFM: (*supra*) desglose de los diferentes objetivos y tareas propuestas; (*infra*) diagrama Gantt correspondiente.

0.6. 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 *R*, 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 *R*, 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 *R*, 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 *R*, 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@*).

0.7. Análisis de riesgos

En esta sección se indicarán algunas de las contingencias que pudieran surgir durante la realización del proyecto, indicando, a su conclusión, 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 han planteado. Ser capaz de maximizar la dedicación con el tiempo disponible, e identificar a tiempo los escollos que pueden atascar el proceso, serán claves para mitigar sus efectos.
- ***Planificación incorrecta:*** una posible mala priorización o asignación de tiempo a las tareas pertinentes puede influir de manera negativa en la consecución de los objetivos, sobre todo si existe dependencia de la tarea afectada por parte de alguna otra. También puede afectar de manera negativa una infravaloración de la dificultad de alguno de los objetivos propuestos, ya sea por el tiempo requerido para su consecución, o por falta de los conocimientos necesarios.
- **Etapa de análisis y pruebas:** en ella pueden surgir diversos contratiempos, como la dificultad en el manejo de los *scripts* de código que se pretenden utilizar, problemas inesperados con el computador utilizado, lentitud de los procesos ejecutados al tratar con grandes cantidades de datos, etc.

0.8. Breve sumario 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*). Así mismo, también se incluye una valoración de los posibles riesgos que pudieran surgir a lo largo de la elaboración del proyecto.
- **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.
- **Producto:** aun sin ser un objetivo principal, puede llegar a obtenerse una iteración mejorada del programario utilizado ya existente.
- **Presentación virtual del TFM:** exposición oral y visual basada en la memoria producida. En ella se resaltarán 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.

Incio de la 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 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* ([1]) o análisis de correlación canónica, y ***MultiPhen*** [2] que utiliza la regresión ordinal) adolecen de la posibilidad de evaluar diseños complejos que presentan mutiples interacciones entre el genotipo y otras covariables.
- Tanto el análisis multivariante de la varianza (***MANOVA***), como el de los modelos multivariantes lineales mixtos (***mvLMMs***) [3], resultan ser más tolerantes a estos diseños complejos al tratar los fenotipos como variables dependientes, introduciendo de froma 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*** [4] o ***MOSTest*** [5] [6] 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* ([7]) 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.* ([8] y [9]), y el trabajo de Monlong, J. *et al.* [10]. Donde, gracias al programa ***MANTA*** ([11], desarrollado principalmente en R), se estudia mediante diversas simulaciones ([12]) 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. ([13], [14]), o aproximaciones numéricas como la de Liu–Tang–Zhang ([15]), 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 ([8], [9]), [10]), 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. ([13], [14]), o aproximaciones numéricas como la de Liu–Tang–Zhang ([15]), 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 siguiendo las pautas marcadas por las diferentes entregas programadas por la UOC (*Pruebas de Evaluación Continua* o *PEC*), estableciendo los siguientes bloques de trabajo:

- **Primera entrega:** *PEC1 - Definición y plan de trabajo* (5 % de dedicación).
- **Segunda entrega:** *PEC2 - Desarrollo del trabajo - Fase 1* (40 % de dedicación).
- **Tercera entrega:** *PEC3 - Desarrollo del trabajo - Fase 2* (40 % de dedicación).
- **Cuarta entrega:** *PEC4 - Cierre de la memoria y de la presentación* (10 % de dedicación).
- **Defensa pública** (5 % de dedicación).

Se puede encontrar una planificación más detallada en la sección *Planificación del trabajo*.

1.4. Planificación del trabajo

Una planificación orientativa de las tareas que conforman cada bloque de trabajo específico ideado, basados en la estructura de las diferentes PEC a entregar y de las necesidades del tema escogido, puede encontrarse en 1.

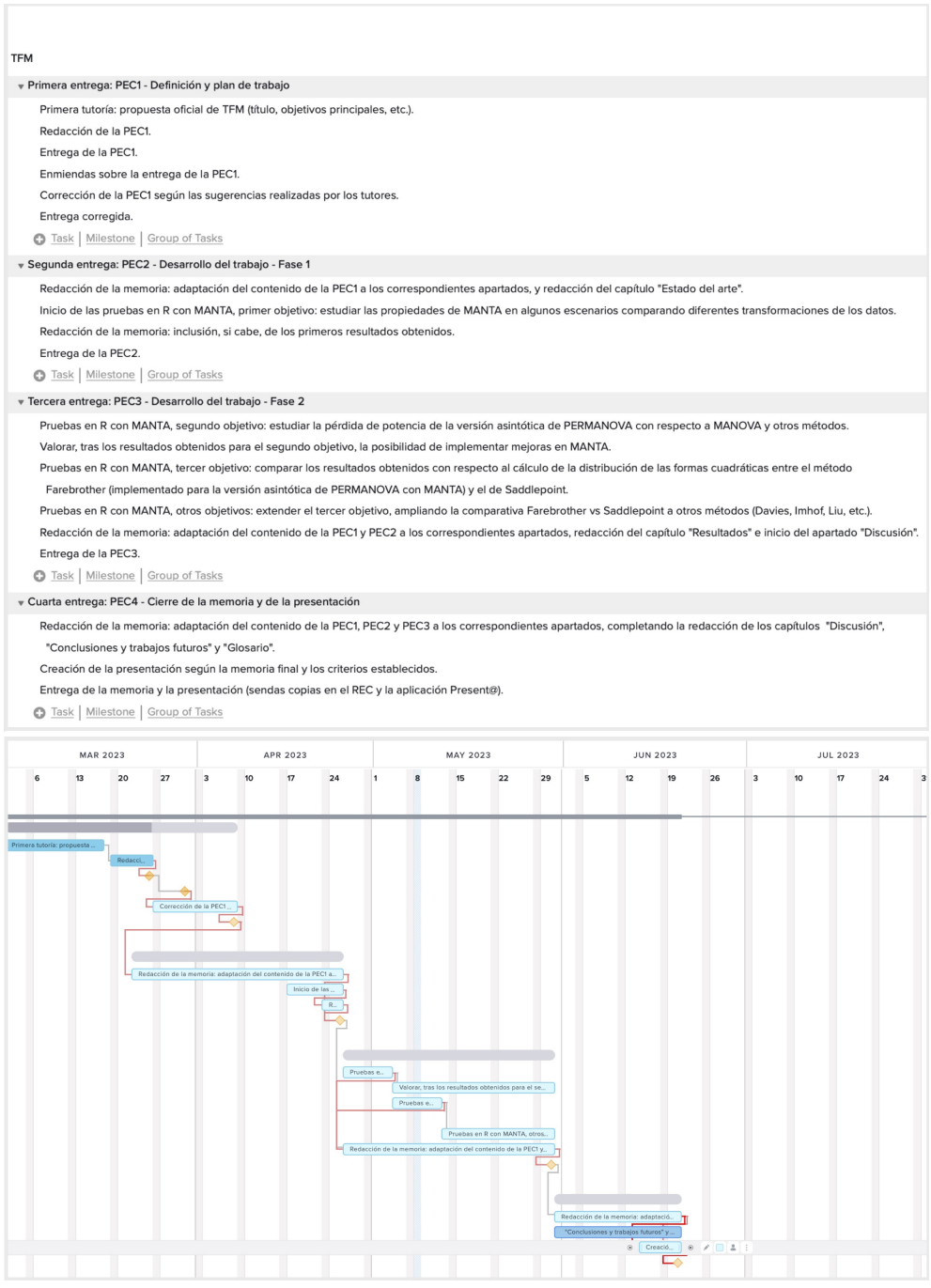


Figura 1.1: Planificación orientativa de las tareas a realizar para la consecución de la memoria y presentación del presente TFM: (*supra*) desglose de los diferentes objetivos y tareas propuestas; (*infra*) diagrama Gantt correspondiente.

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 *R*, 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 *R*, 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 *R*, 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 *R*, 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@*).

1.6. Análisis de riesgos

En esta sección se indicarán algunas de las contingencias que pudieran surgir durante la realización del proyecto, indicando, a su conclusión, 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 han planteado. Ser capaz de maximizar la dedicación con el tiempo disponible, e identificar a tiempo los escollos que pueden atascar el proceso, serán claves para mitigar sus efectos.
- ***Planificación incorrecta:*** una posible mala priorización o asignación de tiempo a las tareas pertinentes puede influir de manera negativa en la consecución de los objetivos, sobre todo si existe dependencia de la tarea afectada por parte de alguna otra. También puede afectar de manera negativa una infravaloración de la dificultad de alguno de los objetivos propuestos, ya sea por el tiempo requerido para su consecución, o por falta de los conocimientos necesarios.
- **Etapa de análisis y pruebas:** en ella pueden surgir diversos contratiempos, como la dificultad en el manejo de los *scripts* de código que se pretenden utilizar, problemas inesperados con el computador utilizado, lentitud de los procesos ejecutados al tratar con grandes cantidades de datos, etc.

1.7. Breve sumario 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*). Así mismo, también se incluye una valoración de los posibles riesgos que pudieran surgir a lo largo de la elaboración del proyecto.
- **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.
- **Producto:** aun sin ser un objetivo principal, puede llegar a obtenerse una iteración mejorada del programario utilizado ya existente.
- **Presentación virtual del TFM:** exposición oral y visual basada en la memoria producida. En ella se resaltarán 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.

1.8. 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

...

Capítulo 3

Resultados

...

Capítulo 4

Discusión

...

Capítulo 5

Conclusiones y trabajos futuros

5.1. Conclusiones

...

5.2. Líneas de futuro

...

5.3. Seguimiento de la planificación

...

Glosario

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. 6, 14

GWAS En genética, un estudio de asociación del genoma completo (Genome-wide association study) o WGAS (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. 6, 14

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. 7, 15

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. 8, 16

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. 7, 15

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. 7, 15

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). 7, 15

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. 7, 15

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. 6, 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). 6, 14

pseudo-F En el análisis multivariante de la varianza con permutacionesa (*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. 7, 15

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'. 6, 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. 6, 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. 6, 14

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. 6, 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. 8, 16

Referencias

[1] 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.

[2] 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

[3] 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

[4] 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

[5] "precimed/mostest," Jan. 2023, original-date: 2019-06-27T12:33:49Z. [Online]. Available: https://github.com/precimed/mostest

[6] 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

[7] 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

[8] 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

[9] 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

[10] 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

[11] D. Garrido-Martín, "MANTA," Feb. 2023, original-date: 2019-02-27T12:09:53Z. [Online]. Available: https://github.com/dgarrimar/manta

[12] ——, "manta-sim (sim)," Dec. 2022, original-date: 2022-02-16T12:40:23Z. [Online]. Available: https://github.com/dgarrimar/manta-sim

[13] 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

[14] 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

[15] 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