Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Departamento de Ecología, Genética y Evolución

Avances y limitaciones en la selección de embriones humanos: aneuploidías, mosaicismos y enfermedades poligénicas

Tesis presentada para optar al título de
Doctor de la Universidad de Buenos Aires en el área de Ciencias Biológicas

Juan Manuel Berros

Director de tesis	Dr. Hernán Dopazo
Consejera de estudios	Mg. Adriana Pérez
Lugar de trabajo	Biocódices S.A.

Buenos Aires, Argentina

Julio de 2022

Índice general

Título de la tesis

Avances y Limitaciones en la Selección de Embriones Humanos: Aneuploidías, Mosaicismos y Enfermedades Poligénicas

Resumen

El test genético preimplantacional (PGT) consiste de un conjunto de análisis genómicos que ayudan a decidir qué embriones transferir al útero durante la fertilización in vitro. En el primer capítulo de esta tesis, analizamos la posibilidad de realizar el PGT de aneuploidías (PGT-A) con datos generados por la tecnología genotyping-by-sequencing (GBS). Desarrollamos modelos matemáticos basados en el BAF (B allele frequency), simulamos aneuploidías de diverso tipo y niveles de mosaicismo, y diseñamos tests estadísticos para su detección. Los tests desarrollados logran una detección de aneuploidías con estándares actuales. En el segundo capítulo, analizamos las limitaciones que impone la pleiotropía en el PGT de riesgo de enfermedades poligénicas (PGT-P), un test nuevo y controversial de selección de embriones según su riesgo de enfermedades multifactoriales. Encontramos 14 pares de enfermedades con puntajes poligénicos de riesgo (PRS) correlacionados negativamente en población europea. Mediante una simulación de embriones, hallamos que en 9 de esos pares la selección del embrión de PRS mínimo de una enfermedad por parte de una pareja implicaría un aumento inesperado del riesgo de otra enfermedad en el embrión elegido.

Palabras clave

test genético preimplantacional, puntaje de riesgo poligénico, selección de embriones, aneuploidía, mosaicismo, rasgo cuantitativo, pleiotropía, correlación genética, genética de poblaciones, medicina reproductiva

---

Thesis title

Advances and Limitations in Human Embryo Selection: Aneuploidy, Mosaicism and Polygenic Disease

Abstract

Preimplantation genetic testing (PGT) consists of a group of genomic analyses which help decide which embryos should be transferred to the uterus during in vitro fertilization. In the first chapter of this thesis, we analyze the possibility of performing a PGT for aneuploidy (PGT-A) from data generated with the technology genotyping-by-sequencing (GBS). We developed mathematical models based on the BAF (B allele frequency), we simulated aneuploidies of multiple types and mosaicism levels, and we designed statistical tests for their detection. The tests detect aneuploidy with the current standards. In the second chapter, we analyze the limitations imposed by pleiotropy on the PGT for polygenic disease risk (PGT-P), a recently introduced and controversial test that selects embryos based on their estimated risk of several multifactorial diseases. We found 14 pairs of diseases that have negatively correlated polygenic risk scores (PRS) in the European population. Through a simulation of embryos, we found that in 9 of those pairs, if a couple selects the embryo with minimum PRS of a disease, this would entail an unexpected increase in the risk of other disease in the chosen embryo.

Keywords

preimplantation genetic testing, polygenic risk score, embryo selection, aneuploidy, mosaicism, quantitative trait, pleiotropy, genetic correlation, population genetics, reproductive medicine

Agradezco al equipo de Biocódices, que ha hecho de su laboratorio un espacio pujante de ciencia, tecnología e innovación, en el que he trabajado con gusto durante algo más de seis años.

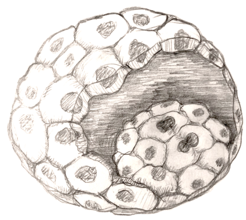
Agradezco a Leonardo Arbiza, Pierre Luisi, Marcos Miretti y Joaquín Dopazo, que han realizado comentarios detallados sobre versiones intermedias del trabajo aquí presentado.

Acerca del formato de esta tesis

Antes de comenzar, aclaramos brevemente algunos aspectos sobre el formato de esta tesis. La narración principal de cada capítulo se encuentra en un párrafo como este, ocupando la columna central de la página. Usaremos además notas numeradas en los márgenes que se mencionan en el texto principal con un superíndice numérico, al modo de nota al pie[1] .

[1] Ejemplo de nota marginal con una ecuación y figura:

$$p + q = 1$$



Las primeras ocurrencias de abreviaturas como polimorfismo de nucleótido simple (SNP) y conceptos como la aneuploidía se marcan con una tipografía distinta y tienen un hipervínculo hacia el glosario. Las definiciones en el glosario, además, tienen en la versión digital un hipervínculo de regreso a la página donde se utilizó el término, para poder continuar la lectura con comodidad. De igual manera, las referencias bibliográficas como [1] también tienen hipervínculos de regreso a las páginas donde son mencionadas.

Numeramos las ecuaciones como la **ecuación** (0.1), las figuras como la **Figura** 1 y las tablas como la **Tabla** 1 según el orden de aparición dentro de cada capítulo.

$$\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma^2\right) \tag{0.1}$$

Figura 1: Ilustramos con esta figura un cromosoma humano cualquiera.

Tabla 1: Tabla de ejemplo con las tipografías utilizadas en la tesis.

Fuente	Categoría	Función
Bembo	Serif	Texto principal
Helvetica	Sans Serif	Notas marginales y epígrafes
`Roboto Mono`	`Monospace`	`Código`

Usamos también algoritmos como el **Algoritmo** 0.1 para describir con pseudocódigo la lógica de un procedimiento, cuando no es de interés la implementación específica.

Algoritmo 0.1 Ejemplo de una función detallada en pseudocódigo.

function Función de Ejemplo(x,y)
    $z \leftarrow x + y$
    return z

Los fragmentos de código como el **Código** 0.1 son diferentes de los algoritmos. Los usamos cuando nos interesa mostrar los parámetros específicos de algún *software* utilizado, especialmente si se trata de programas de uso estándar en bioinformática, como `samtools`, `bcftools`, etc. El código es muchas veces ilustrativo, antes que directamente ejecutable, y presupone interpretar nombres de variables del estilo de `$CHROM` o `$PANEL_BED` que esperamos queden claras en su contexto.

Código 0.1: Ejemplo de una rutina en bash.

```bash
# Rutina de ejemplo
cd $TESIS_DIR
rm -rf *
```

Ocasionalmente, un tema auxiliar debe ser desarrollado de manera más extensa y el margen deja de ser adecuado. En estos casos, incluimos recuadros numerados como el **Recuadro** 0.1.

Usamos el punto como separador de decimales y un espacio pequeño como separador de miles. Por jemplo, $1/2 = 0.5$ y $10^4 = 10\,000$.

En la versión digital, los links en magenta son cliqueables y llevan a una URL externa, mientras que los números cliqueables en azul como (0.1) llevan a referencias numeradas dentro de la tesis.

Adoptamos un abuso de notación de la línea de comando de Unix: cuando decimos "los Paneles {30K, 60K, 100K} SNPs", queremos decir "el Panel 30K SNPs, el Panel 60K SNPs y el Panel 100K SNPs". En cada contexto de uso quedará claro, esperamos.

Decimos Mb por megabases y Kb por kilobases, para referirnos a la extensión de una región genómica. Por comodidad, abreviamos chrN para referirnos al cromosoma N.

Finalmente, adoptamos la siguiente notación para las distribuciones de probabilidad. Sean X, Y y Z variables aleatorias y sea μ una constante. $X \sim \mu$ indica que X se distribuye alrededor de un valor central μ. $X \sim Y$ indica que X se distribuye igual que Y. $X \sim Y \sim Z$ indica que las tres variables se distribuyen idénticamente. Si $\mathcal{F}$ es una distribución, entonces $X \dot{\sim} \mathcal{F}$ significa que X tiene una distribución aproximada $\mathcal{F}$.

Otras inspiraciones **Recuadro 0.1**

Algunas obras influyeron en esta tesis sin ser bibliografía en el sentido tradicional. Las comentamos a continuación.

Numerosas ideas de Edward Tufte se han transformado en el sentido común de los profesionales de la visualización de datos. En particular, los libros *The visual display of quantitative information* (1983) y *Envisioning Information* (1990) nos han inspirado en la realización de las figuras de la tesis. Resaltamos dos conceptos que nos han sido de especial provecho: la maximización del *data-ink ratio* y el recurso del *small multiple.*

También en el ámbito de la visualización de datos, nos han inspirado el celebrado ensayo *Up and down the ladder of abstraction*, de Bret Victor (2011), y las animaciones sobre temas de matemática creadas por Grant Sanderson bajo su pseudónimo 3blue1brown.

Finalmente, no podemos dejar de mencionar que somos herederos de una tradición del pensamiento biológico denominada "*the gene-centered view of evolution*". Nuestra cosmovisión biológica ha sido nutrida por obras como *Adaptation and Natural Selection* (1966), de George C. Williams, *The selfish gene* (1976) y *The extended phenotype* (1982), de Richard Dawkins, y *Darwin's dangerous idea* (1995), de Daniel Dennett.

A mis padres

On peut soupçonner de tout le phénomène. Il en est capable. L'hypothèse dénonce l'infini ; c'est ce qui la fait grande. Derrière le fait apparent elle cherche le fait réel. Elle demande à la création sa pensée, puis son arrière-pensée. Les grands inventeurs scientifiques sont ceux qui tiennent la nature pour suspecte. Suspecte d'accroissement, d'extension, d'exfoliation obscure, de pousses profondes dans toutes les directions, de végétation indéfinie ; suspecte de prolongements dans l'invisible. C'est vers ces prolongements que se dirige le tâtonnement sublime de l'hypothèse. Qui entrevoit ces prolongements dans l'invisible de la création est le mage ; qui entrevoit ces prolongements dans l'invisible de la destinée est le prophète.

[Se puede sospechar todo del fenómeno. De todo es capaz. La hipótesis delata al infinito; es lo que la hace grande. Detrás del hecho aparente, busca el hecho real. Interroga a la creación sobre sus intenciones y luego sobre sus segundas intenciones. Los grandes creadores de la ciencia son quienes toman a la naturaleza por sospechosa. Sospechosa de crecimiento, extensión, exfolación oscura, de brotes profundos en todas direcciones, de vegetación indefinida; sospechosa de prolongaciones hacia lo invisible. Hacia esas prolongaciones se dirige el tanteo sublime de la hipótesis. Quien entrevé esas prolongaciones hacia lo invisible de la creación es el mago; quien entrevé esas prolongaciones hacia lo invisible del destino es el profeta.]

Victor Hugo, *Les travailleurs de la mer*, 1866

Introducción

Tecnologías de test genético y selección de embriones

En el año 1959, el biólogo de la reproducción MC Chang logró, en un trabajo pionero de fertilización *in vitro* (FIV), el nacimiento de conejos negros a partir de conejas blancas [1]. Veinte años después, en 1978, el fisiólogo Robert Edwards intervino en el nacimiento de Louise Joy Brown, la primera bebé humana obtenida por FIV, lo que le valió el Premio Nobel de Medicina.

Diario *Evening News*, 25 de julio, 1978

Desde este caso hace 44 años, millones de bebés han nacido en el mundo gracias a esta técnica [2]. Los avances en cuatro décadas han sido muy importantes y abarcan la totalidad del proceso: la hiperestimulación ovárica, la posibilidad de inyectar un espermatozoide directamente en el óvulo, el congelamiento de los embriones y la mejora y diversificación de los tests genéticos que se les realizan. En su conjunto, los avances han elevado la tasa inicial de implantación en el útero, que inicialmente era menor a 5 % por embrión, hasta el valor actual de más de 50 % [2].

En Argentina, el primer nacimiento por FIV ocurrió en 1986 con dos mellizos tucumanos: Eliana y Pablo Delaporte. Al día de hoy, en nuestro país nacen alrededor de 3 500 bebés al año gracias a esta técnica, realizada en aproximadamente 70 clínicas de reproducción asistida.

Diario *Clarín*, 8 de febrero, 1986

0.1. El Test Genético Preimplantacional

La fertilización *in vitro* consiste en la fecundación de un óvulo con espermatozoides en un medio de cultivo. El cigoto resultante es cultivado durante alrededor de 6 días antes de ser transferido al útero, donde debe implantarse y continuar su desarrollo [3]. Típicamente, varios óvulos son fertilizados en un mismo ciclo, con la idea de incrementar la probabilidad de que alguno pueda lograr un embarazo exitoso.

En el pasado, se transferían varios embriones en simultáneo al útero, pero actualmente se recomienda elegir uno solo con el fin de evitar embarazos múltiples. La decisión de qué embrión transferir se basó durante un tiempo en criterios morfológicos, pero tras varias décadas de intentos no se pudo encontrar un único marcador morfológico que prediga con certeza la implantación y desarrollo exitoso del embrión [2].

En este contexto surge el estudio genético del embrión preimplantacional, para asistir la elección del embrión a transferir. Con diversas metodologías, se ha demostrado que el test genético preimplantacional permite mejorar la tasa de embarazos por transferencia[2], lo que reduce el tiempo de tratamiento hasta lograr el embarazo. Describimos brevemente la historia y los tipos de tests genéticos preimplantacionales a continuación.

[2] Véase la Figura 10 de [2].

0.1.1. Tipos de PGT

Hacia fines de los 80s se introdujo en la práctica clínica el diagnóstico genético preimplantacional (PGD), que consistía en el análisis de los embriones de FIV en busca de mutaciones asociadas a enfermedades monogénicas, particularmente cuando los padres eran potenciales portadores de una mutación [4].

El estudio genético de los embriones se extendió luego a la búsqueda de aneuploidías, consideradas una causa de abortos espontáneos recurrentes. Se distinguió entonces entre el diagnóstico del PGD y el *screening* genético preimplantacional (PGS). En contraste con el PGD, que se indica cuando se sospecha que los padres portan una variante asociada a enfermedad, el PGS no se basa en evidencia de que el embrión o los padres tengan una anomalía genética. El objetivo del PGS es anticiparse a problemas en la implantación y de esta manera reducir el tiempo de tratamiento [5].

En los últimos años se produjo un nuevo cambio de terminología: todos los tests se engloban bajo el concepto de test genético preimplantacional (PGT) y se discriminan diferentes tipos de estudios. Por un lado, lo que antes se llamaba PGD ahora se divide entre PGT de enfermedades monogénicas (PGT-M) y PGT de rearreglos estructurales (PGT-SR). Este último estudio se indica en casos de infertilidad repetida en familias. Por otro lado, lo que se denominaba PGS ahora es llamado PGT de aneuploidías (PGT-A) [5, 6].

Recientemente, se introdujo un nuevo estudio altamente controversial: el test del riesgo poligénico en el embrión.

0.1.2. PGT-P

El PGT de riesgo de enfermedades poligénicas (PGT-P)[3] fue introducido recientemente por un grupo de investigadores ligados a la compañía Genomic Prediction [7-9]. El objetivo del PGT-P es caracterizar el riesgo genético que los embriones tienen de contraer diversas enfermedades comunes y no transmisibles a lo largo de su vida –por ejemplo, diabetes tipo 2, esquizofrenia o enfermedad de arterias coronarias– con la finalidad de priorizar la transferencia del embrión que tenga menor riesgo. A diferencia del PGT-A, el PGT-P discrimina entre embriones *euploides*. A diferencia del PGT-M, las enfermedades analizadas son poligénicas: no existe uno o pocos genes que determinen completamente la condición, sino una multitud de *loci* asociados a lo largo del genoma.

[3] Por comodidad, en ocasiones diremos "riesgo poligénico" por "riesgo *de enfermedades* poligénicas".

El riesgo genético en este test es cuantificado mediante un puntaje de riesgo poligénico (PRS). El PRS es un estimador de la propensión genética de un individuo a contraer una enfermedad. Esta propensión puede interpretarse como la probabilidad que tiene el individuo de contraer la enfermedad en el transcurso de su vida, dado su genoma particular, si se ignora completamente el tipo de ambiente al que será expuesto. A diferencia de una aneuploidía o de una mutación asociada a una enfermedad monogénica, el PRS permite clasificar al individuo según su riesgo, pero no asegura que la enfermedad eventualmente ocurra.

Si bien se sabe que los PRS están asociados significativamente a diversas enfermedades, sus valores bajos de correlación y su perfomance predictiva reducida por fuera de la población en la que fueron desarrollados hacen que todavía no haya consenso para adoptarlos en la práctica clínica.

En este contexto, a pesar de sus limitaciones y de múltiples voces que llaman a la precaución [10, 11], el PGT-P ya es comercializado por algunas compañías. Por un lado, Genomic Prediction lo ofrece a través de su plataforma comercial LifeView para analizar el riesgo de doce enfermedades: esquizofrenia, diabetes, varios tipos de cáncer, hipertensión, infartos y otras. Por otro lado, la compañía Orchid ofrece encontrar al embrión "más saludable" con respecto al riesgo de esquizofrenia, Alzheimer, infartos, cáncer de mama o de próstata, diabetes y otras condiciones.

Desarrollaremos en detalle el modo en el que se calcula el PRS y las posibles limitaciones en su aplicación a embriones en el capítulo 2.

0.2. Las tecnologías del PGT

Las tecnologías utilizadas para el PGT evolucionaron notablemente en las últimas décadas. En esta sección describimos las sucesivas técnicas utilizadas con foco en la búsqueda de aneuploidías y mencionamos algunas de sus limitaciones, pues esto será relevante para los objetivos del capítulo 1.

0.2.1. FISH

Inicialmente, el test genético del embrión consistía en una biopsia seguida del uso de hibridización fluorescente *in situ* (FISH) para detectar deleciones y translocaciones en algunos cromosomas, en particular cuando involucraban genes asociados a la discapacidad intelectual [12]. La técnica de FISH se basa en el uso de sondas de ADN complementarias a regiones específicas del genoma y marcadas con distintos fluoróforos. Al hibridizar las sondas con el ADN de las células analizadas, se puede visualizar la presencia o ausencia de las regiones *target* bajo un microscopio. Con ello, se infiere la presencia o ausencia de los cromosomas analizados.

Retrospectivamente, se ha denominado a este período como **PGS v1**. FISH era utilizado para diagnosticar la pérdida o ganancia de algunos cromosomas asociados a síndromes: típicamente el 13, 18, 21, X e Y, como se ejemplifica en la **Figura** 0.2.

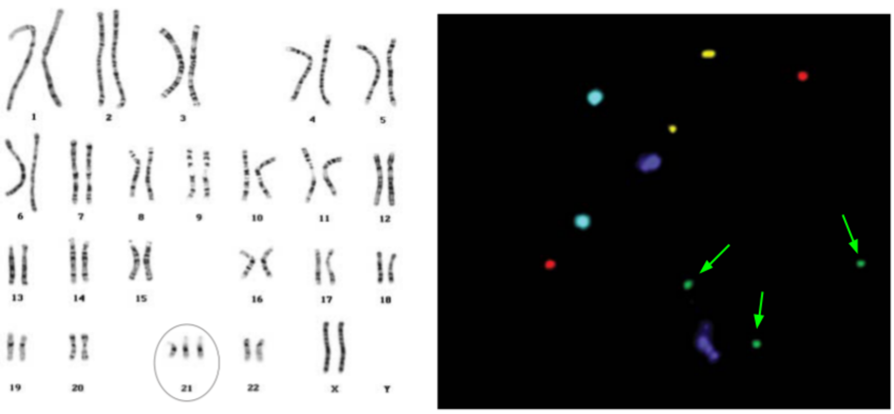


Figura 0.2: Izquierda: cariotipo humano con trisomía del cromosoma 21, tomado de [13]. **Derecha**: FISH con cinco tipos de sondas, para el análisis de aneuploidías en cinco cromosomas, tomado de [14]. Se observan dos regiones de cada color, correspondientes a dos copias de cada uno de los cromosomas analizados, excepto por los tres puntos verdes (señalados con flechas), correspondientes a las tres copias del cromosoma 21.

La técnica tiene algunas limitaciones: no permite analizar cómodamente todos los cromosomas, ni detectar pérdidas o ganancias a nivel sub-cromosómico. Tras algunos años de uso, no se pudo demostrar que FISH mejorara la tasa de nacimientos, lo que llevó a su reemplazo con otras tecnologías [15].

0.2.2. qPCR

La PCR cuantitativa o qPCR es una técnica de laboratorio basada en la PCR (reacción en cadena de la polimerasa) que sirve para determinar la presencia y cuantificar en una muestra el ADN de regiones específicas a medida que es amplificado, ciclo por ciclo [16]. En su aplicación para PGT-A, permite identificar aneuploidía de cromosomas completos de manera relativamente poco costosa y rápida, con la utilización de cuatro sondas por cromosoma [17]. Sin embargo, esto mismo constituye su limitación principal: 96 sondas totales para todo el genoma no permiten detectar mosaicismo, aneuploidías segmentales, ni tampoco translocaciones no balanceadas cuando la sonda no se ubica en el segmento afectado [18].

0.2.3. aCGH

Al igual que la qPCR, la técnica de *array Comparative Genomic Hibridization* (aCGH) permite analizar todos los cromosomas, lo que suele señalarse como la característica principal de la fase **PGS v2**.

El análisis comprehensivo del genoma evidenció que las aneuploidías de embriones preimplantacionales pueden ocurrir en cualquiera de los 24 cromosomas y no únicamente en los que eran observados con FISH [2, 3, 19]. Se comprobó que el *screening* completo mejora la selección de embriones [20], por lo cual el aCGH se volvió la práctica estándar en PGT-A. Otra ventaja del aCGH, tanto frente a FISH como a la qPCR, es que mejora notablemente la resolución del análisis, permitiendo la detección de aneuploidías segmentales y la detección de mosaicismos [15, 21].

El aCGH se basa en una técnica preexistente llamada CGH (*comparative genomic hybridization*), que era utilizada para el análisis del número de copias cromosómicas en tejido tumoral [22]. El principio básico tanto en CGH como en aCGH es el mismo: la cohibridización de fragmentos de ADN muestral y de ADN euploide de referencia o control, etiquetados con distintos fluoróforos, a oligonucleótidos inmobilizados en una plataforma y que representan regiones diferentes del genoma[4]. La fluorescencia del ADN de la muestra hibridizado se expresa, por región genómica, en relación a la fluorescencia del ADN euploide de referencia, mediante un estadístico denominado $\log_2 R$ (el logaritmo del cociente R de intensidades) [23, 24]. El valor de $\log_2 R$ permite detectar ganancias o pérdidas de material genético en la muestra analizada.

[4] Véase la Figura 1 de [23].

En la **Figura** 0.3 puede verse un ejemplo del tipo de gráfico que resulta del análisis de aCGH, con numerosas mediciones por cromosoma, lo que permite la detección de aneuploidías segmentales.

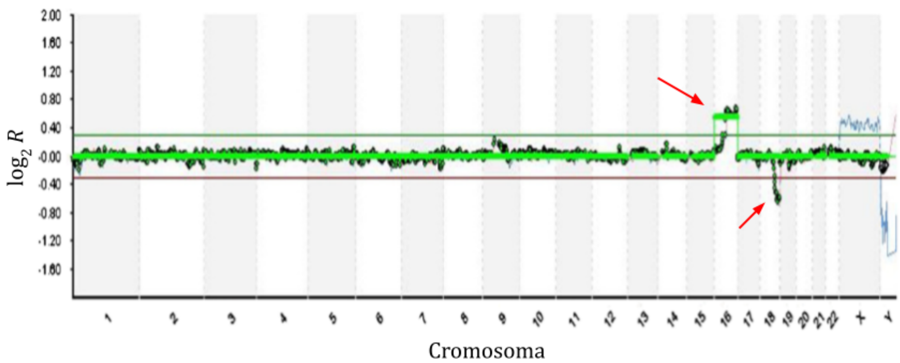


Figura 0.3: $\log_2 R$ en un análisis de aCGH, tomado de [25]. A diferencia de FISH, se analizan los 24 cromosomas y además hay múltiples mediciones por cromosoma. Las aneuploidías completas o segmentales se buscan con la ayuda de umbrales fijos del estadístico (líneas horizontales verde y roja). Las flechas indican una trisomía parcial del chr16 y una posible monosomía parcial del chr18.

El aCGH tiene limitaciones conocidas: no permite detectar poliploidía o disomía uniparental (UPD), porque las proporciones relativas de ADN son iguales a las del ADN control en esos casos, y tiene una capacidad limitada para detectar mosaicismo [15]. Por otro lado, el número de muestras que pueden ser analizadas con una única placa es limitado y cada placa adicional incrementa el costo del test proporcionalmente [26].

0.2.4. SNP *array*

Una tecnología que opera con principios similares al aCGH es el SNP *array*, que permite genotipificar cientos de miles de SNPs bialélicos. Cada SNP está representado en una superficie sólida con dos tipos de oligonucleótido: uno para el alelo A, con un fluoróforo, y otro para el alelo B, con otro fluoróforo. Cuando el ADN muestral hibridiza con los oligonucleótidos de la placa, emite el color de fluorescencia que corresponde a los alelos presentes en la muestra. Así, la obsrevación del color asociado al alelo A indica un genotipo homocigota AA, la observación del otro color indica un genotipo BB, y un color intermedio indica que ambos alelos están presentes, como es el caso de un genotipo heterocigota AB.

Los SNP *arrays* son principalmente utilizados en estudios de asociación de genoma completo, pero también pueden explotarse en PGT-A para generar un cariotipo, si se aprovecha el valor de intensidad de fluorescencia para estimar el número de copias por SNP, con una medida de $\log_2 R$ similar a la obtenida en aCGH [15].

Además, los SNP *arrays* proveen un tipo de dato adicional, que no está en aCGH: la frecuencia del alelo B (BAF), calculada como la intensidad del alelo B relativa a la intensidad total de A+B. En los sitios heterocigotas, el BAF se distribuye de distinta forma según el número de copias de cada región, un fenómeno conocido en la genómica del cáncer que se utiliza para detectar variaciones del número de copias (CNVs)[5]. Adicionalmente, el BAF permite detectar triploidía y UPD, los puntos ciegos del aCGH.

[5] Véanse por ejemplo el Cytoscan HD array [27], que explora el BAF y una medida relacionada, las *allelic differences*, y el *software* SiDCoN [28], que ofrece visualizaciones del BAF en diversos escenarios (Figura 1 y Figura 2).

En la **Figura** 0.4 se ilustran los dos tipos de dato presentes en el SNP *array* y se los compara al dato derivado del aCGH.

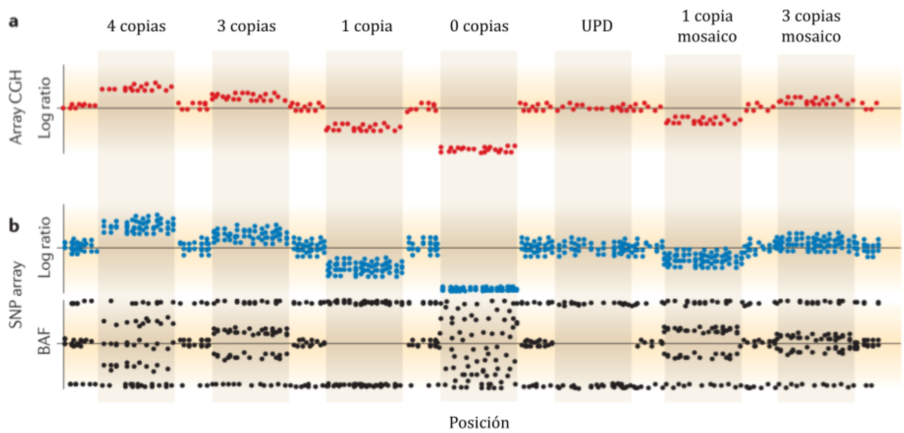


Figura 0.4: Arriba: ejemplo del $\log_2 R$ en aCGH bajo diversos escenarios de aneuploidía. **Abajo**: el $\log_2 R$ y el BAF, obtenidos con SNP *array*, en los mismos escenarios. Tomado de [29].

Los SNP *arrays* fueron validados para PGT y mostraron mejoras sustanciales respecto de FISH en especificidad y sensibilidad, además de la capacidad, compartida con el aCGH, de detectar aneuploidías en cualquiera de los 24 cromosomas y a nivel sub-cromosómico [25, 30, 31]. Puede leerse una comparación detallada con aCGH en [12].

En la actualidad, los SNP *arrays* se utilizan principalmente en la detección de desórdenes monogénicos [2], aunque recientemente Treff y col. [7] validaron la utilización de un SNP *array* de 800 mil marcadores como plataforma *única* para realizar en simultáneo PGT-M, PGT-SR, PGT-A y PGT-P.

Si bien tanto aCGH como los SNP *arrays* permitieron avances importantes en PGT, en los últimos años ambas tecnologías estan siendo reemplazadas por la tecnología de *next generation sequencing*, que describimos a continuación.

0.2.5. NGS

El término secuenciación de próxima generación (NGS) surgió para denominar a las tecnologías de secuenciación *masiva en paralelo* de ADN a mediados de los 2000s. Si bien sería preferible hablar de NGS como secuenciación "masivamente paralela” o secuenciación “de segunda generación” [32], utilizaremos el término NGS porque está muy establecido en la literatura científica.

Los instrumentos de secuenciación NGS comparten una serie de pasos comunes [33]. El proceso comienza con el ligamiento de adaptadores de ADN sintético a los fragmentos de ADN muestral que se quiere secuenciar. Luego, los fragmentos con adaptadores se fijan, con cierta distancia entre sí, a una superficie que tiene oligonucléotidos complementarios a los adaptadores. La fijación a la superficie permite asociar cada fragmento de ADN con una coordenada X-Y de la superficie. Cada fragmento inmovilizado es amplificado en el lugar con una reacción mediada por una polimerasa, lo que da lugar a *clusters* de moléculas idénticas. Este paso de amplificación es fuente de errores de secuenciación que serán arrastrados río abajo en el análisis, puesto que las polimerasas nunca son 100 % precisas.

La secuenciación propiamente dicha ocurre mediante la síntesis de ADN (*sequencing by synthesis*)[6], en una serie de pasos repetidos automáticamente, nucleótido a nucleótido. Se agregan al medio sucesivamente nucleótidos bloqueados químicamente de tal manera que sólo uno puede incorporarse a la vez a cada cadena. Estos nucleótidos se incorporan a las cadenas de ADN muestral sólo en los casos en los que hay complementariedad. La incorporación de cada nucleótido genera fluorescencia, que es registrada por una cámara. Luego, el grupo químico bloqueante es escindido y se repite el ciclo con la adición de otro nucleótido. En cada paso, la fluorescencia emitida por los *clusters* es registrada en cada coordenada. Al final del proceso, las fluorescencias detectadas en cada coordenada se traducen en una secuencia de nucleótidos inferidos [34].

[6] Detallamos el procedimiento centrándonos en la tecnología patentada por la compañía Illumina de secuenciación por síntesis, pero varios de estos conceptos se aplican también a protocolos de otras compañías.

A cada secuencia generada de este modo se la denomina lectura o *read* y, en la actualidad, suele tener entre 50 y 400 nucleótidos dependiendo del instrumento y del tipo de librerías utilizadas. En principio, no se sabe a qué región del genoma pertenece la lectura obtenida. En el caso del ser humano, la secuencia de aproximadamente 3 Gb del genoma ha sido descripta [35] y consta de diferentes versiones, que son denominadas genoma de referencia. Por ende, las lecturas obtenidas por NGS pueden ser alineadas a alguno de estos genomas de referencia, para localizarlas en un cromosoma y posición específicos[7].

[7] Probablemente el mejor modo de comprender el pipeline completo de NGS es recurrir a los recursos propios de Illumina. La Wikipedia de Illumina dye sequencing también es un buen recurso.

En la **Figura** 0.5 describimos esquemáticamente el pipeline de NGS desde la secuenciación hasta el alineamiento.

La tecnología de NGS ha sido utilizada en PGT-A porque presenta numerosas ventajas, principalmente mejores costos y eficiencia que el aCGH, aunque también mayor resolución, mejor detección de mosaicismo y mayor capacidad de automatización [19]. Adicionalmente, si bien los *arrays* individuales son económicos, el equipo necesario para escanear las placas es caro y no tiene otros usos. En cambio, un equipo de NGS tiene múltiples usos en un laboratorio, como tecnología que permite tanto genotipificar como secuenciar, lo que resulta costo efectivo. Por estas razones, las técnicas de *arrays* están siendo reemplazadas por NGS en la clínica.

Existen hoy dos técnicas de utilización de NGS en PGT-A: *non-targeted* y *targeted*. Las describimos a continuación.

0.2.6. *Non-targeted* NGS

La técnica de *non-targeted* NGS consiste en una secuenciación de baja profundidad que genera lecturas espaciadas entre sí a lo largo de todo el genoma. Las lecturas dispersas no permiten determinar genotipos, pero la cantidad de lecturas *en un intervalo* permite saber si esa región está sobre o sub-representada. Estos conteos deben ser comparados con distribuciones conocidas de antemano. Al conteo normalizado de lecturas por intervalo se lo suele graficar bajo el nombre de *copy number* y es análogo al $\log_2 R$ de aCGH: representa el número de copias de cada región cromosómica. Dado el *copy number* por región, se estable-

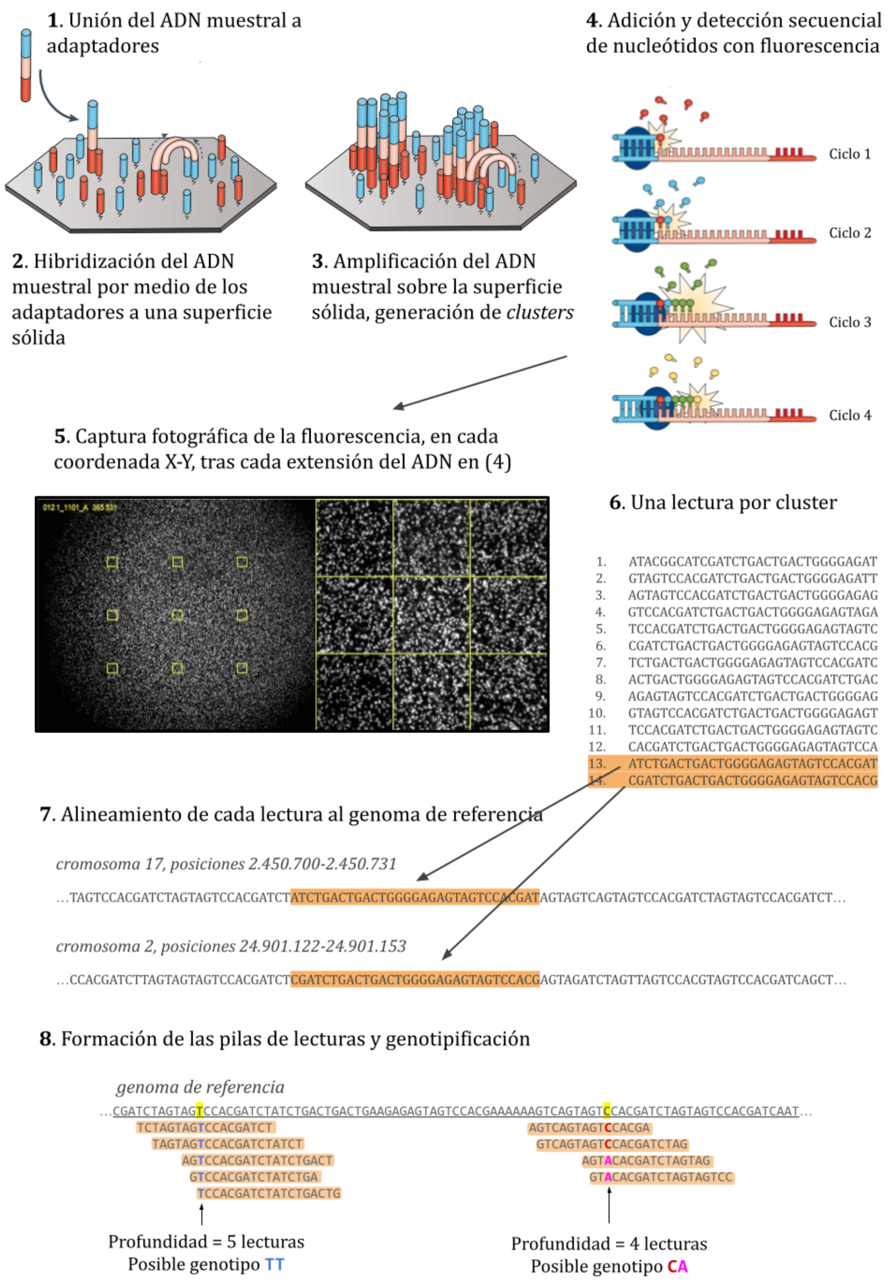


Figura 0.5: Ilustración esquemática del pipeline de NGS de Illumina, adaptado en parte de [36] con elaboración propia. Se observan (**1-3**) la fijación de los fragmentos de ADN muestral a una superficie y su amplificación, que genera *clusters* de moléculas idénticas. La secuenciación propiamente dicha ocurre de manera secuencial, con adiciones y detecciones sucesivas de nucleótidos con fluorescencia (**4**). Por cada adición de nucleótido, se toma una fotografía (**5**); las coordenadas que emiten fluorescencia evidencian la presencia de un nucleótido específico, complementario al nucleótido agregado. En (**6**), las fluorescencias sucesivas detectadas en cada coordenada X-Y de las fotografías se compilan como una secuencia de nucleótidos: la "lectura" de ADN. En el alineamiento, cada lectura es asociada a una región del genoma humano (**7**). Dependiendo de la profundidad de secuenciación, los sitios de interés pueden acumular muchas lecturas solapadas, formando una "pila" que aporta evidencia para genotificar al individuo (**8**).

ce la presencia de ganancias o pérdidas cuando el valor supera ciertos umbrales predefinidos (véase la **Figura** 0.6).

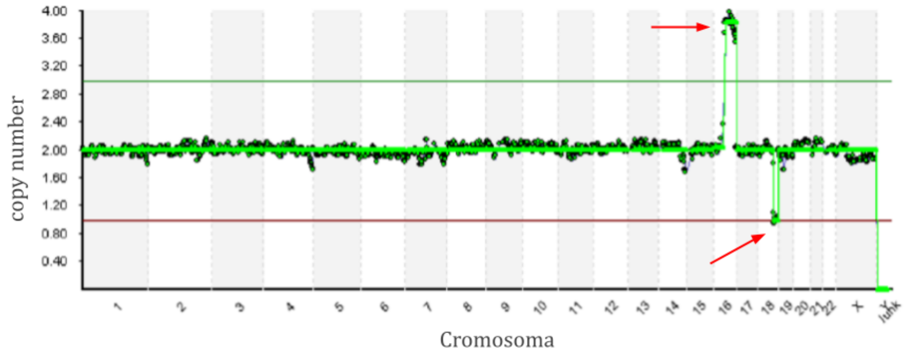


Figura 0.6: Gráfico de *copy number* generado a paritr de datos de NGS, tomado de [37]. El eje Y detalla el número de copias inferido en cada región, a partir del conteo de lecturas normalizado. Las flechas señalan una ganancia parcial del chr16 y una pérdida parcial del chr18.

La técnica de *non-targeted* NGS es la técnica NGS más utilizada hoy en PGT-A. Fue validada en numerosos estudios y ha demostrado tener una mejor resolución para la detección de aneuploidías segmentales, rearreglos estructurales y mosaicismo que aCGH [19, 26, 38].

Como toda técnica, tiene sus limitaciones. La normalización del conteo de lecturas que mencionamos toma como referencia la media de los autosomas, asumiendo que corresponde a un *copy number* de 2. Es decir, se asume que la mayoría del genoma es normal y que la aneuploidía afecta a pocos cromosomas o a regiones acotadas. La idea es buscar un contraste (crecimiento o disminución) del número de lecturas entre regiones euploides y regiones aneuploides del mismo genoma.

Este supuesto tiene algunos puntos ciegos. Por un lado, existen desórdenes genéticos en los que el número de copias no cambia, como la UPD, que resulta en tramos de pérdida de heterocigosis (LOH) y sólo puede detectarse mediante una genotipificación. Por otro lado, cuando el genoma *entero* tiene más copias, como en la triploidía ($3n$), no hay un contraste del conteo normalizado entre regiones. En este caso, la media autosómica es erróneamente utilizada como valor de referencia. Finalmente, al no disponer de genotipos, la metodología no puede extenderse para realizar PGT-M o el reciente PGT-P, tests en los que importa conocer el alelo presente en cada SNP.

Una estrategia diferente de utiliización del NGS, el *targeted* NGS, resuelve algunos estos problemas. Lo describimos a continuación.

0.2.7. *Targeted* NGS, pila y BAF

La técnica de *targeted* NGS parte de la amplificación con profundidad alta de un conjunto de regiones de interés en el genoma llamadas *targets*. La distribución de los *targets* puede variar según el objetivo de la secuenciación. Una distribución posible consiste en localizarlos de manera más o menos regular a lo largo de cada cromosoma, para obtener una muestra representativa de varias regiones sub-cromosómicas. Estos *targets*, además, pueden ubicarse especialmente en SNPs elegidos con algún criterio, con la idea de genotipificar al individuo. En la **Figura** 0.7 se ilustra el contraste entre *non-targeted* y *targeted* NGS.

La técnica de *targeted* NGS permite genotipificar porque genera un gran número de observaciones de cada alelo, como se ilustra en pila de lecturas de la **Figura** 0.8. Nótese que "la pila" aquí refiere a un modo de visualización de los datos típico en NGS, pero no a un alineamiento físico de las lecturas de ADN.

Cuando los *targets* coinciden con un SNP bialélico, el número de observaciones de cada alelo, (n_A, n_B), permite inferir el genotipo del individuo y posibilita el cálculo del BAF como:

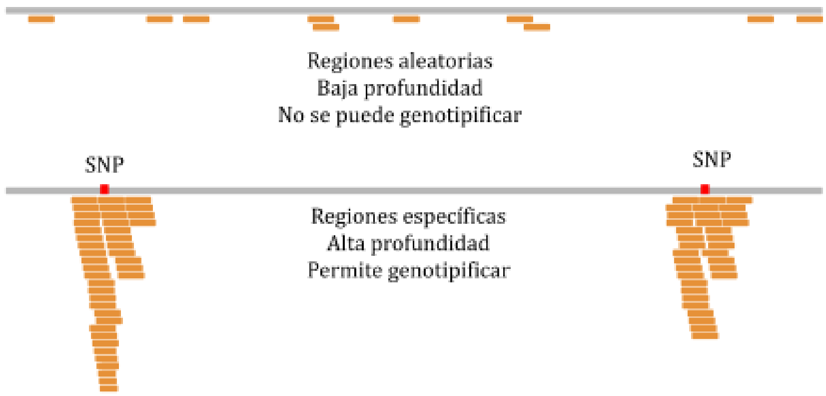


Figura 0.7: Diferencia entre *non-targeted* y *targeted* NGS. **Arriba**: el protocolo de *non-targeted* NGS consite en una secuenciación de regiones aleatorias con baja profundidad. **Abajo**: en *targeted* NGS se busca secuenciar con alta profundidad regiones de interés. Cuando el foco está puesto en SNPs, esta estrategia permite genotipificar.

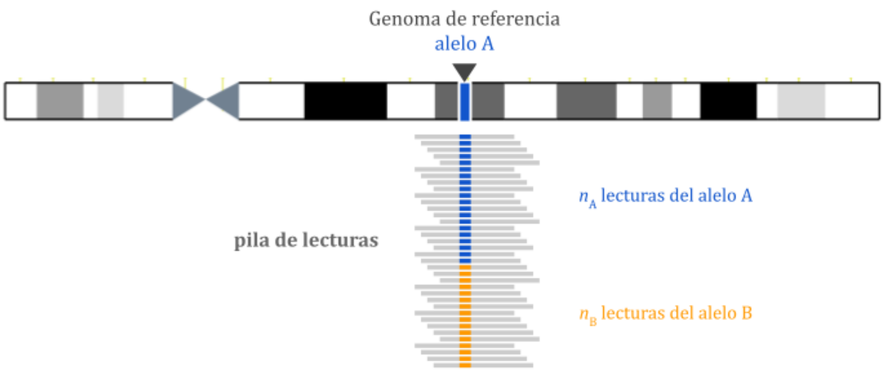


Figura 0.8: Pila de lecturas alineadas al genoma de referencia. La presencia de lecturas con ambos alelos del mismo SNP es evidencia de que el individuo secuenciado tiene un genotipo AB.

$$\mathrm{BAF} := \frac{n_\mathrm{B}}{n_\mathrm{A} + n_\mathrm{B}} = \frac{n_\mathrm{B}}{d} \tag{0.2}$$

donde d es la profundidad. Nótese que este modo de calcular el BAF difiere del cálculo basado en intensidades, propio del SNP *array*, pero el resultado es el mismo: una estimación de las cantidades relativas de cada uno de los alelos en la muestra.

La obtención del BAF complementa el dato del *copy number*, que se obtiene de igual manera que en *non targeted* NGS [39]. En este sentido, el *targeted* NGS es similar al SNP *array*, en que permite obtener dos datos complementarios: número de copias y proporciones alélicas.

La **Figura** 0.9 muestra un ejemplo en el que el BAF permite detectar una triploidía que no es posible detectar con el *copy number*.

Zimmerman y col. [39] validaron el *targeted* NGS para la detección de aneuploidías completas, con una concordancia del 99.2 % utilizando líneas celulares aneuploides de cariotipo conocido. Tiegs y col. [41] validaron una plataforma de *targeted* NGS también para la detección de aneuploidías completas, analizando embriones que fueron transferidos o no según criterios morfológicos en numerosos centros de fertilidad. El resultado fue nuevamente auspicioso para la técnica: el 64 % de los embriones clasificados como euploides por NGS llegaron a gestaciones o nacimientos saludables, mientras que ninguno de los embriones clasificados como aneuploides llegaron al nacimiento.

A pesar de las ventajas de esta técnica, el *targeted* NGS es poco usado aun en PGT-A.

0.2.8. GBS

El genotipado por secuenciación (GBS) es una estrategia de obtención de genotipos por NGS y, en este sentido, se la puede describir como una aplicación

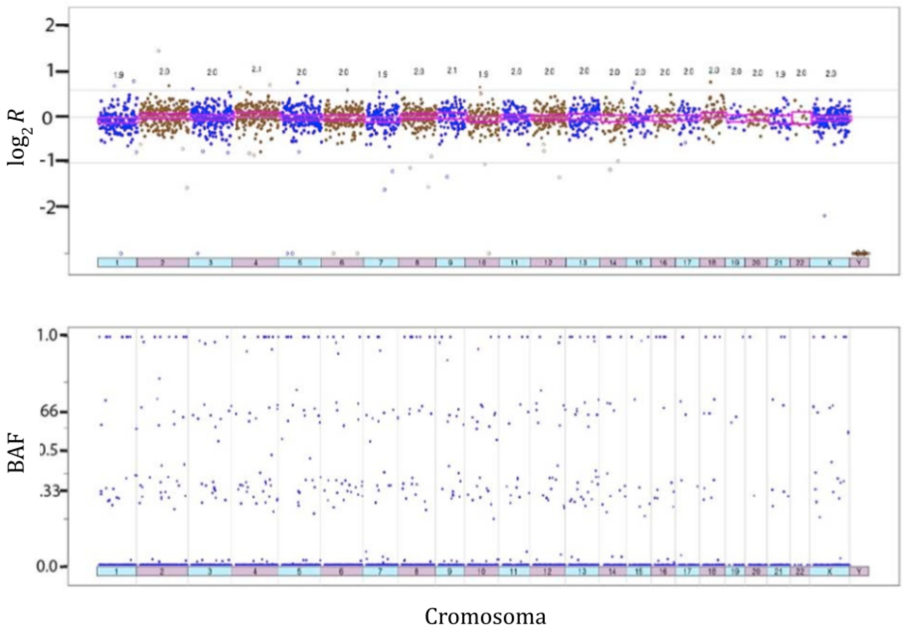


Figura 0.9: Detección de triploidía con datos de *targeted* NGS, tomado de [40]. **Arriba:** Número de copias, representado como $\log_2 R$, de un embrión que aparenta ser euploide 46,XX, pues todos los cromosomas se reparten alrededor de $\log_2 R = 0$. **Abajo**: La frecuencia del alelo B (BAF) se reparte alrededor de $1/3$ y $2/3$ en los sitios heterocigotas de todo en genoma. Esto permite inferir que el embrión es triploide (69,XXX).

particular del *targeted* NGS. Esta estrategia tiene usos en otras áreas de la genómica: los SNPs detectados por GBS son utilizados en GWAS, mapeo de *loci* cuantitativos y predicción genómica en muchas especies de plantas, por ejemplo el maíz [42, 43].

0.3. Biocódices

Esta tesis fue realizada íntegramente en Biocódices S.A, gracias a una Beca Interna de Doctorado en empresa otorgada por CONICET entre agosto de 2017 y julio de 2022. Biocódices es una empresa argentina de base tecnológica dedicada a la genómica reproductiva, oncológica y clínica. La empresa fue fundada en 2014 por el Dr. Hernán Dopazo (director de esta tesis), inicialmente como un emprendimiento de la incubadora de empresas de la Facultad de Ciencias Exactas y Naturales de la Universidad de Buenos Aires. Durante 2021 se estableció un contacto entre Biocódices y una empresa de genómica de EE.UU. para analizar el potencial de GBS para PGT-A, una estrategia que no es explotada por ninguna compañia en la actualidad. Esa empresa amablemente cedió los datos que permitieron el desarrollo del capítulo 1 aquí presentado.

0.4. Capítulos de la tesis

El capítulo 1 se concentra en el test ya establecido de PGT-A y pone a prueba la hipótesis de que un panel de SNPs ubicados cada 100 Kb obtenidos por GBS permite detectar aneuploidías cumpliendo con niveles estándar de detección.

El capítulo 2 se concentra en el test novedoso de PGT-P y pone a prueba la hipótesis de que la arquitectura del genoma humano limitaría las posibilidades de la selección de embriones según su puntaje poligénico, pues esperamos que existan pares de enfermedades con riesgo correlacionado negativamente.

Páginas 11 a 64 eliminadas a pedido del autor.

Capítulo 2

Selección de embriones basada en el puntaje de riesgo poligénico

Pregunta que motiva este capítulo

¿Existen puntajes de riego poligénico (PRS) de enfermedades distintas que estén correlacionados negativamente y que, por ende, planteen un problema en la aplicación del test genético preimplantacional de riesgo poligénico (PGT-P), es decir, en la selección de embriones basada en PRS?

Respuesta resumida

Encontramos 14 pares de enfermedades con puntajes poligénicos anticorrelacionados en la población general. Individuos con PRS bajo de una enfermedad tienen riesgo absoluto aumentado de la otra. En 9 de esos pares, una simulación de embriones mostró que parejas con PRS bajo de una de las enfermedades elegirían, en media, embriones con mayor riesgo absoluto de la otra enfermedad si priorizan al embrión de PRS mínimo.

Qué se conoce

Numerosas correlaciones géneticas entre enfermedades complejas –i.e. multigénicas y multifactoriales– han sido descriptas en humanos, con métodos diversos [76-79]. La búsqueda sistemática de fenotipos correlacionados se basó, en general, en un análisis de *summary statistics* de los GWAS, pero no en los puntajes de individuos de la población o de embriones.

Datos y métodos

Utilizamos los *summary statistics* de 142 GWAS de enfermedades realizados sobre los datos de UK Biobank [80] y los genotipos de 404 individuos de ancestría europea del Proyecto 1000 Genomas [62]. Simulamos con ORIGAMI [81] embriones derivados de parejas de esos individuos. Calculamos puntajes poligénicos con el software LDAK [82]. Comparamos tres estrategias posibles de selección de embriones y su efecto en el riesgo absoluto. Desarrollamos una medida nueva de poligenicidad, $\mathcal{K}_{0.90}$, y otra de autoinmunidad genética, $\mathcal{V}_{\text{CMH}}$, para caracterizar la base poligénica de las enfermedades implicadas en los hallazgos.

Resultados principales

Encontramos 14 pares de enfermedades donde los individuos de bajo PRS de la primera enfermedad tienen riesgo absoluto incrementado de la segunda. Basándonos en simulaciones, en 9 de estos pares observamos que si en cada familia se selecciona el embrión de PRS mínimo de la primera enfermedad, los embriones seleccionados tendrán en media mayor riesgo absoluto de la segunda enfermedad, con respecto a un embrión elegido al azar. Esto se aplica, principalmente, a las familias cuyo PRS parental promedio se ubica en el primer quintil de la población.
Todos los fenotipos involucrados en los hallazgos tienen una base genética caracterizada por una fuerte asociación con el complejo mayor de histocompatibilidad (CMH) y niveles bajos de poligenicidad..

Limitaciones y precauciones

Los efectos alélicos de los puntajes poligénicos fueron ajustados en base a un dataset de referencia relativamente pequeño, de 404 individuos. Por otro lado, las asociaciones en la región del CMH exigen un cuidado especial, debido al alto desequilibrio de ligamiento en la región.

Consecuencias más amplias del hallazgo

Si bien la práctica incipiente de PGT-P es materia de intenso debate en la comunidad científica, ya existen compañías que comercializan este test para seleccionar embriones. Este capítulo caracteriza y delimita un problema posible en esta práctica hasta aquí no descripto y que podría ser evitado con la estrategia adecuada de selección de embriones.

2.1. Introducción

2.1.1. La promesa de la medicina de precisión

Durante las últimas dos décadas, el programa de la medicina de precisión ha movilizado cantidades colosales de esfuerzo científico y de presupuesto internacional, en particular en genómica humana [83]. Ejemplos de estos emprendimientos son el Proyecto Genoma Humano, concluido en 2004 [35, 84], el proyecto HapMap [85] y 1000 Genomas [62], que caracterizaron en conjunto los genomas de 2 504 individuos de 26 poblaciones, el desarrollo del biobanco del Reino Unido (UK Biobank), que recopiló datos genéticos y de salud de medio millón de personas entre 2006 y 2010 [86], la realización de numerosos estudios de asociación de genoma completo durante más de una década [87], y el anuncio en 2015 del programa *All of Us* por parte de Barack Obama en EE.UU., que enlistará a un millón de personas para registrar sus datos genéticos, fenotípicios y de estilo de vida [88].

Una de las principales promesas de la medicina de precisión consiste en la estratificación de los pacientes gracias al uso de ese tipo de datos genéticos, fenotípicos y ambientales de gran escala, con la finalidad de superar el enfoque clínico tradicional basado en "signos y síntomas" [89]. La clasificación de los pacientes en una nueva taxonomía de riesgo ayudaría a guiar las decisiones del cuidado de la salud hacia un tratamiento *a medida*, más efectivo para cada paciente particular [90].

En este contexto se introduce el desarrollo de los PRSs, que buscan cuantificar el riesgo genético o la propensión a una enfermedad que tiene un individuo, en virtud de cientos o miles de variantes genéticas presentes en su genoma. Se espera que estos puntajes permitan mejorar la prediccion del riesgo de enfermedades comunes, ya que capturan un tipo de riesgo complementario a los factores de riesgo tradicionales, como el estilo de vida [91].

2.1.2. Las enfermedades comunes son fenotipos cuantitativos

Más de una década de estudios de asociación de genoma completo (GWAS) dejaron como conclusión clara que muchas enfermedades comunes no transmisibles como el Alzheimer [92], la enfermedad de arterias coronarias [93], el cáncer de mama [94] y la diabetes [95], por nombrar algunos, son lo que se conoce en genética de poblaciones como rasgos cuantitativos del fenotipo (*quantitative traits*) (de aquí en adelante, diremos simplemente "fenotipos cuantitativos") [96].

A diferencia de las llamadas enfermedades monogénicas, donde pocos *loci* tienen un gran peso en el desarrollo de la enfermedad, los fenotipos cuantitativos tienen una base más o menos poligénica, i.e. compuesta por decenas, cientos o miles de *loci*, cada uno de ellos con un efecto individual relativamente pequeño en el valor cuantitativo final [97, 98]. Debido a sus efectos pequeños, las variantes individuales no son muy informativas a la hora de determinar el riesgo total de una enfermedad de este tipo. Por ende, se vuelve necesario combinar de algún modo esos efectos para obtener un número que resuma el riesgo del individuo. Los PRSs cumplen este rol[1].

Para entender la construcción de un PRS, primero debemos explicar cómo se estima el efecto que tiene cada marcador genético. Comentamos esto a continuación.

[1] Cuando se habla de fenotipos en general y no sólo de enfermedades, se suele decir "puntajes poligénicos". En este capítulo nos concentramos en enfermedades, de modo que diremos indistintamente PRS o puntaje poligénico.

2.1.3. Estudios de asociación de genoma completo

El GWAS es un diseño esperimental utilizado desde mediados de los 2000s para detectar asociaciones entre variantes genéticas y fenotipos sin tener una hipótesis previa sobre genes o *loci* candidatos. Este tipo de estudio se basa usualmente en la tecnología de SNP *arrays*, que permite obtener el genotipo de cientos de miles o millones de SNPs típicamente con un MAF $\geq 1\,\%$, que en este contexto suelen denominarse variantes comunes. Nada asegura que esas variantes comunes presentes en un *array* sean las causantes de la enfermedad estudidada, pero los GWAS explotan un fenómeno llamado desequilibrio de ligamiento (LD), la estructura de correlación que existe entre variantes en el genoma humano moderno como resultado de las fuerzas evolutivas que lo moldearon. Gracias al LD, cada SNP común presente en el *array* "etiqueta" a SNPs cercanos en el mismo haplotipo, con los que está correlacionado [96]. Nótese que el conjunto de SNPs a incluir en un *array* puede diseñarse a medida y generalmente se incluyen las variantes descubiertas en GWAS anteriores. Esto no garantiza una representación de ancestrías diferentes de la europea.

Los fenotipos de enfermedad que analizamos en este capítulo son binarios, con individuos clasificados como enfermos o sanos (casos y controles en la terminología de GWAS). El test de asociación para fenotipos binarios puede realizarse con diferentes modelos, pero independientemente del modelo elegido el resultado para cada SNP consiste de dos valores: uno para caracterizar la magnitud del efecto y otro que representa la confianza estadística en la asociación. Estos dos valores se expresan, por un lado, como el *odds ratio* (OR) –o como un coeficiente β que equivale al log(OR)– y, por otro lado, como el P valor de la asociación. En este capítulo nos referimos a β como la magnitud del efecto de un alelo en el riesgo de la enfermedad. Dado que está en escala logarítmica, $\beta < 0$ representa una reducción del riesgo, $\beta > 0$ representa un aumento del riesgo y $\beta = 0$ significa que el efecto del SNP en la enfermedad es nulo. Se ilustra el test de un SNP en la **Figura** 2.1.

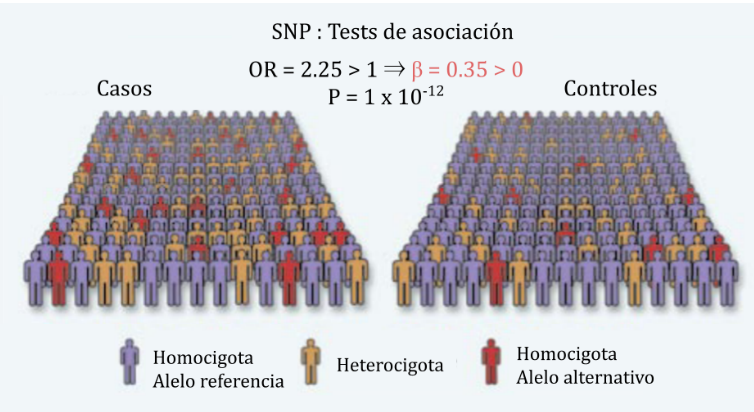


Figura 2.1: Ejemplo de un test de asociación para un único SNP, en un GWAS. En el ejemplo se observa que hay más individuos heterocigotas y homocigotas alternativos en el grupo de casos. El resultado es un $\beta > 0$ (que viene de un OR > 1) y un P valor de la asociación. Modificado a partir de [99].

Se denomina *summary statistics* al resultado de los n tests independientes de asociación genética a un fenotipo (un test por cada SNP). Los *summary statistics* consisten de los n coeficientes $\{\beta_1, \ldots, \beta_n\}$, es decir el efecto de cada SNP en el fenotipo, junto al P valor de cada test.

Existe un problema conocido desde hace años con los GWAS: la falta de diversidad de ancestría en sus participantes [100, 101]. Un estudio reciente muestra que los GWAS entre 2007 y 2017, realizados para 3 508 fenotipos, tuvieron una gran mayoría de participantes de ancestría europea (86 % en promedio), con sólo un 2 % de latinoamericanos [102]. Se resume este problema en los promedios del período completo graficados en la **Figura** 2.2.

La falta de diversidad de los GWAS es problemática porque se ha observado que alrededor del 25 % de las variantes identificadas como asociadas al fenotipo en GWAS europeos tienen efectos de tamaño significativamente diferente en al menos una población no europea, si bien el signo de las asociaciones no cambia [103]. Con todo, se ha observado que al menos las variantes causales subyacentes sí están compartidas en la mayoría de los casos [104].

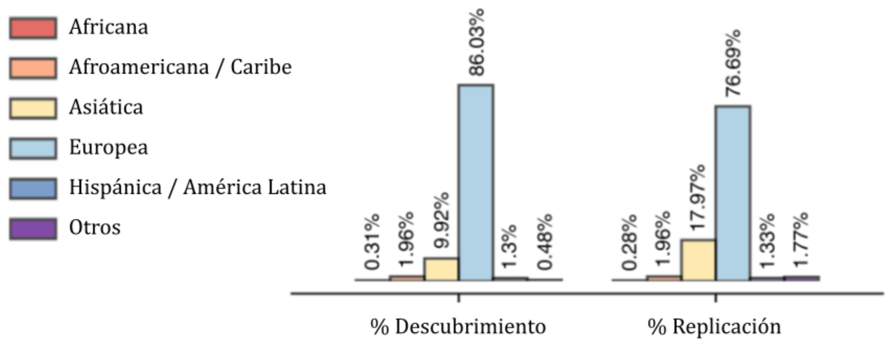


Figura 2.2: Ausencia casi total de latinoamericanos y predominancia de europeos en los participantes de GWAS, período 2007-2017. "Descubrimiento" se refiere al GWAS inicial de detección de SNPs asociados. "Replicación" se refiere a los GWAS de *follow-up* en los que se confirman hallazgos previos. Tomado de [102].

2.1.4. Puntajes de riesgo poligénico y PGT-P

Como dijimos, un puntaje de riesgo poligénico busca resumir, para un individuo dado, el riesgo genético total de una enfermedad. Se suele asumir que el fenotipo tiene una arquitectura genética aditiva, es decir, que los efectos genéticos debido a la interacción entre variantes son mínimos, pues se ha mostrado que este modelo coincide con datos de recurrencia observado en parientes [105].

En su forma más básica, el PRS se computa como una suma ponderada de los alelos de riesgo presentes en n variantes de un genoma,

$$\sum_{i=1}^{n} \beta_i G_i \tag{2.1}$$

donde la dosis alélica $G_i \in \{0, 1, 2\}$ es el número de alelos alternativos (i.e. distintos del genoma humano de referencia) que tiene un individuo en la variante i-ésima y β_i es la magnitud del efecto que tienen esos alelos en el riesgo. El valor de β_i, en primera instancia, proviene de las *summary statistics* de un GWAS. De aquí en adelante, asumiremos que las variantes que componen al PRS son SNPs bialélicos, de modo que hablaremos indistintamente de SNPs y variantes. Se ilustra esquemáticamente este cálculo en la **Figura** 2.3.

Figura 2.3: Ejemplo esquemático de la suma ponderada de genotipos $\sum \beta_i G_i$ que conforman al PRS de un individuo. A y B refieren al alelo de referencia y alternativo, respectivamente, en un SNP bialélico.

La cantidad de puntajes poligénicos asociados a diversos fenotipos crece sin interrupción desde hace varios años: en mayo de 2022, el PGS Catalog ya registra 2 196 puntajes asociados a 524 fenotipos, que se reparten en 18 categorías como cáncer, cardiopatías, enfermedades autoinmunes, desórdenes metabólicos, desórdenes neurológicos y respuesta a drogas. Junto a este creciente catálogo de PRSs, se da una fluida conversación en la literatura acerca de su utilidad clínica [98, 106–108]. Los costos decrecientes de genotipificación y secuenciación, la relativa sencillez del cálculo y el potencial de segmentar la población para asignar tratamientos diferenciales anticipándose al desarrollo de la enfermedad hacen del PRS una herramienta atractiva para la medicina genómica.

Selección y ajuste de los β_i

La pregunta fundamental en la construcción de un PRS es qué SNPs incluir en el cálculo. El primer método utilizado para esta decisión fue el de *pruning and thresholding* (P+T), que conserva únicamente variantes en equilibrio de ligamiento (i.e. no correlacionadas entre sí), para luego aplicar umbrales de significancia progresivamente más laxos, lo que implica conjuntos de SNPs cada vez mayores. Con esta técnica se mostró tempranamente que incluir SNPs no significativos mejora el poder predictivo del PRS [109].

Trabajos posteriores encontraron que el método P+T construye puntajes poligénicos con un poder predictivo menor al teóricamente posible [110] . Así, surgieron métodos para corregir los β_i de los GWAS contemplando el efecto del LD, que lograron incrementos del r^2 entre puntajes y fenotipos [82, 110–112]. Muchos de estos métodos, además, tienen la ventaja de basar la corrección únicamente en las *summary statistics* de los GWAS y en un panel genético de referencia, datos a los que se accede sin dificultad, en contraste con los datos originales del GWAS (genotipos y fenotipos de los participantes), cuyo acceso suele estar regulado muy estrictamente.

El método LDAK-BayesR-SS

Zhang y col. [82] comparan cuatro métodos de ajuste de los β_i basadas en *summary statistics*: lassosum [111], sBLUP [113], LDpred [110] y SBayesR [114], reimplementados en su software LDAK. En esa comparación, concluyen que el modelo de SBayesR (en su reimplementación denominado LDAK-BayesR-SS) supera a los otros en poder predictivo en 223 de 225 fenotipos analizados. Si además se asume el modelo de heredabilidad que denominan BLD-LDAK, se obtiene una nueva mejora en la correlación entre puntaje y fenotipo. La combinación de LDAK-BayesR-SS y BLD-LDAK es por ende recomendada por los autores. Exponemos en la sección A.1.1 con más detalle tanto el método de ajuste de los β_i como el modelo de heredabilidad.

La rutina completa de ajuste de los β_i en el software LDAK es denominada por sus autores MegaPRS. Detallamos los pasos y parámetros de MegaPRS utilizados en este capítulo en la sección A.1.2.

PRS en embriones

Como mencionamos en la introducción general, hace pocos años se discuten las posibilidades y límites del PGT-P, un test altamente controversial que consiste en utilizar los PRSs para guiar la selección de embriones preimplantatorios. Al momento de escribir esta introducción, sabemos de dos compañías que comercializan el PGT-P: LifeView y Orchid, mientras que Turley y col. [11] mencionan otras dos: MyOme y Reprocare Genetics.

Lifeview y Orchid ofrecen una selección de embriones de FIV basada en el riesgo de enfermedades como la diabetes, varios tipos de cáncer, cardiopatías, hipercolesterolemia, hipertensión, esquizofrencia, discapacidad intelectual y Alzheimer, entre otras, como se ve en la **Figura** 2.4. MyOme, según reporta [11], "pareciera proveer" PRSs del nivel educativo, *status* económico, habilidad cognitiva y "sensación de bienestar" (*subjective well-being*) como parte de un protocolo de investigación, aunque no pudimos confirmar esto porque la URL citada como fuente en el artículo ya no es accesible.

El PGT-P es propuesto como un test complementario al PGT-A, que aporta un criterio para guiar la selección entre los embriones euploides. La utilidad clínica de esta técnica ha sido muy discutida, pero escapa a los objetivos de este capítulo recapitular este debate de forma exhaustiva. Ofrecemos un resumen de algunos de sus puntos centrales en el **Recuadro** 2.1.

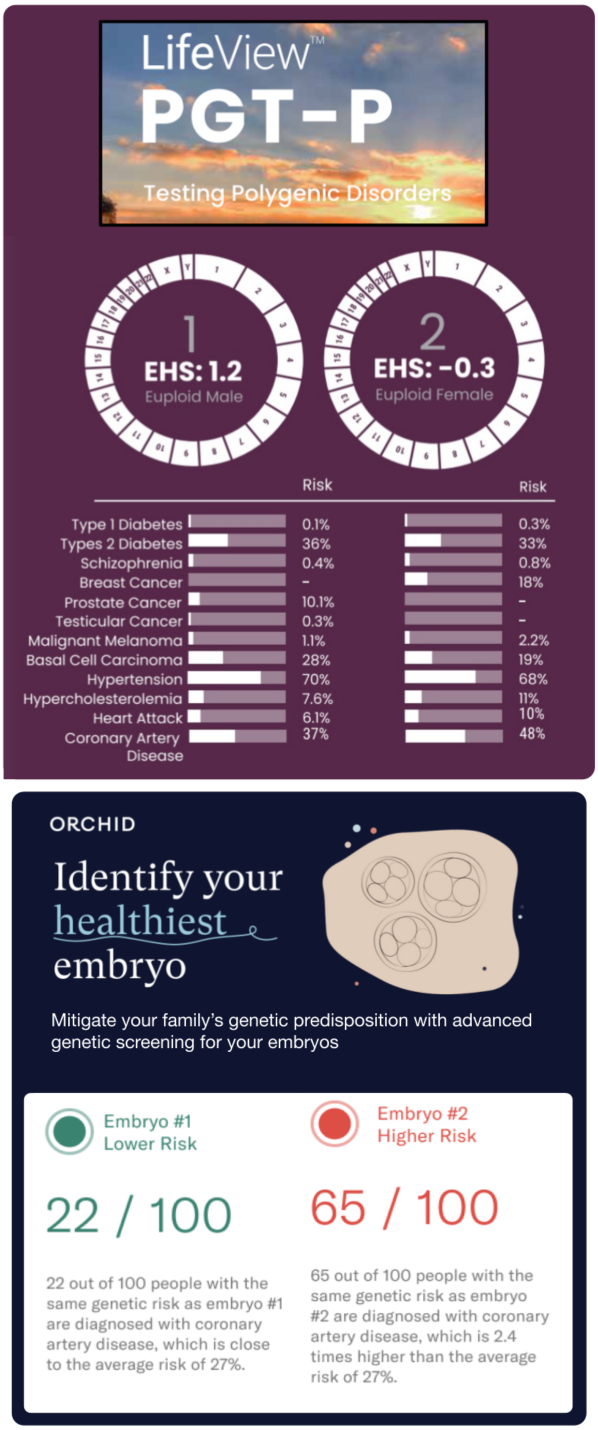


Figura 2.4: Reportes de ejemplo de PGT-P tomados de las webs de LifeView y Orchid (capturas de enero de 2022). **Izquierda**: ejemplo ofrecido por LifeView en su brochure de PGT-P. Se reporta el riesgo de 12 enfermedades comunes en dos embriones y se calcula un Embryo Health Score (EHS) para cada uno. **Derecha**: ejemplo ofrecido por Orchid en sus guías online. Se ven dos embriones con riesgo genético diferente de enfermedad de arterias coronarias.

La controversia del PGT-P **Recuadro 2.1**

La utilidad clínica de los PRSs en la población adulta es materia de debate. Por ejemplo, según la reseña reciente de Torkamani y col. [115], numerosos estudios han mostrado que el PRS de enfermedad de arterias coronarias es útil para identificar individuos en alto riesgo –con riesgo relativos de hasta 400 %–, que se beneficiarían con terapia de estatina, mientras que otros estudios han mostrado que los PRSs de cáncer de mama y de colon permitirían tomar mejores decisiones sobre la edad en la que se deberían iniciar chequeos de rutina en cada individuo. En contraste, un metanálisis reciente [108] afirma que no hay en la literatura *un solo estudio* que demuestre la utilidad clínica del PRS, aunque tal vez sí la validez de los puntajes en ciertas condiciones.

La utilización de los puntajes poligénicos para seleccionar embriones de FIV añade una nueva arista a la discusión: el temor a las aplicaciones eugenésicas, que ocupan rápidamente los títulos de los artículos de divulgación, de la prensa escrita y de los medios de comunicación [116, 117]. Los alegatos en contra del PGT-P han sido recopilados recientemente en un artículo de Turley y col. [11]. Resumimos aquí sus principales puntos, acompañados de algunas interpretaciones propias.

La primera línea argumental se enfoca en la baja ganancia esperada (*expected gain*) debida a la reducida variabilidad genética que existe entre embriones de una misma pareja, un argumento ya expuesto en [10] con un análisis de la altura y el coeficiente intelectual y repetido en [11] con el *educational attainment*. El argumento fue contestado en [8]: es más ético y práctico prevenir enfermedades que aumentar valores de algún fenotipo considerado “positivo”. En [118] se ha mostrado, además, que el PGT-P reduciría el riesgo de numerosas enfermedades incluso cuando la selección se realiza sólo entre dos embriones euploides –i.e. con muy poca variabilidad disponible–, independientemente de la presencia o ausencia de la afección en la familia.

Un segundo problema señalado es la pérdida de poder predictivo del PRS debida a diversos factores. Por un lado, la correlación genes-ambiente infla el poder predictivo del PRS cuando se consideran individuos de diferentes familias, pero esta porción de poder predictivo se pierde al comparar embriones de los mismos padres. Por otro lado, los embriones bajo selección tendrán ambientes necesariamente diferentes que los participantes del GWAS o que la cohorte con la que se validen los puntajes. Este es el alegato más fuerte, a nuestro entender, aunque esa pérdida de poder predictivo sólo parece haber sido cuantificada en el caso de cambio de ancestría [119].

Otros problemas mencionados se resumen en la dificultad de comunicar la naturaleza probabilística del puntaje de riesgo y la incertidumbre consiguiente en el resultado, en parte debido a las ideas erróneas que el paciente podría tener sobre el “determinismo genético”. Tal preocupación, creemos, es difícilmente un argumento en contra de una práctica médica, pero sí un aspecto importante a examinar, donde el rol del médico genetista capacitado en la comunicación de este tipo de datos cobra cada vez más relevancia.

La pleiotropía es una preocupación adicional mencionada en el artículo [11]: el PRS asociado a un fenotipo podría tener consecuencias inesperadas en otro fenotipo. El trabajo de [120] contesta en parte a esta preocupación, notando que los SNPs predictores de un buen número de enfermedades de interés conforman mayoritariamente conjuntos disjuntos. Por otro lado, este capítulo circunscribe de manera constructiva el problema: como veremos, en algunas efermedades autoinmunes los PRSs sí estarían negativamente correlacionados en un sentido relevante para la clínica, lo que nos da una lista de fenotipos con los que ser cautos en PGT-P. En el artículo de Turley y col., sin embargo, el planteo es pesimista antes que programático: el énfasis está en que nunca conoceremos todas las relaciones entre variantes genéticas.

Es reveladora la continuación de esta línea argumental: tras poner en duda el poder predictivo del PRS en pasos anteriores, los autores admiten que, gracias a avances técnicos, los puntajes probablemente mejoren su rendimiento en el futuro. Sin embargo, arguyen, debido a la pleiotropía esto también sería negativo, pues con mejores PRSs *la magnitud de las consecuencias inesperadas* también crecería. Combinando ambos razonamientos, podemos resumir: el PRS es malo si predice poco y es malo si predice mucho.

Otras preocupaciones acucian y las compartimos, como la peligrosa exacerbación de disparidades basadas en el estatus socioeconómico. Consideramos, como los autores, que la solución a este tipo de problemáticas no puede dejarse en manos de una empresa de tests genéticos y que es necesario un desarrollo de protocolos y guías de aplicación del PGT-P por parte de la comunidad científica y médica. Agreguemos que, en última instancia, la lista de fenotipos que sería “aceptable” seleccionar en embriones no es un asunto que se resuelva empíricamente: dependerá de lo que cada sociedad considere ético cuando le llegue el momento de legislar sobre esta materia, si es que esto sucede.

La discusión es vasta, tanto en sus aspectos puramente científicos como en sus aristas bioéticas. Es aparente que, frente a los problemas y desafíos reales que los PRSs deben superar para tener utilidad clínica, algunos autores se retraen con pesimismo hacia una oposición agresiva, de tintes morales, algo que ya describieron Schulman y Edwards en un artículo de 1996 sobre PGD en el que se discutía, como en el presente, si diagnosticar desórdenes genéticos en embriones es eugenésico [121]. Tras mencionar que las tecnologías de diagnóstico preimplantacional “expanden la libertad de padres potenciales para reducir la inmensa carga de la enfermedad genética en sus familias”, los autores concluían que “en contraste con las predicciones catastróficas, la experiencia de aplicaciones de tecnología genética y reproductiva ha tenido en el pasado consecuencias enormemente positivas para la humanidad”. Más de 20 años después, a la luz de inmensos avances en tecnologías de PGT hoy aceptadas en la comunidad médica, renovar el optimismo de la cita no sería del todo infundado.

2.1.5. Estrategias de selección de embriones por PRS

Lencz y col. [122] caracterizan dos posibles estrategias de selección de embriones basadas en PRS. La primera, denominada *high-risk exclusion*, consiste en no transferir embriones de alto puntaje poligénico. En su trabajo, esto consiste en trazar un umbral que defina qué es un PRS alto. Análogamente, nosotros hablaremos de Excluir alto riesgo (EAR) para referirnos a una estrategia parecida, pero algo más agresiva: dados n embriones en una pareja, elegir un embrión al azar entre los $n - 1$ embriones cuando se excluye el de máximo PRS. La segunda estrategia es denominada por los autores *lowest-risk-prioritization* y nosotros la llamaremos Priorizar bajo riesgo (PBR): consiste en transferir el embrión de puntaje poligénico *mínimo* entre los disponibles. En ambos casos, asumimos que no existen "empates" de PRS: el mínimo y el máximo son siempre únicos. Se ilustran estas dos estrategias en la **Figura** 2.5.

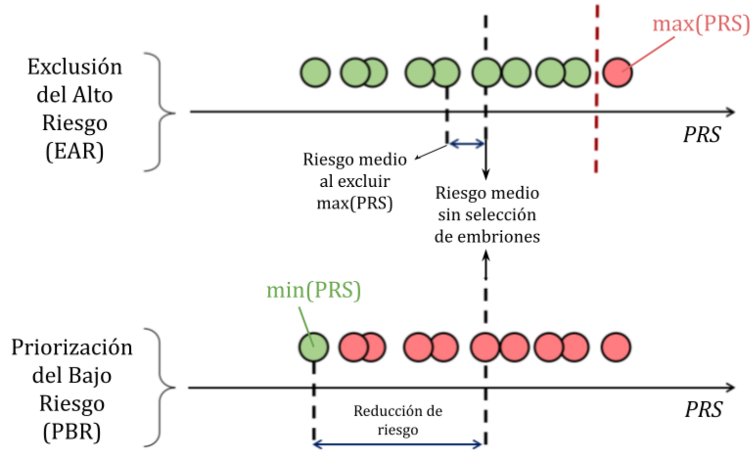


Figura 2.5: Los círculos representan embriones con diferentes PRSs calculados para un fenotipo particular. Se ilustran dos estrategias de selección de embriones en base al puntaje poligénico. En verde: los embriones "seleccionables" según la estrategia; en rojo, los que no se transferirían. Adaptado de [122].

En [122] concluyen que la estrategia de *high-risk exclusion* ofrece reducciones modestas del riesgo relativo, mientras que la estrategia de elegir el embrión de PRS mínimo sí puede resultar en reducciones mayores. Esto se debe a que la distribución del riesgo en los embriones cambia poco al excluir sólo a uno de ellos, mientras que la distribución del PRS mínimo es muy diferente a la de valores de PRS tomados al azar entre los embriones.

2.1.6. Pleiotropía y correlación entre puntajes poligénicos

La pleiotropía es un fenómeno en el que el mismo gen, *locus* o SNP tiene efectos fenotípicos diferentes y en apariencia no relacionados entre sí. Se ha estimado en algunas especies modelo que el grado de pleiotropía tiene una distribución en forma de L: la mayoría de los genes afectan pocos fenotipos, pero existe una cola que decae exponencialmente de pocos genes que afectan a muchos fenotipos [123, 124]. Se ha descripto también la pleiotropía en bacterias, con diversas hipótesis de su papel en la evolución de la arquitectura genética [125].

En humanos, se ha visto que la pleiotropía es un fenómeno común en numerosos fenotipos [79, 126, 127]. La pleiotropía es denominada "horizontal", cuando una misma variante genética tiene efectos directos sobre dos fenotipos diferentes, o efectos indirectos en dos fenotipos a través de un endofenotipo único, o "vertical", cuando el fenotipo afectado tiene un relación causal con un segundo fenotipo [128].

Se ha observado que ciertos grupos de fenotipos tienden a co-variar entre individuos de la misma especie o incluso entre distintas especies, conformando "módulos de variación" [129]. En este contexto, un módulo está conformado por un grupo de fenotipos que es afectado en conjunto por un grupo correspondiente de genes. Este fenómeno, denominado pleiotropía modular, fue descripto en levaduras, nematodos y ratones [124]. La pleiotropía modular puede representarse mediante una red bipartita de genes y fenotipos, donde los genes de un módulo raramente afectan a fenotipos de otro módulo, como se ilustra en **Figura** 2.6.

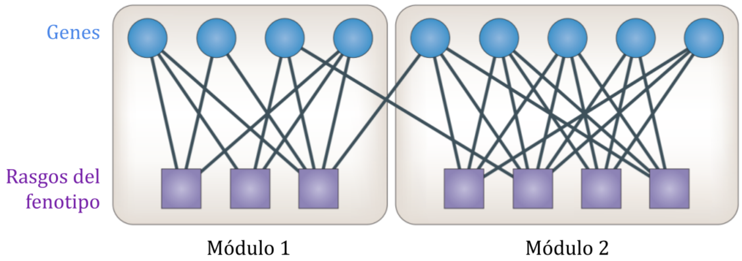


Figura 2.6: Red bipartita altamente modular de genes (círculos) y fenotipos (cuadrados). Las aristas representan que un gen afecta al fenotipo. Tomado de [129].

Un fenómeno relacionado a la pleiotropía es el de correlación genética, que describe el efecto *promedio* de la pleiotropía a nivel local, es decir, cómo varían dos puntajes genéticos en virtud de sus variantes asociadas en común y los efectos de esas variantes en cada fenotipo. Nótese que dos PRSs que compartan SNPs pleiotrópicos podrían no estar correlacionados. Esto ocurre porque regiones con pleiotropía de diferente signo podrían compensarse, como se ve en el ejemplo **D** de la **Figura** 2.7. Por otro lado, es interesante notar que un valor dado de correlación genética entre fenotipos puede resultar de arquitecturas de correlaciones locales muy diferentes, como se ve en los ejemplos **A**, **B** y **C** **Figura** 2.7 **B** [128].

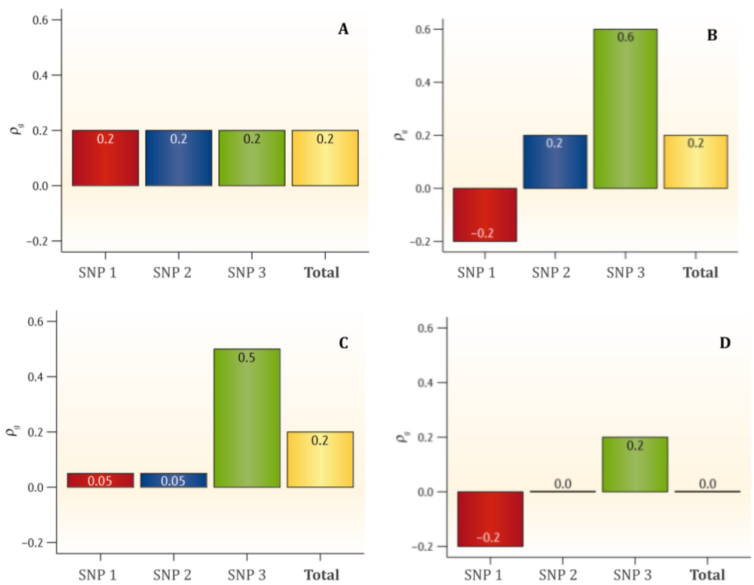


Figura 2.7: Ejemplos de correlación genética entre fenotipos. La correlación genética total (columna amarilla) resume en cada caso a la pleiotropía o correlación local de tres SNPs (columnas roja, azul y verde), expresada como valores de ρ_g en el eje Y. En **A**, **B** y **C**, la correlación total es 0.2, pero las correlaciones locales varían. En **D**, la correlación total es 0 a pesar de que existen correlaciones locales no nulas. Tomado de [128].

Así pues, la correlación entre fenotipos se basa en la pleiotropía, pero podemos decir que es una condición aun más fuerte, pues implica que *la dirección* (el signo) de los efectos también debe estar consistentemente alineado [77].

Varios trabajos han mostrado que existen numerosos pares de fenotipos complejos correlacionados, tanto positiva como negativamente [77, 78, 130, 131]. En [128] se da un extenso listado de métodos de estimación de la correlación, ya sea basados en *summary statistics* o en datos de individuos. Existe además un catálogo exhaustivo de correlaciones genéticas entre fenotipos de UK Biobank, basadas en la regresión con LD *score* [132].

2.1.7. El complejo mayor de histocompatibilidad

Un caso particular en donde la pleiotropía tiene efectos conocidos es en el complejo mayor de histocompatibilidad (CMH), una región del genoma de los vertebrados también conocida en humanos como HLA (*human leukocyte antigen*). El CMH alberga varios clusters de genes que codifican proteínas de superficie implicadas en la respuesta inmune adaptativa. Se trata del *locus* más polimórfico del genoma humano y se caracteriza por altos niveles de desequilibrio de ligamiento [133].

Se ha observado que numerosas enfermedades autoinmunes comparten SNPs asociados en esta región, lo que sugeriría un mecanismo común subyacente [76]. No es seguro, sin embargo, que los SNPs en común impliquen verdadera pleiotropía, como se afirma en [134].

2.2. Objetivos

El objetivo general de este capítulo es entender si el fenómeno común de la pleiotropía en el genoma humano y la correlación genética resultante entre fenotipos representan un problema para la aplicación del PGT-P. En particular, analizamos si existen pares de enfermedades cuyos puntajes poligénicos estén correlacionados negativamente de tal manera que, al elegir embriones de bajo PRS de una enfermedad, se incremente el riesgo de otra enfermedad. Para esto, el primer objetivo particular es encontrar pares de enfermedades en las que, al seleccionar individuos de bajo PRS en la población general para la primera enfermedad, se observe en ese subgrupo un incremento en el riesgo de la segunda enfermedad. El segundo objetivo es simular embriones y comparar las estrategias EAR y PBR, para entender si el problema observado en la población general persiste, en esas enfermedades, al seleccionar embriones según su PRS.

2.3. Datos, Métodos y Simulaciones

2.3.1. UKB-GWAS

Utilizamos las *summary statistics* de 142 enfermedades denominadas *self-reported* (de aquí en adelante, "auto-reportadas") de GWAS realizados por el Neale Lab [80], basados en los datos de genotipos y fenotipos de la población del UK Biobank [86]. Llamaremos UKB-GWAS a este dataset.

Elegimos las enfermedades auto-reportadas porque dentro de los fenotipos disponibles con GWAS de UK Biobank, se trata de los estudios con mayor número de casos. En contraste, algunos fenotipos de enfermedades "diagnosticadas por un médico" (*doctor-diagnosed*) tienen un número de casos muy bajo. De las enfermedades auto-reportadas, conservamos únicamente las que tienen valores de heredabilidad y prevalencia disponibles en el catálogo de [135]. Excluimos también fenotipos que consideramos poco interesantes: 25 dentro de la categoría *fracture* y 2 descriptos como *unclassifiable*.

2.3.2. Genotipos de 1KGP

Utilizamos los genotipos del Proyecto 1000 Genomas (1KGP) fase 3 [62] en formato `pgen` del sitio web de PLINK2 [136]. Filtramos con PLINK2 un total de 404 individuos de cuatro poblaciones de ancestría europea (EUR): 91 británicos de Inglaterra y Escocia (GBR), 107 individuos de poblaciones ibéricas de España (IBS), 107 toscanos de Italia (TSI) y 99 residentes de Utah con ancestría del norte y oeste de Europa (CEU). Finalmente, construimos el panel de referencia en formato `{bed,bim,fam}` de PLINK con las instrucciones de la guía de MegaPRS.

2.3.3. Control de calidad de los SNPs

Filtramos las variantes del dataset UKB-GWAS, conservando únicamente SNPs bialélicos no ambiguos[2] de MAF < 0.01, con un *call rate* mayor a 95 %, un puntaje de imputación (INFO *score*) mayor a 0.80 en el caso de las variantes imputadas, y que no se desvíen significativamente del equilibrio Hardy-Weinberg. Este dataset tiene variantes identificadas por posición genómica, mientras que en el dataset de 1KGP las variantes están identificadas por su RefSeq ID. Por ende, generamos un archivo en format `bed` con las posiciones por cromosoma, en base al cual extrajimos los genotipos de los 404 individuos con PLINK. Al dataset de genotipos resultantes le aplicamos nuevos filtros: conservamos todas las variantes con más de 95 % de *call rate* en 1KGP y nos aseguramos que todos los individuos tienen más del 90 % de los SPNs presentes. El resultado final fue un conjunto de 6 708 984 SNPs en 404 individuos. Llamaremos de aquí en adelante a este dataset como 1KGP-EUR.

[2] Llamamos no ambiguo a un SNP bialélico cuando sus dos alelos no son A/T ni C/G. Por ejemplo, un SNP con alelos A/G no es ambiguo.

2.3.4. Efecto de los SNPs y cálculo del PRS

Distinguimos aquí dos procedimientos diferentes. Por un lado, hablamos de la *construcción* del PRS de un fenotipo dado para referirnos a la *elección del efecto de cada SNP*, es decir, la elección y ajuste de los β_i a utilizar en $\sum \beta_i G_i$. Esto usualmente implica una subselección de SNPs, que serán aquellos con $\beta_i \neq 0$. Por otro lado, hablamos del *cálculo* del PRS de un fenotipo dado para referirnos a la obtención de un valor particular de PRS *correspondiente a un individuo*, con los β_i elegidos en el paso previo y los G_i correspondientes a los genotipos del individuo.

Durante los últimos años, se han desarrollado métodos para construir el PRS a partir únicamente de los *summary statistics* del GWAS, sin la necesidad de un dataset de genotipos y fenotipos de *testing* en los que minimizar el error de predicción. Entre estos métodos, la rutina MegaPRS implementada en LDAK puede generar un conjunto óptimo de β_i –es decir, construye un PRS– tras comparar la performance predictiva de numerosos modelos en competencia, a partir de los *summary statistics* de un GWAS [82].

La posibilidad de optimizar un PRS sólo a partir de *summary statistics*, la extensa documentación online y la robustez del software nos inclinaron por la elección de LDAK. Resumimos en sección A.1.2 la rutina de MegaPRS utilizada, en primer lugar, para la construcción de los PRSs de 142 fenotipos de UKB-GWAS y luego para el cálculo de puntajes en los 404 individuos de 1KGP-EUR.

2.3.5. Conversión del PRS al riesgo absoluto

La conversión del PRS de un individuo al riesgo de adquirir una enfermedad requiere un cálculo adicional con ciertos supuestos y parámetros, que describimos en esta sección.

Nos basamos en el modelo de umbral de propensión (*liability threshold model*) [137] y partes del desarrollo de Lencz y col. [122] para transformar los valores de PRS en la probabilidad que tiene un genoma particular de adquirir una enfermedad a lo largo de su vida. Este modelo clásico, que se sigue utilizando para el análisis de datos modernos[3], es relativamente simple pues depende sólo de dos parámetros: la heredabilidad, que definiremos en breve, y el riesgo de la enfermedad en la población [139].

[3] Su adecuación a datos de parientes ha sido mostrada en [105], bajo el nombre de *Probit model*. Un ejemplo de uso es [138].

Comencemos por recordar que los fenotipos que analizamos en este capítulo son binarios, que dividen a la población en dos clases: enfermos y sanos. El modelo de umbral de propensión propone un fenotipo continuo L subyacente, que en el contexto de enfermedades humanas se suele describir como la susceptibilidad a la enfermedad (*liability*). En este modelo, los individuos enferman cuando su valor particular de L supera cierto umbral ℓ a determinar.

Se suele modelar a L como una variable de distribución normal y se la descompone como una suma de dos componentes independientes: el genético y el ambiental[4]. Aquí limitaremos el componente genético al puntaje poligénico, que llamaremos S (por *score*) en las ecuaciones. Todos los otros factores que influyen en el fenotipo –efectos genéticos no incluidos en el PRS y efectos ambientales– son agrupados en una variable de error E. Tanto S como E se distribuyen normalmente y son independientes. El modelo se resume como la suma:

[4] El supuesto de normalidad asume que la propensión es multifactorial, es decir, que es afectada por un gran número de factores genéticos y ambientales, con efectos relativamente pequeños [137, 139].

$$L = S + E \tag{2.2}$$

Ahora bien, el valor S de (2.2) no es excatamente el puntaje que obtenemos de LDAK. Llamemos S^* a los puntajes obtenidos con este software. LDAK calcula los puntajes centrándolos en cero a partir del panel de referencia, es decir que su media es conocida: $\mathbb{E}[S^*] = 0$. No conocemos de antemano su varianza, de modo que la estimamos a partir de la muestra de $n = 404$ individuos para los que calculamos el PRS: $\sigma^2_{S^*} \approx \frac{1}{n}\sum_{i=1}^{n} S_i^{*2}$. Tenemos pues que $S^* \sim \mathcal{N}\left(0, \sigma^2_{S^*}\right)$.

Definimos S como el puntaje reescalado de la siguiente manera:

$$S := S^* \cdot \frac{h_g}{\sigma_{S^*}} \tag{2.3}$$

donde h_g es la raíz cuadrada de h_g^2, un parámetro poblacional dependiente del fenotipo, que definiremos en breve. La distribución del puntaje reescalado es:

$$\mathbb{E}[S] = \mathbb{E}\left[S^* \cdot \frac{h_g}{\sigma_{S^*}}\right] = \frac{h_g}{\sigma_{S^*}}\mathbb{E}[S^*] = 0$$

$$\mathrm{var}(S) = \mathrm{var}\left(S^* \cdot \frac{h_g}{\sigma_{S^*}}\right) = \frac{h_g^2}{\cancel{\sigma^2_{S^*}}}\cancel{\mathrm{var}(S^*)} = h_g^2$$

$$S \sim \mathcal{N}\left(0, h_g^2\right)$$

Por otro lado, modelamos al componente ambiental como una normal cuya varianza captura *el resto* de la varianza del fenotipo. Dado que queremos que la varianza total sea 1, usamos $\mathrm{var}(E) = 1 - h_g^2$.

$$E \sim \mathcal{N}\left(0, 1 - h_g^2\right)$$

Con S y E definidas de este modo y recordando que S es independiente de E, obtenemos dos resultados deseables. Por un lado, la propensión L se distribuye como una normal estándar, lo que facilita algunos cálculos:

$$\mathbb{E}[L] = \mathbb{E}[S + E] = \mathbb{E}[S] + \mathbb{E}[E] = 0$$

$$\mathrm{var}(L) = \mathrm{var}(S + E) = \mathrm{var}(S) + \mathrm{var}(E) = h_g^2 + \left(1 - h_g^2\right) = 1$$

$$L \sim \mathcal{N}(0, 1)$$

Por otro lado, la proporción de la varianza total capturada por el PRS es el valor h_g^2.

$$\frac{\mathrm{var}(S)}{\mathrm{var}(L)} = h_g^2$$

Con esta definición, h_g^2 representa el concepto de heredabilidad limitado a los SNPs que entran en el puntaje poligénico, conocido como *SNP heritability* o h^2_{SNP}. El dato de h_g^2 es poblacional, pues tanto la varianza del componente genético como la varianza total del fenotipo pueden cambiar entre poblaciones.

Para completar el modelo, necesitamos definir el umbral ℓ que separa a los enfermos de los sanos. Para cada fenotipo, conocemos la probabilidad de enfermar que tiene un individuo sólo por pertenecer a una población: es el valor K de prevalencia, la proporción de individuos afectados. Tomamos este dato del catálogo de [135]. Luego, elegimos el umbral ℓ que cumple $\Pr\{L > \ell\} = K$. Dado que $L \sim \mathcal{N}(0,1)$, este valor es z_K, el cuantil $1-K$ de la normal estándar. Se ilustra el modelo en la **Figura** 2.8.

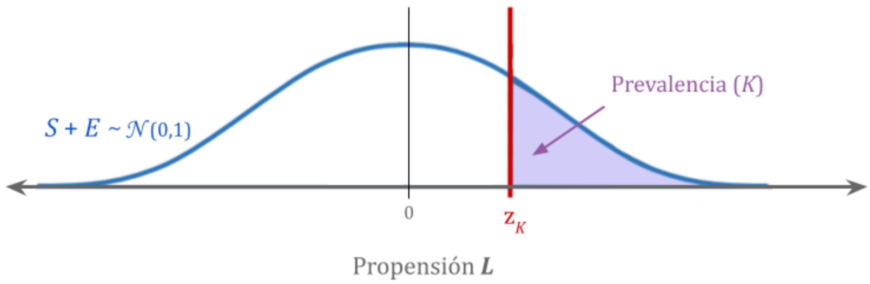


Figura 2.8: En azul, la densidad de la propensión $L = S + E$, con una distribución normal estándar. En rojo, el umbral que determina la enfermedad. En lila, el área bajo la curva que equivale a la prevalencia K.

El área bajo la curva ilustrada en la **Figura** 2.8 representa la probabilidad de enfermar *para un individuo cualquiera* de la población, en virtud de la variación tanto de los genomas como de los ambientes posibles. Sin embargo, si conocemos el genoma de un individuo particular, podemos calcular su PRS reescalado con (2.3) y computar la probabilidad de que *ese individuo* enferme como una probabilidad condicional. De aquí en adelante, llamaremos riesgo absoluto (RA) a este valor:

$$\begin{aligned}
\mathrm{RA} := \Pr\{\text{enfermedad} \,|\, S = s\} &= \Pr\{L > z_K \,|\, S = s\} \\
&= \Pr\{S + E > z_K \,|\, S = s\} \\
&= \Pr\{s + E > z_K\} \\
&= \Pr\{E > z_k - s\} \\
&= \Pr\left\{\frac{E}{\sqrt{1-h_g^2}} > \frac{z_K - s}{\sqrt{1-h_g^2}}\right\} \\
&= \Pr\left\{Z > \frac{z_K - s}{\sqrt{1-h_g^2}}\right\} \\
&= 1 - \Pr\left\{Z \leq \frac{z_K - s}{\sqrt{1-h_g^2}}\right\} \\
&= 1 - \Phi\left(\frac{z_K - s}{\sqrt{1-h_g^2}}\right)
\end{aligned} \tag{2.4}$$

donde $s \in \mathbb{R}$ es un valor particular de PRS, Z es una v. a. normal estándar, Φ es la función de distribución acumulada de la normal estándar y z_K es el cuantil $1-K$ de esta distribución. Como mencionamos, tanto la heredabilidad h_g^2 como la prevalencia K son parámetros poblacionales, específicos de cada fenotipo, cuyos valores empíricos tomamos de [135].

Reforcemos la intuición en este punto. Un genoma particular determina un valor particular de PRS, lo que señalamos como $S = s$. Dado que no conocemos de antemano los ambientes a los que ese genoma estará expuesto, su propensión a la enfermedad es una variable aletoria. Aquí, la aleatoriedad proviene enteramente de E, ya que el puntaje poligénico está fijado. La distribución de la propensión *dado un genoma* es

$$L_{|S=s} = E + s \sim \mathcal{N}\left(s, 1 - h^2\right) \tag{2.5}$$

donde $L_{|S=s}$ significa L condicionada al valor particular de $S = s$. Como se ve, la propensión condicionada a s se centra en s.

Así, un genoma cuyo PRS para la enfermedad sea menor a la media poblacional, es decir $s < 0$, implica un "buen comienzo genético", pues aleja a la distribución de $L_{|S=s}$ del umbral z_K, de modo que incluso ambientes muy perjudiciales podrían no resultar en la enfermedad. Por el contrario, un genoma con un PRS alto acerca la distribución condicional al umbral, de modo que sólo ambientes muy favorables podrían prevenirla. El peso relativo del componente genético y del ambiental en este balance se cifra en el valor h_g^2, que varía según el fenotipo en cada población.

En la **Figura** 2.9 ilustramos esta idea de buen o mal comienzo genético con tres genomas distintos, que determinan tres puntajes poligénicos diferentes. El valor más bajo, s_1, implica que la distribución condicional L_1 se aleja del umbral z_K y resulta en un riesgo absoluto reducido en ese individuo, respecto de la población ($\mathrm{RA}_1 < K$). El puntaje poligénico intermedio, s_2, implica que la propensión L_2 correspondiente a ese individuo será parecida a la prevalencia de la enfermedad en la población. El valor más alto, s_3, implica una distribución L_3 desplazada hacia la derecha, mucho más cerca del umbral z_K y, por ende, con riesgo absoluto incrementado. Se puede observar que la relación entre S y el RA no es lineal: a medida que el puntaje poligénico crece y se acerca a z_K, incrementos pequeños conllevan un crecimiento cada vez más pronunciado del riesgo absoluto, pues las regiones centrales de la densidad Gaussiana comienzan a asomar tras el umbral.

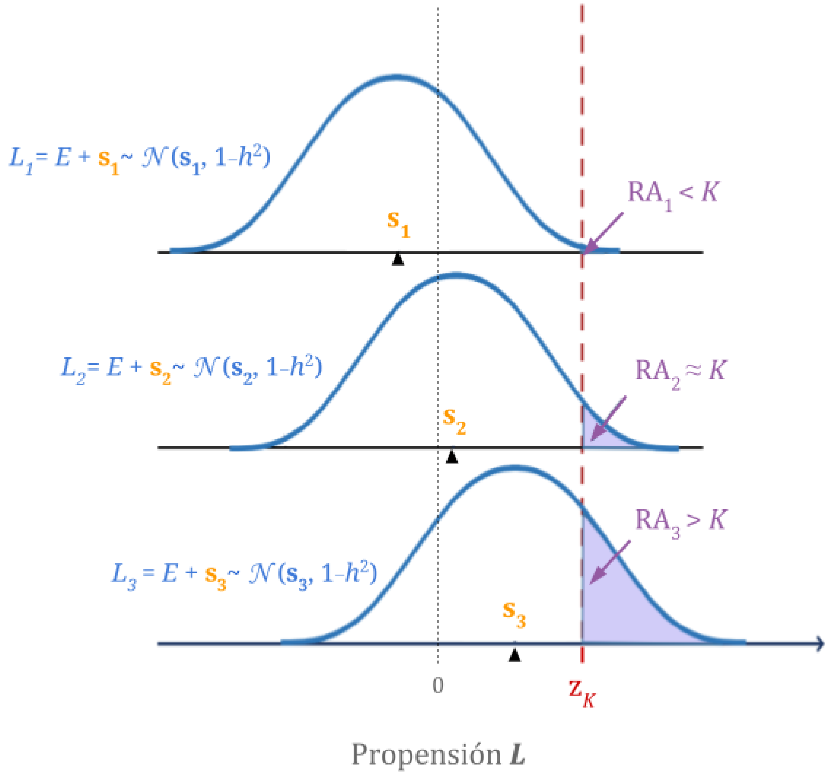


Figura 2.9: El eje x representa la propensión a una enfermedad, L. En línea punteada roja el umbral que determina la afección, z_K. Los tres triángulos negros son tres puntajes poligénicos posibles: s_1, s_2, s_3, que determinan tres distribuciones condicionales posibles de propensión: L_1, L_2, L_3 (curvas azules). Modificado a partir de [122].

Se resume el camino completo desde el PRS de un individuo particular hasta su riesgo absoluto de enfermar en la **Figura** 2.10.

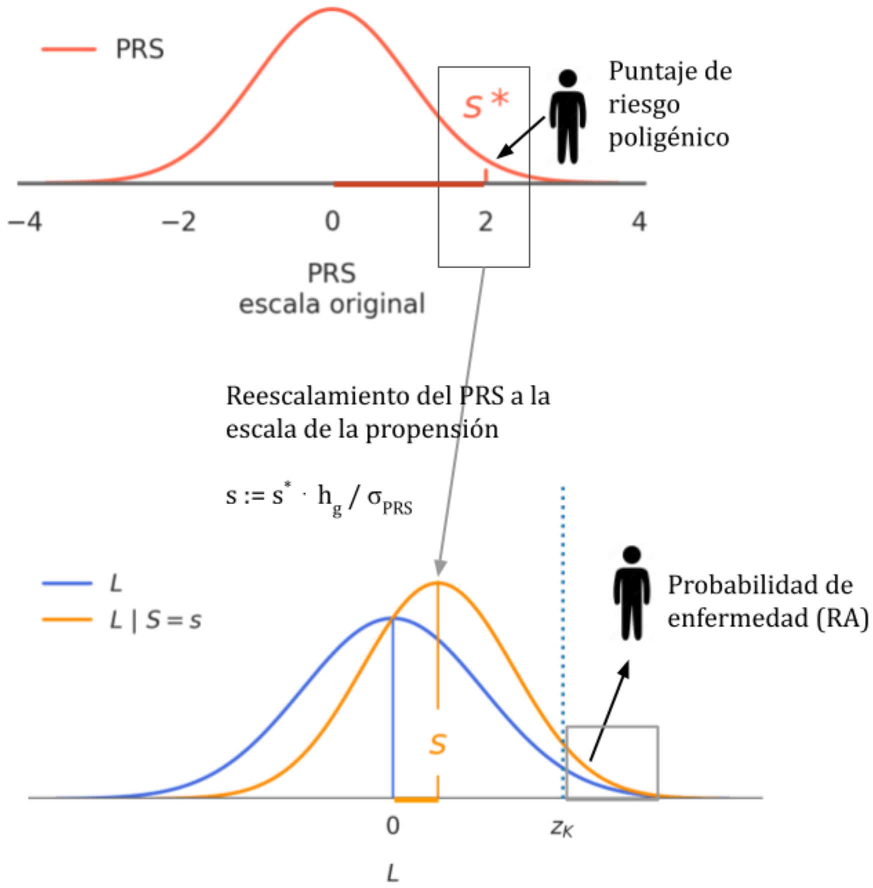


Figura 2.10: El PRS de un individuo particular, s^*, es obtenido con LDAK y debe ser reescalado como s teniendo en cuenta la heredabilidad h_g^2 del fenotipo analizado en la población de interés. Luego, el cálculo del riesgo absoluto de ese individuo es el área bajo la curva de la distribución $L \mid S = s$ tras el umbral z_K.

Algunas medidas relacionadas al RA serán de utilidad en los análisis. Las definimos a continuación.

2.3.6. IRA y NNH

Si uno quiere comparar el riesgo absoluto que resulta de dos tratamientos diferentes, digamos un tratamiento de base b y un tratamiento alternativo a, una medida natural es calcular la diferencia entre ambos riesgos. Cuando el foco está puesto en si el tratamiento alternativo produce o no *un aumento* en el riesgo, se suele calcular el incremento de riesgo absoluto (IRA), definido como:

$$\text{IRA} := \text{RA}_a - \text{RA}_b \tag{2.6}$$

En nuestro contexto, el IRA mide la probabilidad *adicional* de enfermar en el grupo que recibe el nuevo tratamiento, respecto del grupo de base. Por ejemplo, si $\text{RA}_b = 0.05$ (i.e. hay un 5 % de afectados en el grupo de referencia) y $\text{RA}_a = 0.06$, esta diferencia se mide como $\text{IRA} = 0.06 - 0.05 = 0.01$. Esto significa que en el grupo a se da un incremento de *un punto porcentual* en el riesgo de enfermar.

Un modo interesante de entender el impacto de un determinado valor de IRA es calcular el número necesario para hacer daño (NNH), definido como

$$\text{NNH} := \frac{1}{\text{IRA}} \tag{2.7}$$

Este valor representa la cantidad de individuos que deben someterse al nuevo tratamiento para que se produzca un "evento adicional", que en nuestro caso es la aparición de la enfermedad. Por ejemplo, consideremos que el tratamiento de base es la selección de un embrión al azar y que el tratamiento alternativo es la selección del embrión de PRS mínimo. Si los embriones seleccionados con el tratamiento alternativo tienen un IRA $= 0.01 = {}^{1}/_{100}$ de una enfermedad, el NNH $= \frac{1}{1/100} = 100$ se puede interpretar como la cantidad de embriones que habría que elegir con ese criterio para que aparezca entre ellos un enfermo *adicional*, es decir, un embrión que eventualmente desarrollará la enfermedad y que habría sido sano si hubiéramos elegido embriones con el método anterior.

2.3.7. Test *t* de diferencia de riesgo absoluto

Este capítulo parte de la idea de que la correlación negativa entre puntajes podría generar incrementos de riesgo inesperados. Formalizamos aquí el test que pone a prueba esta idea.

Consideremos pares ordenados de enfermedades. Nos interesa entender si los individuos de bajo puntaje *del primer fenotipo* tienen riesgo incrementado *del segundo fenotipo*. Para definir qué es un "bajo puntaje" del primer fenotipo, creamos una grilla de umbrales posibles de selección $U = \{0.1, 0.2, 0.3, 0.4, 0.5\}$, con la idea de investigar los resultados bajo definiciones progresivamente más laxas. Llamaremos a cada valor de U un "umbral de selección". La unidad de estos valores es el cuantil de PRS o qPRS. Los cuantiles fueron calculados para cada fenotipo por separado, con el panel completo de 404 individuos. Para cada $u \in U$, etiquetamos como "seleccionados" a los individuos con qPRS $< u$ y al resto como "no seleccionados". Por ejemplo, si $u = 0.30$, entonces los individuos con el 30 % de puntajes más bajos –i.e. con valores de S en los tres primeros deciles– serían los seleccionados y el 70 % restante serían los no seleccionados.

Calculamos luego el RA de todos los individuos *para el segundo fenotipo del par* con la fórmula de (2.4) y realizamos un test *t* de diferencia de medias para ver si el riesgo absoluto de la segunda enfermedad está significativamente incrementado en el grupo de seleccionados, que es el grupo de bajo PRS de la primera enfermedad, con respecto al grupo de no seleccionados. En la sección A.1.4 mostramos que encontrar esta diferencia implica que el grupo de seleccionados tiene el riesgo incrementado respecto de la población general.

Como dijimos, consideramos pares *ordenados* de fenotipos, es decir, las mismas dos enfermedades E_1 y E_2 se testean dos veces: como (E_1, E_2) y como (E_2, E_1). Ambas alternativas deben testearse por separado porque la relación buscada no es simétrica. La razón de que la relación no sea simétrica reside en que la conversión de PRS a RA no es igual para todos los fenotipos, pues depende de la heredabilidad y la prevalencia específicos de cada enfermedad. Por ende, la distribución conjunta de $(\text{PRS}_1, \text{RA}_2)$ no es la misma que la de $(\text{PRS}_2, \text{RA}_1)$. La diferencia de riesgo medio buscada depende de esa distribución conjunta.

Como mencionamos anteriormente, basamos el análisis en 142 fenotipos de enfermedad. Luego, el número de pares ordenados de enfermedades diferentes es $142 \cdot 141 = 20\,022$. Para cada par, analizamos 5 umbrales posibles de selección, lo que da un número de total de $20\,022 \cdot 5 = 100\,110$ tests realizados. Obsérvese que consideramos los pares *ordenados* de fenotips porque nada indica que el test sea simétrico.

Aplicamos el procedimiento de Benjamini-Hochberg [140] implementado en el paquete `statsmodels` de Python para obtener, a partir de los 100 110 P valores de los tests *t*, el mismo número de P valores ajustados por tests múltiples. Llamaremos P_{adj} (P ajustado) a estos valores:

$$P_{\text{adj}} := P \cdot \frac{m}{\text{rank}\,(P)}$$

donde m es el número total de tests, rank(P) es el *ranking* de P, es decir, la posición que ocupa cuando se ordenan los P valores de menor a mayor.

Este procedimiento controla la tasa de descubrimientos falsos (FDR) para un número m de tests independientes o positivamente correlacionados y es menos conservador que la elección de un umbral de Bonferroni en α/m.

Al usar esta corrección por tests múltiples, asumimos que todos los tests son independientes con excepción de dos casos. Por un lado, los pares ordenados que llevan los mismos fenotipos pero en distinto orden, es decir (F_i, F_j) y (F_j, F_i) para cualquier $i \neq j$, que asumimos como positivamente correlacionados. Por otro lado, para cada par ordenado particular, los 5 tests correspondientes a los cinco umbrales $u \in U$, que también asumimos como positivamente correlacionados.

Finalmente, filtramos los pares ordenados de enfermedades para los que el test *t* obtuvo un $P_{\mathrm{BH}} < 0.01$ con al menos un umbral $u \in U$, lo que indica un incremento significativo del riesgo absoluto de la segunda enfermedad en los individuos con bajo PRS de la primera. Cuando el mismo par de enfermedades tuvo un test significativo con más de un umbral u, nos quedamos con el umbral que produjo el menor P_{BH}.

2.3.8. Autoinmunidad $\mathcal{V}_{\mathbf{CMH}}$

Como dijimos en la introducción, es sabido que las enfermedades autoinmunes tienen variantes asociadas principalmente en el complejo mayor de histocompatibilidad. Con el fin de cuantificar cuánto peso relativo tiene esta región genómica en el puntaje poligénico de cada enfermedad, desarrollamos un índice de autoinmunidad genética, calculado como la proporción de la varianza del PRS que se debe a las variantes del CMH. Lo definimos a continuación.

Sean $\{\beta_i\}_{i=1}^n$ los efectos asociados al alelo alternativo de los n SNPs asociados a un fenotipo particular. Sean $\{f_i\}_{i=1}^n$ las frecuencias poblacionales de esos mismos alelos (en nuestro caso, utilizamos frecuencias de la población EUR de 1KG).

Sea G_i una variable aleatoria que representa la cantidad de alelos alternativos presentes en un genoma en el SNP i-ésimo. Asumiendo que los padres del individuo no están emparentados, la distribución de la variable es $G_i \sim \mathrm{Binom}\,(2, f_i)$, pues el alelo materno y el alelo paterno pueden pensarse como el resultado de dos muestreos independientes de los alelos presentes en la población, en la que la frecuencia de alelo alternativo es f_i.

Suponiendo que los n genotipos $\{G_i\}_{i=1}^n$ son independientes entre sí[5], la varianza del PRS para el fenotipo bajo análisis puede calcularse como

[5]Véase la sección A.1.3

$$\begin{aligned}
\mathrm{var}\,(S) &= \mathrm{var}\left(\sum_{i=1}^{n} \beta_i G_i\right) \\
&= \sum_{i=1}^{n} \mathrm{var}\,(\beta_i G_i) \\
&= \sum_{i=1}^{n} \beta_i^2 \mathrm{var}\,(G_i) \\
&= \sum_{i=1}^{n} 2\beta_i^2 f_i(1-f_i) \qquad (2.8)
\end{aligned}$$

La varianza del PRS debida a variantes del CMH puede calcularse de la misma manera, limitando la sumatoria a los índices de SNPs que se ubican en esa región.

Definimos entonces el índice de autoinmunidad genética, $\mathcal{V}_{\text{CMH}}$, como la proporción de varianza total del PRS debida a los SNPs del CMH:

$$\mathcal{V}_{\text{CMH}} = \frac{\text{var}\,(\text{PRS}_{\text{CMH}})}{\text{var}\,(\text{PRS})} = \frac{\sum_{j \in C} 2\beta_j^2 f_j(1-f_j)}{\sum_{i=1}^{n} 2\beta_i^2 f_i(1-f_i)} \in [0,1] \tag{2.9}$$

donde C es el conjunto de índices de los SNPs ubicados en el CMH[6].

[6]Este valor puede pensarse como una instancia particular de la ecuación (5) de [141].

2.3.9. Poligenicidad $\mathcal{K}_{0.90}$

Nos interesa cuantificar la poligenicidad de los fenotipos analizados, es decir, cuán diversa es su base genética. Intuitivamente, un fenotipo afectado por 5 SNPs del genoma es poco poligénico, mientras que un fenotipo afectado por 500 SNPs es muy poligénico. Así, una medida ingenua de la poligenicidad podría consistir en contar cuántas variantes genéticas afectan al fenotipo. Más formalmente, el criterio consistiría en contar en cuántas variantes se cumple que $|\beta_i| > 0$.

Sin embargo, una característica del modelo de SBayesR que utilizamos al correr LDAK es que cientos de miles de variantes genéticas reciben efectos β_i ajustados no nulos, aunque en su mayor parte ínfimos (en contraste, por ejemplo, con el modelo *lasso*, que "desactiva" a la mayoría de las variantes asignándoles $\beta_i = 0$). Por ende, el criterio sencillo de contar variantes con efecto no nulo no nos sirve.

Queremos además que nuestra definición de poligenicidad tome en cuenta la diferencia entre las magnitudes de los β_i. Intuitivamente, un fenotipo con 100 variantes no nulas de efectos similares es más poligénico que otro con 100 variantes no nulas, pero en el que las primeras cinco variantes tienen un efecto mucho mayor al resto.

Llamemos $\beta_{(k)}$ al efecto de la variante k-ésima si ordenamos a las variantes según la magnitud de su efecto. Es decir, $|\beta_{(1)}| \geq |\beta_{(2)}| \geq \cdots \geq |\beta_{(n)}|$, si n es el número total de variantes disponibles. Definimos $\mathcal{V}_{(k)}$, la proporción de varianza del PRS capturada cuando se incluyen únicamente los k SNPs de mayor efecto:

$$\mathcal{V}_{(k)} := \frac{\text{var}\left(\text{PRS}_{(k)}\right)}{\text{var}\,(\text{PRS})} = \frac{\sum_{i=1}^{k} \beta_{(i)}^2 f_{(i)}(1-f_{(i)})}{\sum_{j=1}^{n} \beta_j^2 f_j(1-f_j)} \in [0,1]$$

donde $f_{(i)}$ es la frecuencia alélica correspondiente al SNP con efecto $\beta_{(i)}$. Al igual que la definición de $\mathcal{V}_{\text{CMH}}$ en (2.9), esta medida es una proporción de varianza capturada por un subconjunto de SNPs.

Llamemos ahora $\mathcal{K}_{(p)}$ al número mínimo de SNPs necesarios para explicar la proporción p de varianza total del PRS:

$$\mathcal{K}_p := \text{mín}\{k \in \mathbb{N}^+ \mid \mathcal{V}_{(k)} \geq p\}$$

Con esta definición, calculamos el $\mathcal{K}_{0.90}$ de todos los fenotipos como un resumen de la poligenicidad. El $\mathcal{K}_{0.90}$ indica cuántos de los SNPs de mayor efecto se requieren para capturar el 90 % de la varianza del PRS.

2.3.10. Correlaciones genéticas ya descriptas entre fenotipos

El test t descripto nos permitió centrar la atención en algunos pares de fenotipos que llamaremos "fenotipos anticorrelacionados". Para estos pares, obtuvimos también la correlación genética r_g del trabajo de [132]. Este valor fue calculado en base a los mismos *summary statistics* de GWAS que utilizamos en este capítulo, con el método de regresión de LD *score* [77]. Nuestro objetivo es contrastar los hallazgos de esta tesis, basados en el test t descripto, con hallazgos previos de correlación genética entre enfermedades.

2.3.11. Simulación de embriones

Nuestra primera búsqueda de fenotipos anticorrelacionados se basó en los puntajes poligénicos de individuos de la población general. Esto nos permitió centrar la atención en algunos pares de fenotipos. Sin embargo, en el contexto del PGT-P, los embriones entre los que una pareja puede elegir durante la FIV tienen en común al menos el 50 % de sus genomas. Esto implica, en particular, que sus puntajes poligénicos serán en media mucho más parecidos que los de dos indivuos no relacionados. Por ende, nos interesa también explorar si el incremento de riesgo de enfermedad ocurre al analizar la distribución de puntajes de embriones emparentados.

El software ORIGAMI (*Offspring Risk Inference through Gamete Simulation*) [81] fue concebido específicamente para analizar la distribución del riesgo poligénico en embriones de una misma pareja. ORIGAMI permite simular genomas de embriones a partir de los genotipos parentales. Internamente, utiliza mapas genéticos para simular de manera realista eventos de recombinación en las gametas parentales y ensamblar los genomas haploides resultantes de la recombinación. Luego, simula el genoma del embrión como una combinación de dos gametas tomadas al azar. Se resume el proceso en la **Figura** 2.11.

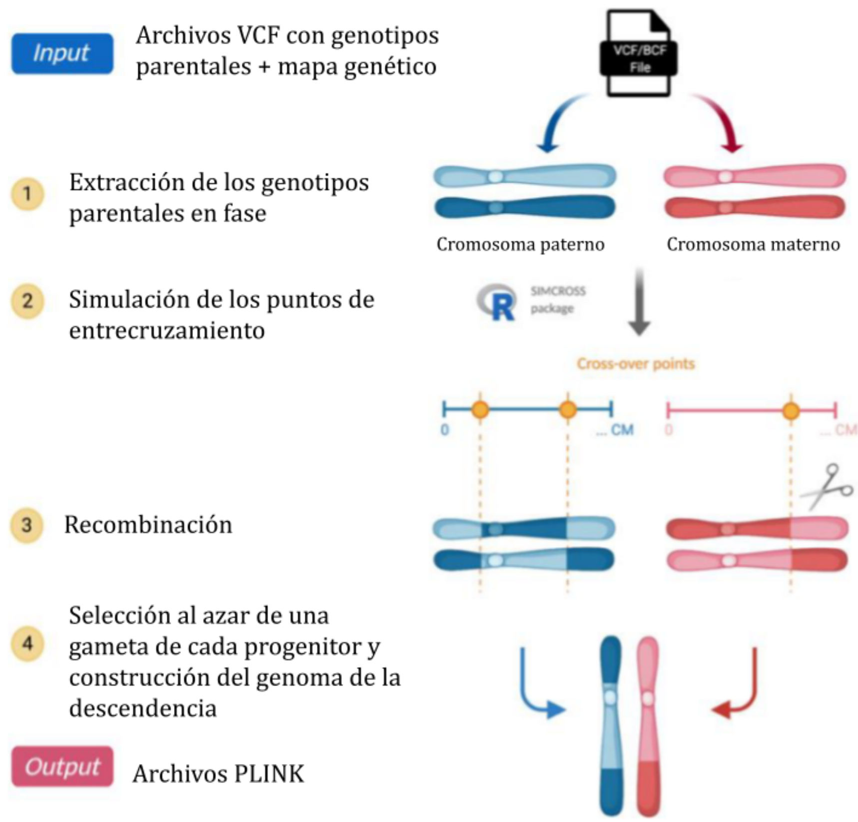


Figura 2.11: Resumen del *pipeline* implementado por ORIGAMI para simular genomas de embriones a partir de genotipos parentales.

Antes de continuar, introduzcamos una notación que facilitará la descripción del experimento. Llamaremos Fenotipo$_1$ al primer fenotipo de cada uno de los pares ordenados de fenotipos. Por extensión, si hablamos de PRS_1, nos referimos al puntaje poligénico del Fenotipo$_1$. Análogamente, Fenotipo$_2$ y RA_2 se refieren al segundo fenotipo del par y su riesgo absoluto.

El primer paso en la simulación de embriones fue generar una lista de las $202^2 = 40\,804$ parejas posibles del dataset de 404 individuos (donde 202 son XX y 202 son XY). Luego, para cada par de fenotipos anticorrelacionados, asignamos a cada pareja un qPRS midparental del Fenotipo_1, calculado como el qPRS_1 promedio de padre y madre. A continuación, elegimos en cada decil de PRS_1 midparental $k = 50$ parejas al azar (entre todas las parejas posibles en cada decil, que son más de 50). Es decir: elegimos 50 parejas con qPRS_1 midparental en $[0, 0.1)$, 50 parejas con qPRS_1 midparental en $[0.1, 0.2)$, etc. Esto da un total de $10k = 500$ parejas distribuidas uniformemente en los diez deciles de PRS_1, como se ilustra en la **Figura** 2.12.

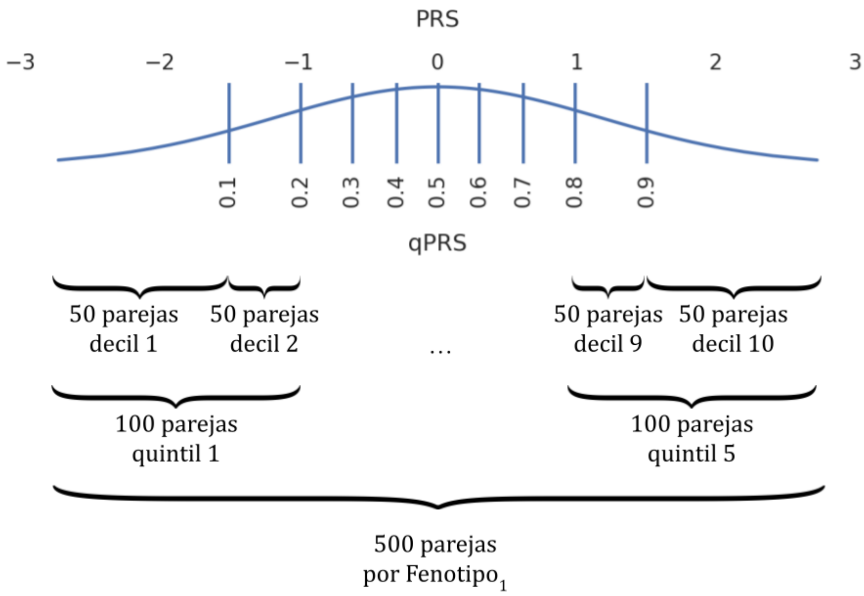


Figura 2.12: Simulación de 50 parejas por decil de PRS_1. Agrupamiento de 100 parejas por quintil para los tests.

Luego, para cada pareja, simulamos $n = 5$ embriones usando ORIGAMI. Esto produjo $10kn = 2\,500$ embriones simulados por cada Fenotipo_1. Dados los 8 Fenotipos_1 que participan de los 14 pares anticorrelacionados, simulamos un total de $8 \cdot 2\,500 = 20\,000$ embriones en $8 \cdot 500 = 4\,000$ parejas formadas *in silico*.

Finalmente, calculamos para cada embrión el PRS_1 y el RA_2.

2.3.12. Estrategias de selección de embriones

Estudiamos tres estrategias posibles de selección de embriones:

1. Elección *random*: entre los 5 embriones disponibles de cada pareja, elegimos uno al azar. Equivale a no usar el PRS como criterio de selección.

2. Priorización del Bajo Riesgo (PBR): elegimos el embrión de PRS mínimo entre los embriones disponibles de cada pareja.

3. Exclusión del Alto Riesgo (EAR): elegimos un embrión al azar entre los 4 embriones que quedan al excluir el embrión de PRS máximo.

De las dos estrategias que utilizan el PRS, es claro que PBR es la más agresiva, en el sentido de que busca reducir lo más posible el riesgo de enfermedad en la descendencia, mientras que EAR es más cauta, pues sólo intenta evitar la situación en la que se transfiere por azar el embrión de riesgo máximo.

Por cada par de fenotipos, en cada una de las 500 parejas definidas en función del PRS_1, elegimos tres embriones, uno por cada estrategia. Los llamaremos embrión *random*, embrión PBR y embrión EAR. Nótese que, por azar, el embrión elegido en estrategias diferentes puede coincidir.

A continuación, analizamos las estrategias PBR y EAR por separado, contrastando cada una con la estrategia *random*. Describimos primero el procedimiento para los embriones PBR. Por cada uno de los 14 pares de fenotipos anticorrelacionados en la población general, agrupamos a las parejas por *quintil* de PRS_1 midparental, es decir, en cinco grupos que dependen del Fenotipo_1 y que llamamos Q_1, …, Q_5. La lógica de este agrupamiento es una intuición previa de que el problema del riesgo incrementado debería ser más grave en los PRS_1 bajos, de modo que nos interesa enfocar el test allí y no disminuir el poder de detección mezclando a esos casos con los de PRSs intermedios. Agrupar por decil implicaría multiplicar demasiado los tests, de modo que encontramos en el agrupamiento por quintil un buen balance, aunque también podría haber sido por cuartil o tal vez por tercil. Nuestra atención estará puesta, por ende, en el primer quintil de cada par, i.e. en los "embriones Q_1", que son aquellos cuyos padres tienen un PRS_1 midparental en el 20 % más bajo de la población.

Por cada par de fenotipos y por cada quintil, realizamos un test de rangos con signo de Wilcoxon, donde comparamos el valor de RA_2 de los embriones *random* y el valor de RA_2 de los embriones PBR del quintil. El test puede pensarse como "de muestras apareadas", pues la serie a la que se le aplica el test consiste de las diferencias entre el RA_2 del embrión PBR y el RA_2 del embrión *random*, calculada *dentro de cada familia*.

La hipótesis nula del test es que las diferencias de RA_2 se reparten con igual probabilidad por encima y por debajo del cero, es decir, que la estrategia PBR no implica un aumento del riesgo, lo que esperaríamos si los puntajes de riesgo no están correlacionados, mientras que la hipótesis alternativa utilizada es que el RA_2 es *mayor* en los embriones PBR respecto de los embriones *random*. Se ilustra esto en la **Figura** 2.13. El mismo procedimiento fue utilizado para analizar los embriones embriones EAR.

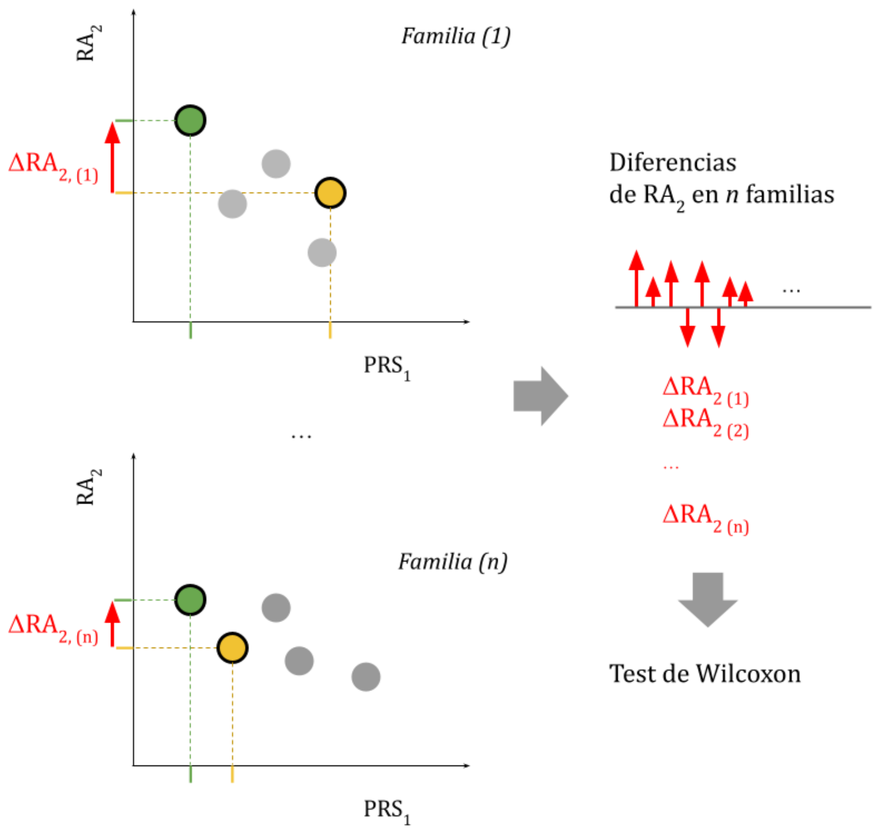


Figura 2.13: Ilustración de cómo se obtienen las diferencias de RA_2 que permiten realizar el test de Wilcoxon.

Nótese que la decisión de seleccionar embriones basándonos en el PRS_1 se debe a que así se propone el PGT-P hoy en algunas clínicas, mientras que la decisión de buscar si hay un incremento de RA_2 como contrapartida (en lugar de buscar un incremento de PRS_2) se debe a que el riesgo absoluto es un valor de probabilidad de enfermedad que puede traducirse en consecuencias numéricas concretas para la población, a diferencia del PRS, que sólo nos habla del riesgo relativo de un individuo pero no de su probabilidad de enfermar.

Para realizar el test de Wilcoxon utilizamos la implementación del paquete `scipy`, en Python. El número total de tests dadas 2 estrategias de selección, 5 quintiles y 14 pares de fenotipos es de $2 \cdot 5 \cdot 14 = 140$. Los P valores de esta segunda ronda de tests fueron corregidos para controlar el FDR con el método de Benjamini-Hochberg antes citado.

La decisión de utilizar el test de Wilcoxon (en lugar del test t de diferencia de medias) nos permite no asumir ninguna distribución en los riesgos absolutos comparados, pues nada indica que el valor del RA correspondiente al embrión de mín(PRS_1) o al embrión EAR se distribuya normalmente.

2.4. Resultados

2.4.1. Enfermedades con riesgo anticorrelacionado

Encontramos 14 pares ordenados de enfermedades en los que los individuos con bajo riesgo de una enfermedad tienen riesgo significativamente incrementado de otra enfermedad (FDR, nivel de significancia $\alpha = 0.01$).

Por comodidad, llamaremos de aquí en adelante a los 14 pares de este primer hallazgo como los “pares anticorrelacionados” y a su relación como “anticorrelación”. Recuérdese, sin embargo, que el hallazgo se basa en el incremento de riesgo medio detectado con un test t, es decir, no se trata de un análisis de correlación genética como los de la regresión de LD *score* usuales en la literatura.

Este primer análisis fue realizado en la población general de ancestría europea, representada por el dataset 1KGP-EUR de 404 individuos.

La lista completa de los pares anticorrelacionados se presenta en la **Tabla** 2.1, con el P valor del test t y su ajuste por FDR (P_{adj}). Para cada par, se lista únicamente el umbral u que produjo la mayor diferencia de riesgos absolutos, es decir, el umbral con el que el resultado es más grave en términos de incremento del riesgo, de los cinco umbrales analizados ($u \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$).

Asignamos a cada par significativo una letra {A, …, N}, con la que nos referimos al par en figuras subsiguientes.

En la **Figura** 2.14 tomamos como ejemplo el par con la anticorrelación más significativa, artritis reumatoidea y esclerosis múltiple. El resto de los pares se grafican en las **Figuras** 2.15 a 2.27. Como antes, el subíndice 1 refiere a la primera enfermedad del par, en base a cuyo puntaje poligenico (PRS_1) se realiza la selección, mientras que el subíndice 2 refiere a la segunda enfermedad del par, en la que el riesgo absoluto (RA_2) se ve incrementado.

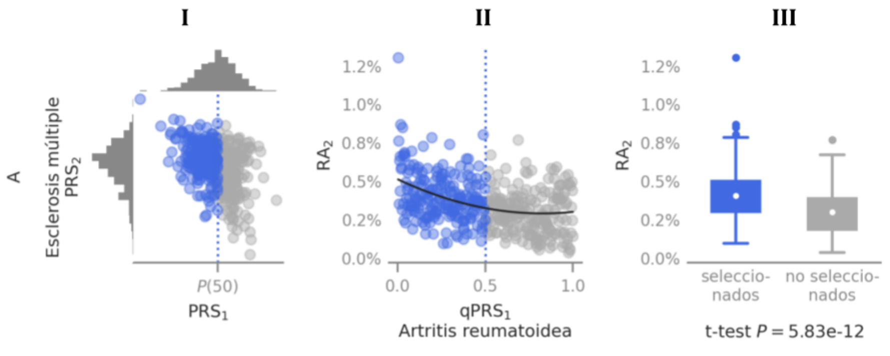


Figura 2.14: Par “A” de fenotipos anticorrelacionados en la población general: esclerosis múltiple y artritis reumatoidea. Explicación detallada en el cuerpo del texto.

	Fenotipo 1	Fenotipo 2	u	P	P_{adj}
A	Artritis reumatoidea	Esclerosis múltiple	0.5	5.8e-12	5.8e-07
B	Celiaquía	Espondilitis anquiliosante	0.4	1.9e-10	6.2e-06
C	Artritis reumatoidea	Celiaquía	0.4	2.5e-10	6.2e-06
D	Hipertiroidismo	Psoriasis	0.1	1.0e-08	0.0001
E	Sarcoidosis	Espondilitis anquiliosante	0.3	6.8e-08	0.0005
F	Esclerosis múltiple	Artritis reumatoidea	0.3	9.8e-08	0.0006
G	Hipertiroidismo	Espondilitis anquiliosante	0.4	1.2e-07	0.0007
H	Addison	Espondilitis anquiliosante	0.1	4.4e-07	0.0020
I	Esclerosis múltiple	Espondilitis anquiliosante	0.4	5.3e-07	0.0023
J	Espondilitis anquiliosante	Celiaquía	0.5	6.6e-07	0.0027
K	Celiaquía	Artritis reumatoidea	0.5	1.2e-06	0.0043
L	Sarcoidosis	Psoriasis	0.1	3.2e-06	0.0089
M	Celiaquía	Psoriasis	0.1	3.6e-06	0.0096
N	Psoriasis	Hipertiroidismo	0.4	3.7e-06	0.0096
	Addison	Psoriasis	0.2	4.1e-06	0.0105
	Espondilitis anquiliosante	Esclerosis múltiple	0.4	4.6e-06	0.0112
	Artritis reumatoidea	Sarcoidosis	0.2	2.1e-05	0.0404
	Celiaquía	Rosacea	0.5	2.1e-05	0.0404

Tabla 2.1: Pares anticorrelacionados en población europea del dataset 1KGP-EUR (404 individuos). La línea horizontal deja arriba a los pares que cumplen $P_{\text{adj}} < 0.01$, con el umbral de selección u indicado. Debajo de la línea, pares adicionales que cumplen $P_{\text{adj}} < 0.05$.

En **I** de la **Figura** 2.14 se observa la distribución conjunta de los PRSs de ambas enfermedades, es decir, PRS_1 vs. PRS_2. El umbral u (en este caso $u = 0.50$) se muestra como una línea punteada vertical, que deja a la izquierda al $(100 \cdot u)\,\% = 50\,\%$ de los individuos con menor riesgo poligénico, que en este contexto llamamos "individuos seleccionados" y coloreamos en azul. En unidades del PRS_1, este umbral es el cuantil 0.50 o percentil 50, anotado en el eje X como $P(50)$. A la derecha del umbral quedan los individuos "no seleccionados", en gris. Los histogramas marginales dan cuenta de la distribución aproximadamente simétrica de cada PRS. En los PRSs de celiaquía y de psoriasis, sin embargo, se observa cierta asimetría (véase la **Figura** 2.26).

En **II** se grafica la relación entre el cuantil de PRS_1 (denominado qPRS_1) y el riesgo absoluto de la enfermedad correlacionada, RA_2. La línea punteada equivale a la línea punteada del primer gráfico, en el umbral que deja a la izquierda al 50 % de los individuos. Se añade una regresión polinómica de orden 2 para facilitar la visualización de la tendencia de la nube de puntos, pero nótese que el hallazgo no depende de dicha regresión. Una inspección visual ya evidencia que el riesgo absoluto parece incrementado en los individuos seleccionados.

El riesgo absoluto o RA_2 es la medida cuyo incremento nos interesa detectar. El método de detección se ilustra en **III**, donde se describe el riesgo en los individuos seleccionados y en los no seleccionados con boxplots separados, haciendo más evidente lo que ya se observaba en el gráfico previo. Se muestra además el P valor del test *t* de diferencia de riesgo medio entre ellos. El punto blanco en cada boxplot es la media que se compara en el test *t*.

Adicionalmente, para cada uno de los 14 pares de nuestro hallazgo, listamos en la sección A.1.5 el valor de correlación genética r_g calculado con una regresión de *LD score*, tomadas del trabajo de [132]. Nótese que nuestro hallazgo, basado en el riesgo aumentado en individuos del dataset 1KGP-EUR, contrasta con las correlaciones genéticas no significativas derivadas de la regresión *LD score*.

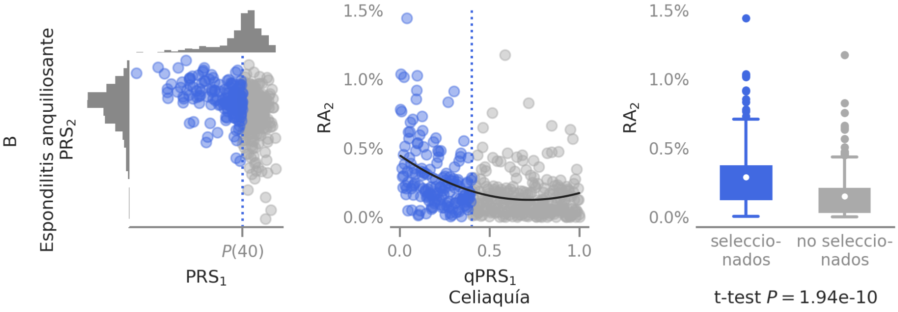


Figura 2.15: Par B: Celiaquía vs. Espondilitis anquilosante.

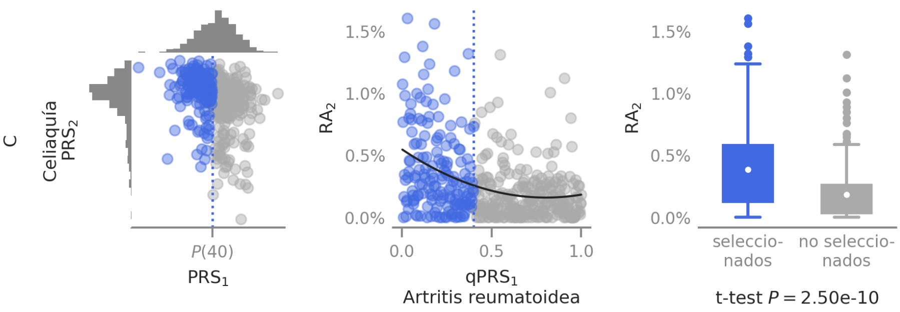


Figura 2.16: Par C: Celiaquía vs. Artritis reumatoidea.

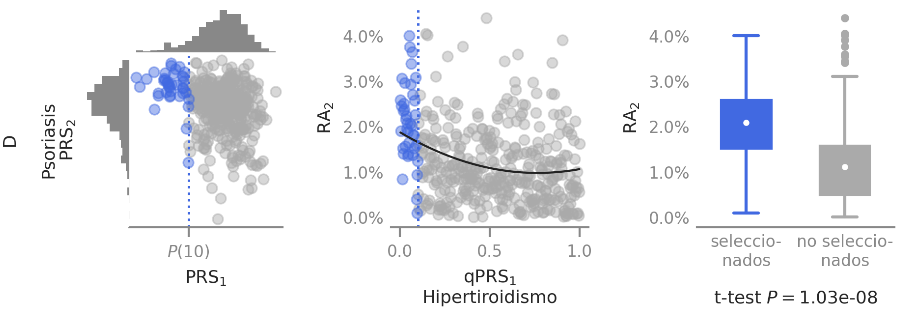


Figura 2.17: Par D: Hipertiroidismo vs. Psoriasis.

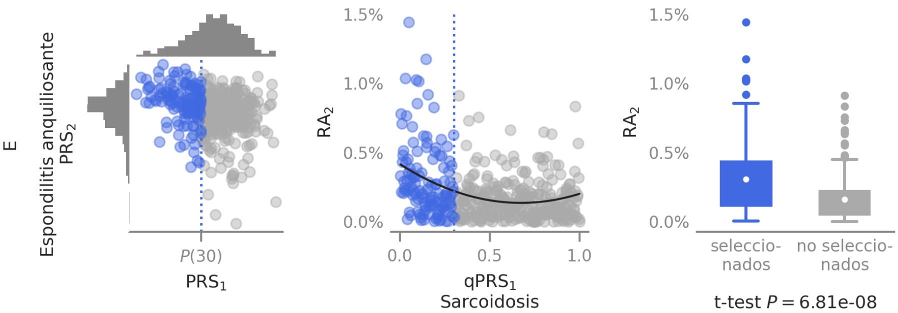


Figura 2.18: Par E: Sarcoidosis vs. Espondilitis anquilosante.

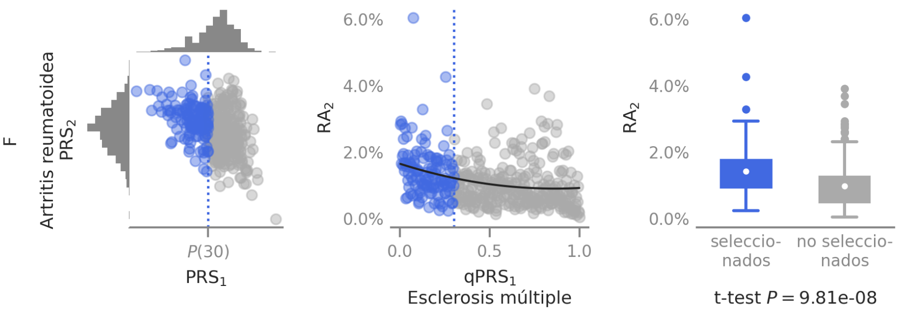


Figura 2.19: Par F: Esclerosis múltiple vs. Artritis reumatoidea.

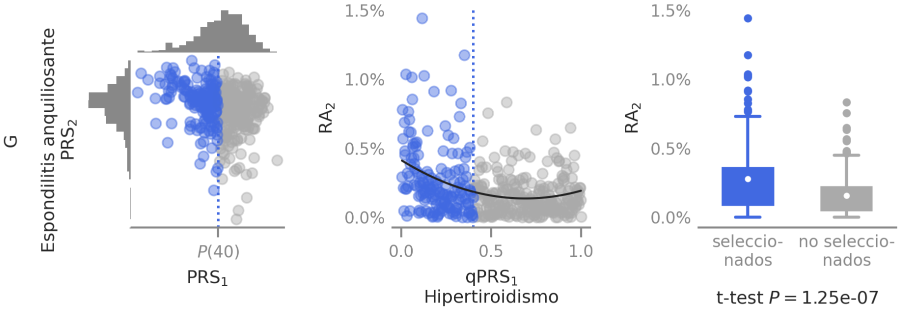


Figura 2.20: Par G: Hipertiroidismo vs. Espondilitis anquilosante.

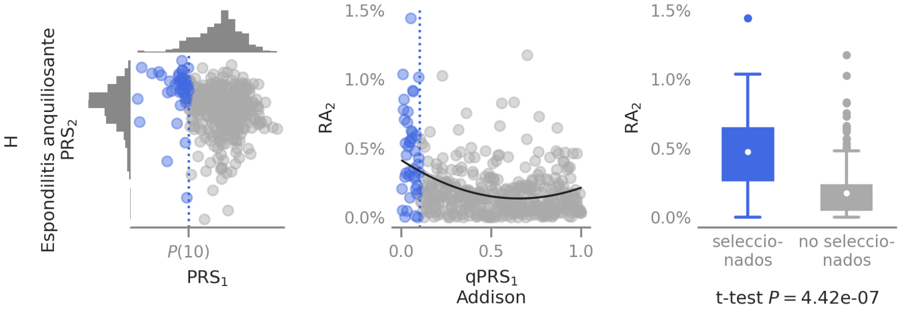


Figura 2.21: Par H: Addison vs. Espondilitis anquilosante.

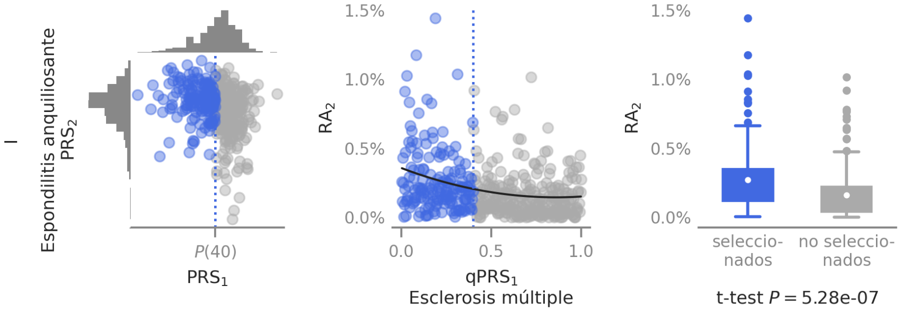


Figura 2.22: Par I: Esclerosis múltiple vs. Espondilitis anquilosante.

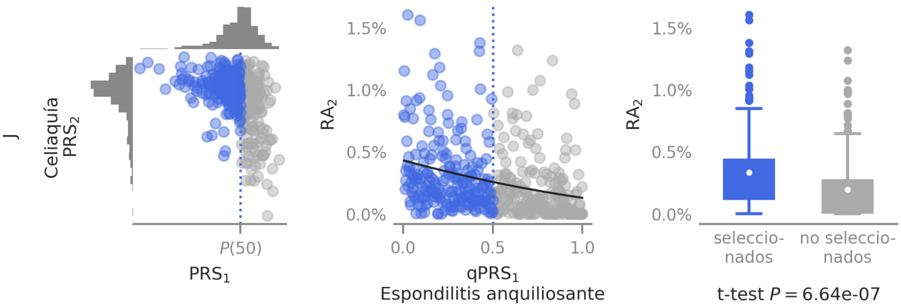


Figura 2.23: Par J: Espondilitis anquilosante vs. Celiaquía.

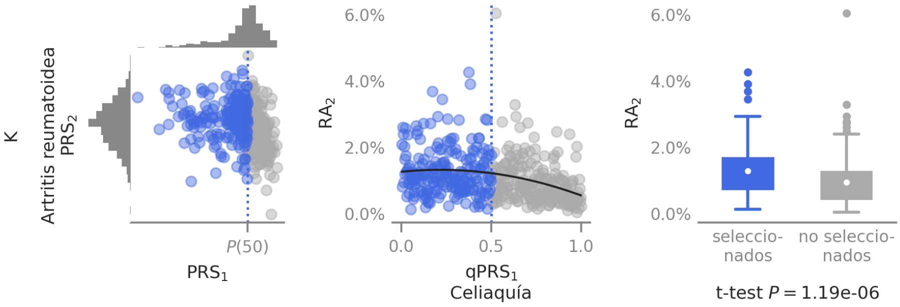


Figura 2.24: Par K: Celiaquía vs. Artritis reumatoidea.

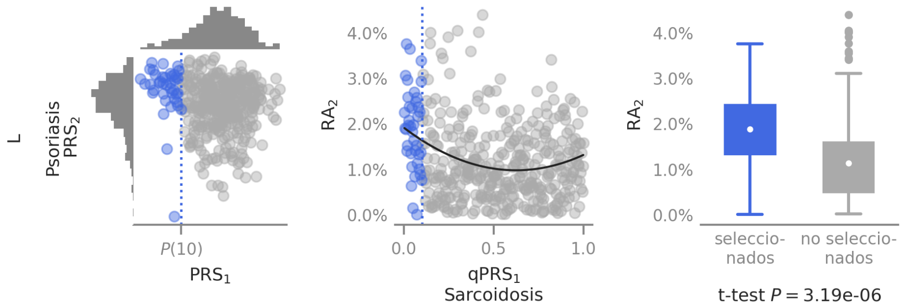


Figura 2.25: Par L: Sarcoidosis vs. Psoriasis.

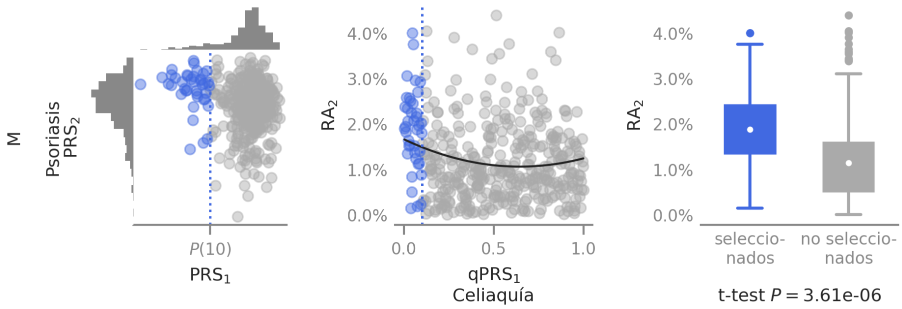


Figura 2.26: Par M: Celiaquía vs. Psoriasis.

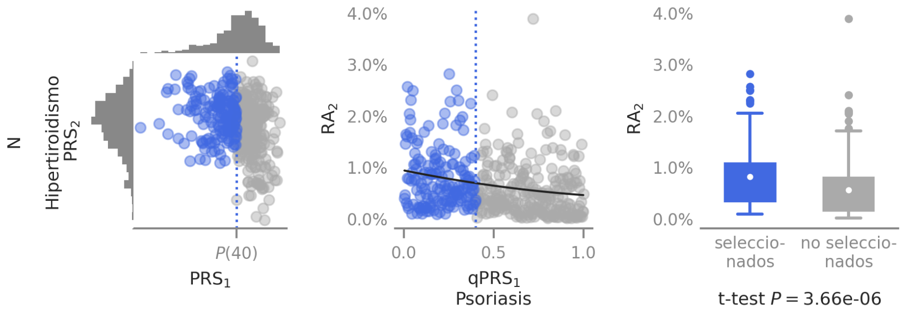


Figura 2.27: Par M: Hipertiroidismo vs. Psoriasis.

2.4.2. Análisis de los fenotipos involucrados en anticorrelaciones

En los 14 pares descriptos, hay algunas enfermedades que se repiten –por ejemplo, la espondilitis anquilosante participa en seis pares. Observamos un total de 8 fenotipos únicos en las anticorrelaciones, enumerados en la **Tabla** 2.2 junto a algunas métricas que serán discutidas en las secciones siguientes.

	Fenotipo	Código UK Biobank	Poligenicidad $\mathcal{K}_{0.90}$	Autoinmunidad $\mathcal{V}_{\text{CMH}}$	Prevalencia (%) $100 \cdot K$	Heredabilidad h_g^2
1	Espondilitis anquilosante	20002_1313	22	0.79	0.29	0.14
2	Artritis reumatoidea	20002_1464	5 802	0.69	1.11	0.06
3	Esclerosis múltiple	20002_1261	39	0.60	0.37	0.03
4	Hipertiroidismo	20002_1225	25	0.38	0.76	0.12
5	Addison	20002_1234	420	0.26	0.04	0.25
6	Sarcoidosis	20002_1371	57	0.11	0.20	0.17
7	Psoriasis	20002_1453	78	0.04	1.16	0.11
8	Celiaquía	20002_1456	600	0.02	0.44	0.26

Tabla 2.2: Detalle de los ocho fenotipos involucrados en anticorrelaciones significativas en población general.

Poligenicidad

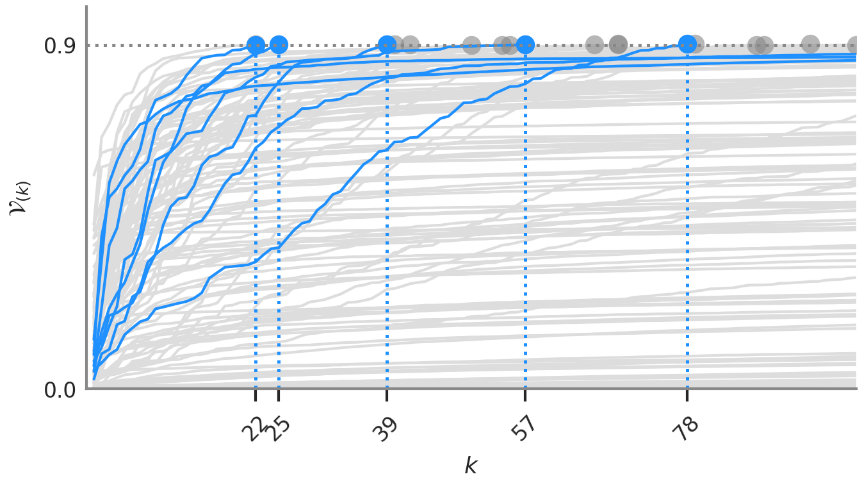


Figura 2.28: Eje x: número k de SNPs. Eje y: proporción $\mathcal{V}_{(k)}$ de varianza del PRS explicada por los k SNPs de mayor magnitud de efecto en cada fenotipo. Se resaltan en azul las series de los fenotipos involucrados en anticorrelaciones. Los círculos marcan el valor de $\mathcal{K}_{0.90}$ de los fenotipos de poligenicidad menor a 100.

En la **Figura** 2.28 se ilustra el método utilizado para hallar el valor $\mathcal{K}_{0.90}$, nuestra medida de poligenicidad. El gráfico muestra la proporción de varianza capturada a medida que se incluye un número creciente de SNPs en un puntaje, con una trayectoria por cada uno de los puntajes. Esta proporción de varianza es el $\mathcal{V}_{(k)}$, representado en el eje Y del gráfico, a medida que el número k de SNPs aumenta en el eje X. Se resaltan en azul los ocho fenotipos implicados en el primer resultado, mientras que el resto se ilustra en gris. El $\mathcal{K}_{0.90}$, representado con puntos azules o grises, corresponde al primer k de cada trayectoria que llega a $\mathcal{V}_{(k)} \geq 0.90$. Cuanto menor es este valor, menos poligénica es una enfermedad.

La espondilitis anquilosante, por ejemplo, es el fenotipo de menor poligenicidad entre los anticorrelacionados, pues con sólo 22 SNPs se captura el 90 % de la varianza de su PRS. Le sigue el hipertiroidismo, con 25 SNPs.

En la **Figura** 2.29 se ve que cinco de los ocho fenotipos del primer hallazgo tienen una poligenicidad relativamente baja, con valores de $\mathcal{K}_{0.90} < 100$. Nótese que la mayoría de las enfermedades del dataset UKB-GWAS tienen valores de $\mathcal{K}_{0.90} > 10^5$ (esto se ve en la mayor barra horizontal mayor del histograma marginal). Por ende, nuestro hallazgo de anticorrelaciones se basa en fenotipos de poligenicidad mayoritariamente baja, algo atípicos en el dataset, con la artritis reumatoidea como la enfermedad más poligénica ($\mathcal{K}_{0.90} = 5\,802$ SNPs).

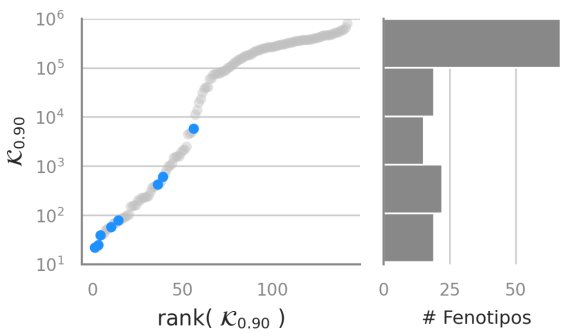


Figura 2.29: Distribución de poligenicidad en los 142 fenotipos del dataset. Se destacan en azul los ocho fenotipos involucrados en anticorrelaciones. **Izquierda:** Los 142 fenotipos ordenados por ranking de poligenicidad (eje X). **Derecha**: Histograma de los valores de poligenicidad.

Autoinmunidad genética

En la **Figura** 2.30 se grafica autoinmunidad genética medida como $\mathcal{V}_{\mathrm{CMH}}$, la proporción de varianza del PRS explicada por los SNPs del complejo mayor de histocompatibilidad. Puede apreciarse que los ocho fenotipos anticorrelacionados tienen valores de $\mathcal{V}_{\mathrm{CMH}}$ excepcionalmente altos en relación al resto del dataset. Mientras que en la mayoría de los fenotipos los SNPs del CMH explican menos del 1 % de la varianza (valores de $\mathcal{V}_{\mathrm{CMH}} < 10^{-2}$ en el gráfico), en seis de los ocho fenotipos involucrados en anticorrelaciones este valor supera el 10 %, con el caso extremo de la espondilitis anquilosante en 80 %. La psoriasis y la celiaquía tienen valores más pequeños (2 % y 4 % respectivamente), pero que también son altos en el contexto del dataset completo.

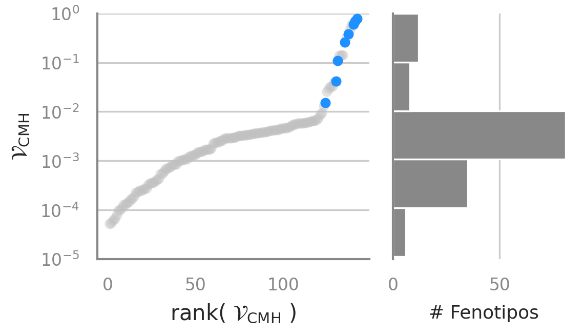


Figura 2.30: Distribución de autoinmunidad genética en los 142 fenotipos del dataset. Se destacan en azul los ocho fenotipos involucrados en anticorrelaciones. **Izquierda:** Los 142 fenotipos ordenados por autoinmunidad. **Derecha**: Histograma de los valores de autoinmunidad.

Prevalencia

En la **Figura** 2.31 se grafica la prevalencia de cada enfermedad en la población europea, estimada en base a los individuos de UK Biobank [135]. Puede observarse que las prevalencias se encuentran principalmente entre un enfermo cada cien y un enfermo cada mil individuos. En este aspecto los fenotipos en anticorrelaciones tienen valores típicos, con la única de excepción de la enfermedad de Addison, de prevalencia muy baja: está presente sólo en 4 de cada 10 000 individuos.

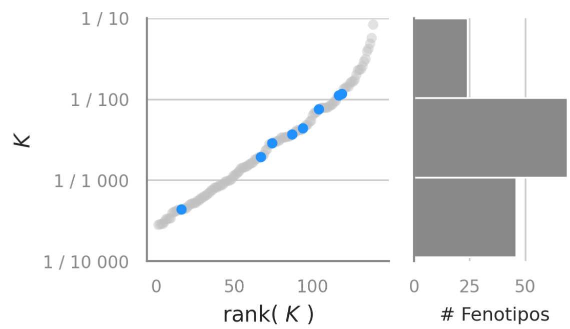


Figura 2.31: Distribución de prevalencia en los 142 fenotipos del dataset. Se destacan en azul los ocho fenotipos involucrados en anticorrelaciones. **Izquierda:** Los 142 fenotipos ordenados por prevalencia. **Derecha**: Histograma de los valores de prevalencia.

Heredabilidad

En la **Figura** 2.32 se grafica la heredabilidad de SNPs en la escala de la propensión de los fenotipos, estimada por [135] al igual que la prevalencia. Observamos también aquí que los valores de h_g^2 en los ocho fenotipos del hallazgo son típicos en el dataset de UKB-GWAS, con la mayoría ubicado por debajo debajo de 0.20. La celiaquía se destaca como el fenotipo de mayor heredabilidad entre ellos, con $h_g^2 = 0.264$, pero no se trata de un valor extremo. La esclerosis múltiple, con $h_g^2 = 0.028$ es el fenotipo menos heredable de los ocho, pero tampoco es un valor extremo. Esto puede notarse en el ranking (eje X): el primer punto azul, correspondiente a la esclerosis múltiple, está en el puesto 19 de 142, es decir que tiene una heredabilidad mayor al 13 % de los fenotipos menos heredables del dataset.

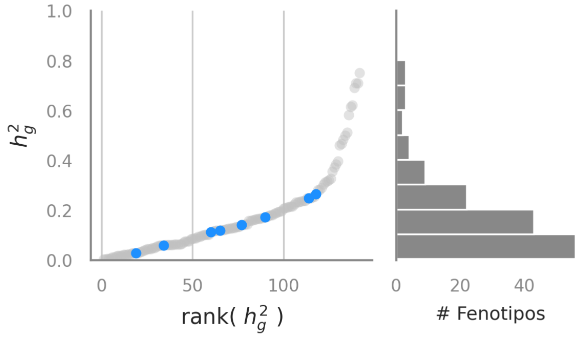


Figura 2.32: Distribución de heredabilidad en los 142 fenotipos del dataset. Se destacan en azul los ocho fenotipos involucrados en anticorrelaciones. **Izquierda:** Los 142 fenotipos ordenados por heredabilidad. **Derecha**: Histograma de los valores de heredabilidad.

2.4.3. Otras anticorrelaciones sugeridas

Las 14 anticorrelaciones encontradas no están "aisladas", sino que, como mencionamos, comparten fenotipos. Por ejemplo, la artritis reumatoidea se anticorrelaciona con la esclerosis múltiple y también con la celiaquía. Estos dos hallazgos sugieren, por ejemplo, que la celiaquía y la esclerosis múltiple no deberían anticorrelacionarse entre sí. Se pueden hacer inferencias similares con los otros pares, siguiendo la misma lógica.

Para entender si este tipo de inferencia es correcta, realizamos el siguiente procedimiento. Partimos del fenotipo más repetido, la espondilitis, y lo ubicamos "a un lado" –digamos, en el grupo I. Todas las enfermedades anticorrelacionadas con la espondilitis se ubican "del otro lado" –digamos, en el grupo II. Continuamos con los demás pares, evaluando si es posible mantener una distribución en dos grupos, con las anticorrelaciones siempre entre fenotipos de grupo distinto. La intuición se cumple: las anticorrelaciones encontradas parecen describir naturalmente dos grupos, donde las enfermedades del mismo grupo no se correlacionan negativamente entre sí.

Esta idea es más clara al ilustrarla como un grafo dirigido, como hacemos en la **Figura** 2.33. Cada nodo es un fenotipo y cada flecha es una anticorrelación, apuntando en dirección a la segunda enfermedad del par. Las 14 anticorrelaciones significativas con $\alpha = 0.01$ se dibujan como flechas azules. Las flechas bidireccionales (e.g. entre artritis reumatoidea y esclerosis mútliple) indican que la anticorrelación fue significativa en ambas direcciones.

Dada la disposición natural en dos grupos sugerida por los hallazgos, faltarían otras anticorrelaciones inter-grupo para completar el cuadro. Fuimos a buscarlas directamente a la lista de 20 022 pares analizados y descubrimos algunas de ellas tienen los P valores que siguen tras el umbral de significancia trazado. En la **Tabla** 2.1 se listan bajo la línea, como pares no significativos cuando el nivel de significancia es $\alpha = 0.01$, pero que serían significativos con $\alpha = 0.05$. Estas anticorrelaciones adicionales, sugeridas por los datos pero no significativas con el α elegido, se representan con tres flechas color naranja en la **Figura** 2.33.

2.4.4. Estrategias de selección en embriones

El hallazgo de los 14 pares anticorrelacionados, como vimos, se basa en 404 individuos no relacionados entre sí, del dataset 1KGP-EUR. En 9 de esos pa-

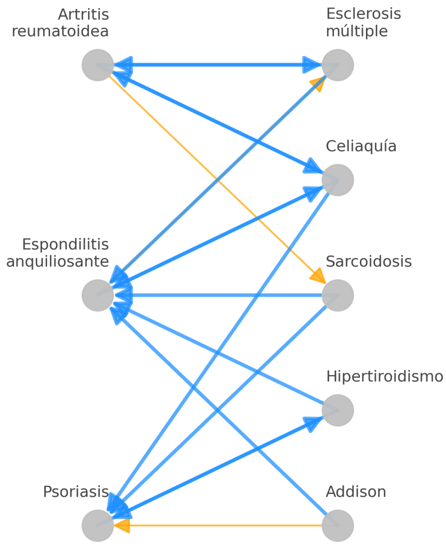


Figura 2.33: Grafo dirigido de las anticorrelaciones encontradas. En azul, relaciones significativas con $\alpha = 0.01$. En naranja, relaciones adicionales que cumplen $0.05 > P_{\text{adj}} > 0.01$.

res de enfermedades, hallamos también que en familias de PRS_1 midparental bajo (particularmente, las familias de lo que llamamos quintil Q_1), la elección del embrión que minimiza el PRS_1 (estrategia PBR) tiene por consecuencia un aumento significativo del riesgo absoluto (RA_2) en los embriones elegidos, cuando se los contrasta con embriones elegidos al azar en las mismas familias. Es decir, si una pareja de individuos con PRS promedio bajo elige a su embrión de menor riesgo poligénico, tiene probabilidades incrementadas de que el embrión elegido desarrolle otra enfermedad. El hallazgo es significativo tras una corrección FDR con nivel de significancia $\alpha = 0.01$.

En 6 de estos pares de enfermedades, el incremento resultó significativo con la estrategia PBR también al considerar familias del quintil Q_2, es decir, padres con PRS de valor bajo a intermedio. En 3 pares, el resultado es significativo hasta el quintil Q_3, en familias de PRS intermedio. En el resto de los quintiles, los resultados fueron mayoritariamente no significativos: no hubo aumentos del riesgo respecto de los embriones elegidos al azar.

Por otro lado, en ningún caso hubo incremento significativo del riesgo al usar la estrategia EAR, que se limita a excluir al embrión de máximo riesgo.

En la **Figura** 2.34 se resume el resultado de los 70 tests de la estrategia PBR (5 quintiles en 14 pares de fenotipos). El hallazgo principal se observa en los tests correspondientes al quintil Q_1 (la primera columna del *heatmap*). Se incluyen los valores de incremento promedio del riesgo absoluto en los embriones PBR respecto de los embriones *random*, expresados como puntos porcentuales ($100 \cdot \text{IRA}_2$). Este número se muestra sólo en los casos en los que rechazamos la hipótesis nula del test de Wilcoxon con un $P_{\text{adj}} < 0.01$, es decir, en los casos significativos tras corrección por FDR. En la **Figura** A.6 y **Figura** A.7 del Apéndice se muestran los P valores de los tests correspondientes.

El mayor valor de incremento de riesgo absoluto o IRA_2 se observa en el par anticorrelacionado L, de sarcoidosis y psoriasis. Aquí, $100 \cdot \text{IRA}_2 = 0.43$ significa que hay un aumento promedio de un poco menos de medio punto porcentual en los embriones seleccionados. El resto de los valores se reparten entre 0.09 y 0.34 puntos porcentuales de incremento, es decir, de riesgo *adicional* con respecto al riesgo de los embriones elegidos al azar en las familias correspondientes. En esta escala, $\text{IRA}_2 = 0$ significa riesgo equivalente a no utilizar el PRS para seleccionar embriones.

En la **Figura** 2.35 se resaltan los subgrupos de embriones de los que depende

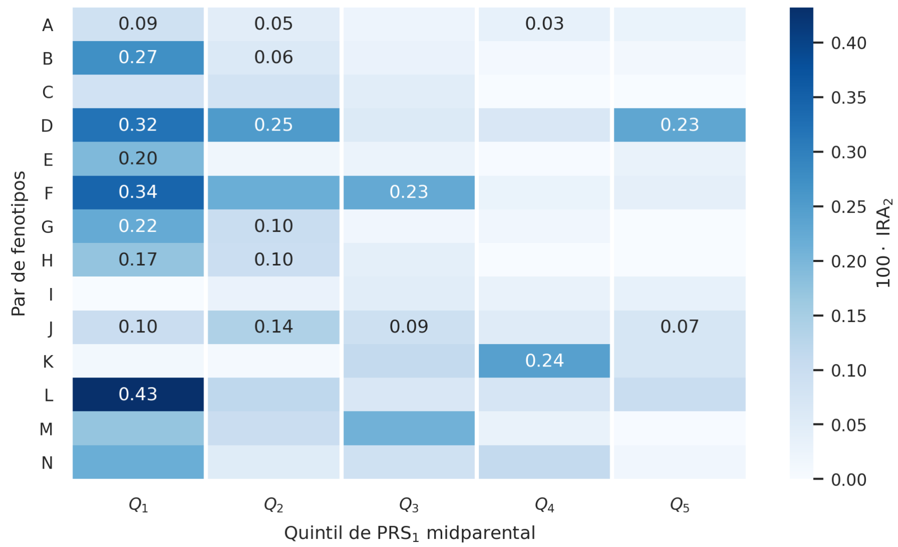


Figura 2.34: Por cada par de fenotipos (eje Y) y quintil de PRS midparental (eje X), se grafica la diferencia de RA medio en 50 familias entre los embriones PBR y los embriones *random*. Se muestran únicamente los valores de IRA donde el test de Wilcoxon fue significativo tras una corrección por FDR.

cada test. El panorama general de los 2 500 embriones simulados está dado por la nube gris de puntos de fondo. En cada fila, se resalta con naranja a los embriones disponibles para la selección en las familias del quintil correspondiente ($Q_1, \ldots, Q_5$) de PRS midparental.

Luego, se resalta en azul el conjunto de embriones seleccionados con alguna de las estrategias: PBR en la primera columna, EAR en la segunda, *random* en la tercera. Así, por ejemplo, en el gráfico de arriba a la izquierda se ve que al utilizar la estrategia de minimización del riesgo (PBR), los embriones seleccionados (azul) están más a la izquierda que el resto de los disponibles para la selección (naranja), pues dentro de cada familia son los que tienen el PRS mínimo.

Como ilustramos en la **Figura** 2.13, la reducción del PRS significa desplazarse a la izquierda en el eje X y conlleva un incremento del riesgo absoluto de la segunda enfermedad del par. En cada gráfico de la **Figura** 2.35 puede compararse el riesgo absoluto promedio de los embriones seleccionados (línea azul horizontal) con el riesgo absoluto promedio de los embriones *random* (línea negra horizontal). Cuanto más se separan estas líneas, mayor es el incremento de riesgo en los embriones seleccionados con respecto a los embriones elegidos al azar. Se resalta con línea roja punteada a los casos significativos según el test de Wilcoxon.

En la secciónA.2 se muestra el mismo gráfico para el resto de los pares de fenotipos.

En la **Figura** 2.36 se muestra el número necesario para hacer daño, NNH_2, correspondiente a los valores de IRA_2 (incremento de riesgo absoluto) de los tests significativos del primer quintil, siempre considerando la estrategia PBR. Como vemos, el número de embriones que deben ser seleccionados con la estrategia PBR para que ocurran enfermedades adicionales no es muy alto para las parejas en el quintil Q_1: entre 200 y 1 250, según el par de fenotipos. Recordemos que este número corresponde a la *cantidad de parejas* en las que la estrategia PBR debe aplicarse para obtener un "nuevo enfermo", no al número de embriones de FIV en una pareja particular.

Nótese que los números de NNH_2 ilustrados describen el problema cuando nos enfocamos en las parejas de PRS midparental en el primer quintil poblacional, Q_1. Sin embargo, como se explica en la sección A.1.6, sólo el 8 % de las parejas posibles de la población tienen puntajes midparentales en ese quintil. Por ende, si nos interesa conocer la cantidad de parejas *de la población general* (y no sólo de Q_1) que deben aplicar la estrategia PBR para que ocurra un evento

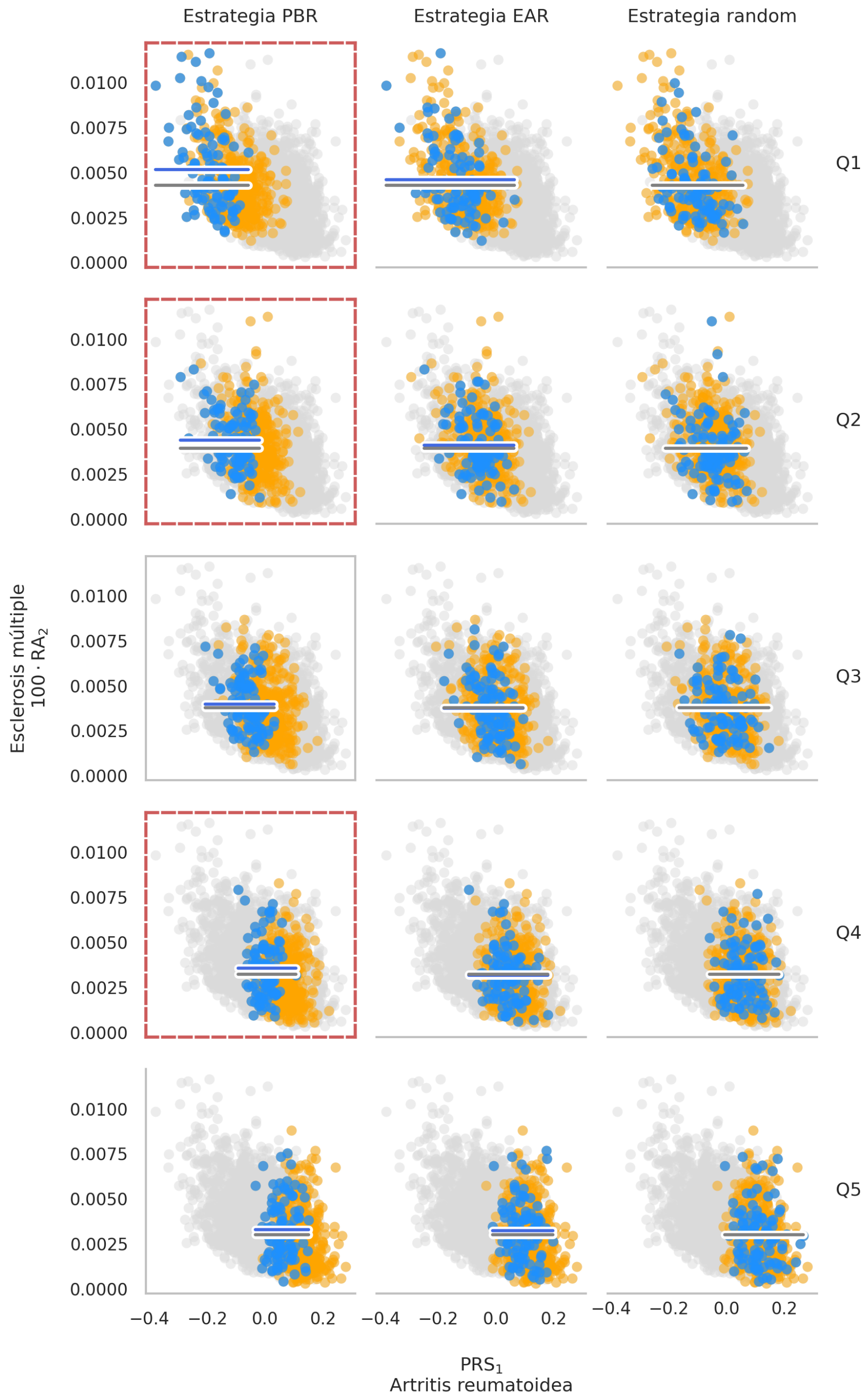


Figura 2.35: Par de fenotipos A: esclerosis múltiple y artritis reumatoidea. Se muestran en azul los embriones seleccionados en cada quintil de PRS midparental, según la estrategia, en naranja todos los embriones de familias en el quintil, y en gris el resto de los embriones. Explicación detallada en el cuerpo del texto. PBR: priorización del bajo riesgo, EAR: exclusión del alto riesgo, *random*: elección de un embrión al azar, RA: riesgo absoluto.

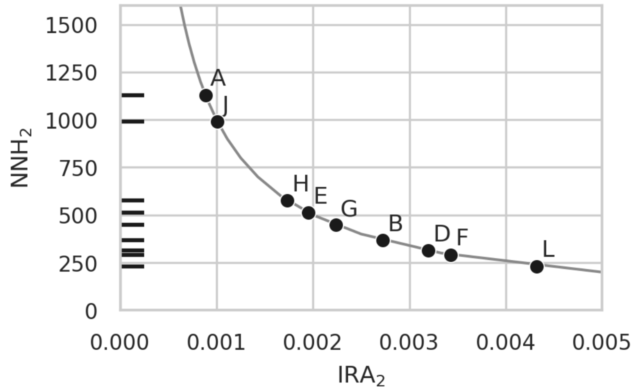


Figura 2.36: Se muestra la relación NNH $=$ IRA^{-1} con los valores de IRA$_2$ de los embriones PBR en parejas Q_1, para los nueve pares que pasaron el test de Wilcoxon.

adicional de enfermedad, tenemos que incrementar los NNH$_2$ con un factor de 100/8. Estos valores se muestran en la **Tabla** 2.3, junto a los incrementos de riesgo asociados y el P valor de cada test de Wilcoxon.

	Fenotipo 1	Fenotipo 2	P	P_{adj}	IRA$_2$	NNH$_2$ en Q_1	NNH$_2$ poblacional
L	Sarcoidosis	Psoriasis	6.73e-05	1.05e-03	0.0043	231	2 887
F	Esclerosis múltiple	Artritis reumatoidea	1.19e-05	3.33e-04	0.0034	291	3 637
D	Hipertiroidismo	Psoriasis	3.13e-04	3.13e-03	0.0032	312	3 900
B	Celiaquía	Espondilitis anquiliosante	2.52e-10	3.52e-08	0.0027	367	4 587
G	Hipertiroidismo	Espondilitis anquiliosante	2.57e-09	1.80e-07	0.0022	447	5 587
E	Sarcoidosis	Espondilitis anquiliosante	1.85e-05	3.70e-04	0.0020	511	6 387
H	Addison	Espondilitis anquiliosante	2.94e-06	1.37e-04	0.0017	576	7 200
J	Espondilitis anquiliosante	Celiaquía	5.01e-04	4.39e-03	0.0010	989	12 362
A	Artritis reumatoidea	Esclerosis múltiple	2.00e-04	2.34e-03	0.0009	1 129	14 112

Tabla 2.3: Pares con test de Wilcoxon significativo al utilizar la estrategia PBR en parejas del quintil Q_1 de PRS midparental. En la tercera columna se muestra el P valor del test y en la cuarta el P ajustado por FDR. IRA: incremento de riesgo absoluto, NNH es el número necesario para hacer daño.

2.5. Discusión

En este capítulo nos propusimos entender si el fenómeno común de la pleiotropía en el genoma humano y, en particular, la anticorrelación entre el componente genético de fenotipos complejos, presenta un problema para la práctica del PGT-P. Partimos de la idea de que la pleiotropía es tan común en el genoma de nuestra especie que, en consecuencia, los puntajes poligénicos estarían interrelacionados de tal manera que sería difícil realizar una selección segura de embriones basada en ellos.

En la primera parte del experimento, buscamos enfermedades en las que al seleccionar individuos de bajo PRS de una enfermedad, el grupo seleccionado tuviera riesgo incrementado significativamente de otra enfermedad. Encontramos 14 pares de enfermedades en los que esto ocurre, basándonos en el dataset 1KGP-EUR de 404 individuos no relacionados entre sí.

Este primer hallazgo nos permitió centrar la atención en un conjunto muy reducido de enfermedades autoinmunes. Sin embargo, no es seguro que el resultado basado en individuos no emparentados se aplique al escenario del PGT-P, pues durante la fertilización *in vitro* la selección se realiza entre embriones de una misma pareja, que comparten en media el 50 % de su genoma y por ende tienen mucha menos variabilidad genética que los individuos de la población general. Es decir, en los embriones hay un rango de PRS menor para seleccionar.

Así pues, en la segunda parte del experimento, formamos –para cada enfermedad de los pares anticorrelacionados– 500 parejas que cubrieran en conjunto el rango completo del PRS. Luego, simulamos 5 embriones por pareja, una cantidad estándar de embriones de PGT-A en la actualidad. Finalmente, realizamos un procedimiento de selección de un embrión para transferencia en cada pareja, con tres estrategias posibles a comparar.

Si bien 5 embriones es un número estándar en PGT-A, a futuro el número de embriones disponibles para PGT-P podría ser mayor, por ejemplo 10 embriones por pareja, dado que los pacientes de esta técnica no se limitarían a parejas con problemas de fertilidad. En este caso, creemos que los resultados serían similares a los presentados aquí o tal vez más pronunciados, pues con más embriones para elegir, habrá menos probabilidad de que el embrión *random* coincida con el embrión de PRS mínimo.

Tras simular las diferentes estrategias de selección, observamos que en 9 pares de enfermedades (de los 14 antes descriptos), elegir el embrión de PRS mínimo entre los 5 disponibles produce un incremento inesperado en otro riesgo, cuando se lo compara con elegir un embrión al azar, sin considerar el PRS. El problema se limita, mayormente, a las parejas de PRS midparental bajo, es decir, parejas donde alguno de los padres o ambos tienen un PRS_1 bajo. Una explicación posible de esto es la evidente relación no lineal entre PRS_1 y RA_2, que puede observarse en las regresiones graficadas en el segundo gráfico de las **Figuras** 2.14 a 2.27. En casi todos los casos, los primeros deciles de PRS_1 están asociados a un incremento más pronunciado del riesgo absoluto. Para valores intermedios y altos de PRS_1, la pendiente de la regresión decrece. Entre los individuos de puntajes poligénicos bajos se observan, incluso, numerosos *outliers* de riesgo mucho mayor a la media.

Denominamos aquí como "anticorrelación" el fenómeno donde los embriones de PRS mínimo tienen riesgo incrementado de otra enfermedad. Como vimos, en algunos casos la anticorrelación afecta a los embriones de parejas de PRS midparental intermedio, en los quintiles Q_2 y Q_3. Esto implica que el problema no se limita a los extremos de la distribución de PRS.

Por otro lado, también vimos que la anticorrelación no se observa si la estrategia se limita a excluir al embrión de máximo PRS. En estos casos, la selección de un embrión al azar entre $n-1$ en lugar de entre n embriones disponibles –que en nuestra simulación implica elegir uno entre cuatro en lugar de uno entre cinco– no modifica sustancialmente el riesgo de las enfermedades correlacionadas.

Recordemos que en los pares ordenados que hallamos, algunas enfermedades se repiten: por ejemplo, la espondilitis anquilosante participa en más de una correlación. Por ende, nuestro hallazgo de 9 pares anticorrelacionados en embriones implica que hay *6 enfermedades* (seis Fenotipos_1 en nuestros términos) en base a las que la selección de embriones de bajo PRS debería realizarse con cautela. Estas enfermedades son: sarcoidosis, esclerosis múltiple, hipertiroidismo, celiaquía, enfermedad de Addison, espondilitis anquilosante y artritis reumatoidea. Si existiera un interés en seleccionar embriones de bajo riesgo de alguna de ellas, una estrategia agresiva de minimización del riesgo podría tener el resultado negativo de aumentar el riesgo de otras enfermedades. Por otro lado, una estrategia menos agresiva que se limite a excluir el embrión de máximo riesgo no tendría ese resultado negativo. Sin embargo, esta última estrategia logrará una reducción menor del riesgo poligénico de la enfermedad en base a la que se realiza la selección.

El incremento de riesgo que describimos no es necesariamente grande *para una pareja particular*. Los incrementos de riesgo absoluto (IRA_2) que calculamos son menores a un punto porcentual. Por ejemplo, el embrión seleccionado con PRS mínimo de sarcoidosis tendría, en media, alrededor de medio punto porcentual extra de riesgo de psoriasis ($IRA_2 = 0.0043$ en la primera fila de la **Tabla** 2.3). Visto de otro modo, entre 200 embriones de estas características, sólo uno tendría psoriasis a lo largo de su vida, por haber sido seleccionado por PRS mínimo de sarcoidosis. A nivel individual, este número puede no parecer preocupante.

Desde una perspectiva poblacional y epidemiológica, sin embargo, ese número no es despreciable. Dijimos que el número necesario para hacer daño o NNH es una medida de cuántos tratamientos son necesarios para producir nuevos enfermos. Como vimos en la **Tabla** 2.3, el número de parejas que deben someterse a una selección basada en PRS no es alto a nivel poblacional, cuando se considera que las técnicas de PGT son utilizadas en miles de parejas por año. Si el PGT-P se populariza, además, como un método que no se limita a parejas con problemas de fertilidad, el número de tratamientos podría crecer mucho más. Así pues, si un IRA de medio punto porcentual equivale a un NNH de alrededor de 200 tratamientos necesarios para generar nuevos casos de enfermedad, este número es bajo (es decir, preocupante) si se piensa en términos poblacionales. Los otros NNH de la tabla citada están en el orden de 100 a 1000, de modo que la misma conclusión se aplica.

Sin embargo, estas conclusiones no implican que el programa del PGT-P sea inviable. Las conclusiones expuestas se basan en el escenario de selección de embriones basada en un PRS y "a ciegas" respecto de los otros puntajes. Conociendo la existencia de las anticorrelaciones descriptas aquí, lo más recomendable sería simplemente calcular ambos puntajes de cada par y seleccionar los embriones que minimicen ambos riesgos combinados. En esta dirección, desarrollos como el de [118] podrían ser el camino a seguir: un *index score* por embrión que pondere los riesgos de diferentes enfermedades con el impacto que cada una de ellas tiene en la expectativa o en la calidad de vida. Un puntaje de ese tipo, que resuma los diferentes *trade offs* dados por la arquitectura genética, permitiría evitar las consecuencias negativas antes mencionadas. Más aun, consideramos que siempre será preferible conocer estos *trade offs* mediante un PGT-P antes que realizar una selección al azar.

Estamos al tanto de al menos dos menciones de la pleiotropía como posible impedimento en la selección de embriones por PRS: [10] y [11]. En esos trabajos la pleiotropía no es analizada, sino que es mencionada al pasar como un problema posible que se suma a múltiples otros problemas y que en conjunto deberían alejarnos del programa del PGT-P. Asimismo, en el origen de esta tesis nos preguntamos si la pleiotropía en el genoma humano, un fenómeno conocido y descripto en múltiples ocasiones, volvería *impracticable* al PGT-P.

Sin embargo, creemos que nuestro hallazgo –y también lo que *no* hallamos– puede sugerir una interpretación contraria. Como vimos, los fenotipos involucrados en las anticorrelaciones descriptas comparten dos características: baja poligenicidad y concentración de efectos en una región genómica acotada, el CMH. Es decir, en el panorama de las enfermedades complejas, los resultados aquí expuestos parecieran sugerir que el problema de la correlación negativa como impedimento del PGT-P es un fenómeno marginal, circunscripto a un número pequeño de fenotipos particulares. Este resultado contradice, al menos en el frente de la pleiotropía, el horizonte catastrófico que postulan algunos oponentes del PGT-P.

El hecho de que los fenotipos involucrados en anticorrelaciones correspondan a fenotipos de baja poligenicidad merece especial atención y consideramos que no es casual. Es posible que la baja poligenicidad sea lo que posibilita las anticorrelaciones descriptas, puesto que la posesión de unos pocos alelos de efecto relativo muy grande puede determinar casi completamente la posición de un individuo en la distribución poblacional de PRS. Si esos alelos de efecto grande en el riesgo de una enfermedad tienen, en simultáneo, un efecto *protector* muy grande para otra enfermedad, entonces el hecho de que un individuo tenga muy bajo riesgo de un fenotipo y muy alto riesgo del otro depende de una combinación particular de pocos SNPs y, por ende, no demasiado improbable. En fenotipos más poligénicos, esta combinación simultáneamente extrema (muy bajo PRS_1, muy alto PRS_2) sería menos probable, pues dependería de un mayor número de SNPs. En este sentido, no parece casual que la espondilitis anquilosante se lleve la mayoría de las anticorrelaciones: es la enfermedad de mayor autoinmunidad ($\mathcal{V}_{\text{CMH}}$) y menor poligenicidad ($\mathcal{K}_{0.90}$) de todo el dataset. Se trata de un riesgo genético que con pocos SNPs del CMH se "activa" o "desactiva". Otro hecho que merece consideración es que los fenotipos sean enfermedades autoinmunes, con base genética en el CMH. Aquí, los hallazgos son consistentes con la hipótesis de pleiotropía modular [129].

Sin embargo, hay que tener cuidado con las conclusiones que conciernen a fenotipos de baja poligenicidad, porque cuanto más nos alejamos del modelo infinitesimal, el modelo aditivo y de propensión resulta menos adecuado [139]. En este punto cabe enfatizar que los fenotipos analizados tienen valores de $\mathcal{K}_{0.90}$ en general mayores a 40 SNPs y con 22 SNPs en el caso menos poligénico. Consideramos que el modelo de propensión no es inadecuado, pero sí ha de tomarse con más cuidado, pues estamos acaso en la frontera de su aplicabilidad.

Otro dato llamativo fue el contraste entre los hallazgos de la primera parte del capítulo (los 14 pares anticorrelacionados en población general) y los valores no significativos de correlación genética (r_g) para esos mismos pares tomados de [132]. Nuestro hallazgo, cabe recordar, no es una medida de correlación genética, sino algo mucho más específico: la existencia de un umbral de selección de individuos de bajo PRS que tiene por consecuencia un incremento del *riesgo absoluto* de una segunda enfermedad en el grupo seleccionado. Enfatizamos que el r_g estima una correlación entre entre PRSs, mientras que nosotros analizamos una relación entre PRS y probabilidad de enfermar. Además, la estimación de r_g se basa en los *summary statistics* de dos GWAS, mientras que nuestro hallazgo se basa en puntajes poligénicos calculados para genomas específicos (reales en el primer hallazgo, simulados en el segundo). Estos elementos podrían explicar la diferencia.

Como mencionamos, los valores de heredabilidad de SNPs y prevalencia fueron tomados de estimaciones realizadas por el laboratorio de Benjamin Neale [135] en base a la población de UK Biobank. Este biobanco captura un subconjunto de británicos sesgado hacia individuos de un rango etario particular y, en particular, *más saludables* que el promedio. Esto implica que los K utilizados para el cálculo de RA podrían ser ligeramente bajos.

Más aun, es probable que tanto los h_g^2 como los K de población británica no coincidan exactamente con los de la población europea en general, si bien esperamos que no se alejen demasiado. En este punto, sólo contamos con aproximaciones. Las correlaciones encontradas aquí, por ende, podrían cambiar si se modifican estos estimados.

En resumen, concluimos que la técnica de PGT-P puede tener consecuencias negativas si se aplica a ciegas sobre fenotipos de enfermedad autoinmune, en particular si la estrategia consiste en elegir el embrión de mínimo riesgo entre los disponibles. Sin embargo, el riesgo desaparece si se calculan en conjunto los puntajes anticorrelacionados para realizar una selección informada, o si se utiliza una estrategia menos agresiva de selección, que se limite a excluir el embrión de PRS máximo.

2.6. Direcciones futuras

Una continuación natural de este capítulo consiste en replicar el resultado con otros datasets de población europea. Sería especialmente deseable abordar el problema en población latinoamericana, pero esto presentará nuevos desafíos no triviales: la construcción de los PRSs basados en GWAS de ancestría diversa que, como vimos, escasean.

El acceso a nuevos datasets de población europea habilitaría varios experimentos posibles. En primer lugar, podemos proceder al cálculo relativamente sencillo de los PRSs ya construidos con LDAK, ahora en nuevos individuos, para verificar que el problema descripto aquí se replica en la población general (el primer hallazgo descripto). En segundo lugar, podríamos reajustar los valores β_i, es decir, reconstruir los PRSs ahora con un mayor número de individuos. Es una pregunta abierta cuánto cambiarían esos coeficientes y, por ende, los resultados río abajo tras el reajuste. Esta opción requiere más tiempo y cierta capacidad computacional. En tercer lugar, se podría avanzar en la simulación de embriones en base a los individuos de otro dataset y replicar el segundo hallazgo del capítulo. Este paso se acerca más al problema del PGT-P, pero también sería un desafío de alto costo computacional.

Nuevas estimaciones de h_g^2 y K permitirían un recálculo del riesgo absoluto. Nuevamente, es una pregunta abierta cuán sensibles son los hallazgos del presente capítulo a modificaciones de esos parámetros.

Otra dirección interesante y ortogonal a las mencionadas sería enfocarse en el detalle de la base genética de los pares anticorrelacionados, analizar la localización de los SNPs de cada puntaje e identificar las regiones pleiotrópicas, a la manera de [78]. La anotación de los SNPs pleiotrópicos que subyacen a la correlación podría arrojar nuevos datos para entender el grado de solapamiento o de varianza compartida por categoría genómica o por región, como aquí hicimos exclusivamente para el CMH.

Apéndice A

Material suplementario

A.1. Desarrollos auxiliares

A.1.1. BLD-LDAK y LDAK-BayesR-SS

Los puntajes poligénicos de este capítulo fueron construidos con el software LDAK siguiendo las recomendaciones de Zhang y col. [82].

En primer lugar, utilizamos el modelo BLD-LDAK, que describe la heredabilidad esperada de *cada SNP*, $\mathbb{E}\left[h_i^2\right]$, en función de su frecuencia alélica y 64 anotaciones funcionales que incluyen, entre otras características, información sobre la tasa de recombinación de la región, si es o no una región codificante, si hay un promotor en las cercanías, si es UTR o si es un sitio sensible a DNAsa I. Cada uno de estos datos recibe un peso específico, que es recalibrado durante el modelado (Tablas suplementarias 6 y 7 de [82]). El efecto principal de esta ponderación es incrementar el efecto de las variantes de mayor MAF y de regiones exónicas o conservadas (**Figura** A.1).

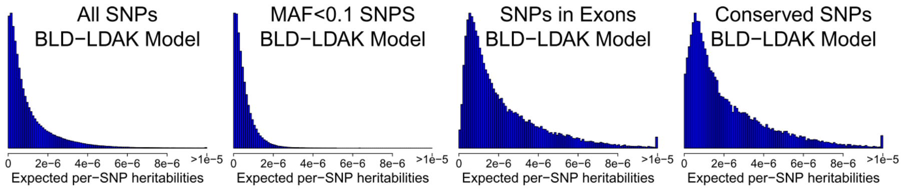


Figura A.1: Modelo BLD-LDAK de $\mathbb{E}\left[h_i^2\right]$ para todos los SNPs ("*All SNPs*") y para algunas subcategorías. Figura suplementaria 10 de [82]

En segundo lugar, el método SBayesR, en su implementación LDAK-BayesR-SS, parte de una distribución *a priori* de los efectos β_i como una mezcla de un punto de masa en cero y tres componentes normales centrados en cero y con varianzas que dependen de $\mathbb{E}\left[h_i^2\right]$ (es decir, cada SNP tiene su propio *prior*). Se contrasta este *prior* con otros tres modelos posibles en la **Figura** A.2. Los cuatro componentes de la mezcla son interpretables como las probabilidades de que se trate de un SNP con impacto nulo, bajo, moderado o alto. Las proporciones de cada componente, π_1, π_2, π_3 y π_4, son estimadas a partir de una grilla de valores posibles (Tabla suplementaria 8 de [82]). Todas las combinaciones posibles de los π_i ensayadas dan un total de 84 modelos en competencia, entre los que se elige el de predicción óptima como modelo definitivo. Los pesos elegidos representan la arquitectura genética particular de un fenotipo, i.e., su combinación particular de variantes de diferente efecto. La validación se basa en un método novedoso de estimación del poder predictivo de un puntaje utilizando únicamente *summary statistics* como *input*, introducido previamente en [142].

A.1.2. Rutina de MegaPRS

La rutina de MegaPRS que utilizamos para la construcción de los PRS fue tomada en gran medida del sitio web de LDAK. Antes de la construcción de cada

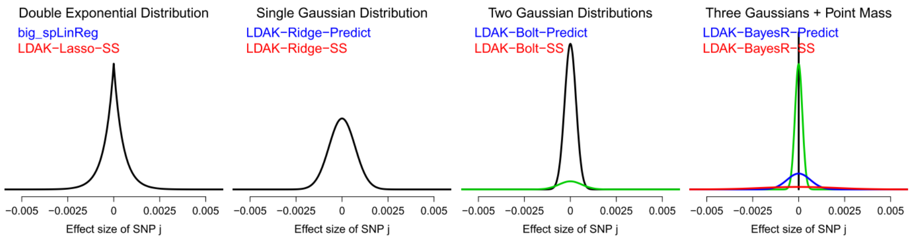


Figura A.2: Cuatro posibles distribuciones *a priori* de los β_i en LDAK. Los autores recomiendan la última, mezcla de 4 componentes. Figura suplementaria 11 de [82]

puntaje, estimamos la contribución de cada SNP a la heredabilidad esperada, $\mathbb{E}\left[h_i^2\right]$, con el comando `--sum-her`. Esto fue relaizado *para cada fenotipo* de UKB-GWAS (**Código** A.1). Usamos el modelo BLD-LDAK en este paso, que tiene en cuenta anotaciones funcionales de cada SNP además de su MAF. Como mencionamos en la introducción, este modelo es el recomendado por los autores del software para datos de SNPs en humanos. Utilizamos el parámetro `--cutoff 0.01` para que los SNPs que explican más del 1 % de varianza fenotípica no afecten el análisis de heredabilidad, según se sugiere en la guía asociada a este paso.

Código A.1: Estimación de la contirbución de cada SNP a la heredabilidad esperada con LDAK.

```
ldak-5.1
    --sum-hers
    --summary $GWAS_FILE
    --tagfile $TAG_FILE
    --matrix $COV_MATRIX
    --cutoff 0.01
    --check-sums NO
```

La construcción del PRS consiste en tres pasos. Primero, se estiman las correlaciones SNP-SNP. Véase el **Código** A.2.

Código A.2: Cálculo de las correlaciones SNP-SNP con LDAK.

```
ldak-5.1
    --calc-cors $OUT_FILE
    --bfile $EUR_GROUP1
    --window-kb 3000
```

Segundo, se simulan lo que los autores denominan *partial pseudo summary statistics* de *training* y de *testing* (**Código** A.3.). Se trata de un conjunto de *summary statistics* (en particular, de valores de β_i) que se distribuyen como si hubieran sido obtenidas con un GWAS del 90 % y del 10 % de las muestras del GWAS original, respectivamente. El primer conjunto se utiliza para entrenar 84 modelos alternativos, que corresponden a asignaciones diferentes de los pesos de los tres componentes gaussianos y del punto de masa en cero que conforman al modelo mixto de LDAK-BayesR-SS (**Código** A.4).

Código A.3: Simulación de las *partial pseudo summary statistics* de *training* y de *testing* con LDAK.

```
ldak-5.1
    --pseudo-summaries $GWAS_FILE
    --bfile $EUR_GROUP2
```

```
    --extract $GWAS_VARIANT_LIST
    --summary $GWAS_FILE
    --training-proportion 0.9
    --allow-ambiguous NO
```

Código A.4: Construcción de 84 modelos de BayesR basados en *summary statistics*.

```
ldak-5.1
    --mega-prs
    --model bayesr
    --bfile $EUR_GROUP1
    --extract $GWAS_VARIANTS_LIST
    --cors $SNP_CORRELATIONS
    --ind-hers $SNP_HERITABILITIES
    --summary $SUMMARY_STATS_FULL
    --summary2 $SUMMARY_STATS_TRAINING
    --window-cm 1
    --allow-ambiguous NO
```

Tercero, se testean los modelos entrenados con una evaluación *out of sample* de la precisión de cada modelo. Se estima R (la correlación entre fenotipo observado y predicho) utilizando el dataset de *partial pseudo summary statistics* de *testing* creado antes (**Código** A.5) y se elige el modelo que maximiza R. El conjunto definitivo de β_i ajustados son los efectos del modelo seleccionado, tras un reentrenamiento final que utiliza los *summary statistics* completos.

Código A.5: Selección del conjunto de β_i que mejor predice el fenotipo.

```
ldak-5.1
    --calc-scores $PHENO_LABEL
    --bfile $EUR_GROUP3
    --scorefile $EFFECTS_ALL_MODELS
    --power 0
    --final-effects $EFFECTS_SELECTED_MODEL
    --allow-ambiguous NO
    --save-counts YES
```

Los autores señalan que los estimadores de R son poco confiables si el panel de referencia es compartido entre pasos, de modo que dividimos a las 404 muestras de 1KGP-EUR en tres conjuntos disjuntos. En los fragmentos de código listados previamente, estos tres grupos se simbolizan con las variables `EUR_GROUP1`, `EUR_GROUP2` y `EUR_GROUP3`, que corresponden a subconjuntos del dataset EUR sin individuos en común, con 142, 130 y 132 individuos respectivamente, repartidos al azar.

Para el tercer paso, el argumento `--extract highld/genes.predictors.used` fue incluido, de modo que el programa excluya los SNPs contenidos en regiones de alto desequilibrio de ligamiento cuando calcula la precisión de los modelos construidos con *pseudo summary statistics*. Esta exclusión está recomendada en las guías de LDAK, pero sólo afecta a la evaluación de los modelos y no al PRS final, que incluye esas regiones.

Tras obtener el conjunto de β_i óptimos para cada fenotipo, calculamos los PRSs de los 404 individuos de 1KGP-EUR (**Código** A.1.2).

```
ldak
    --calc-scores $OUT_LABEL
    --bfile $GENOS_LABEL
    --scorefile $EFFECTS_SELECTED_MODEL
    --power 0
```

A.1.3. Supuesto de independencia de los G_i en el cálculo de varianza del PRS

En (2.8) asumimos que los G_i son independientes. No es claro que esto sea así, pues el conjunto de SNPs del dataset 1KGP-EUR no fue *pruneado* por LD. Sin embargo, esperamos que esto no sea un problema. Obsérvese que si los G_i no son independientes, entonces la varianza es:

$$\begin{aligned}
\operatorname{var}(S) &= \sum_{i=1}^{n} 2\beta_i^2 f_i(1-f_i) + 2\sum_{i<j} \operatorname{cov}\left(\beta_i G_i, \beta_j G_j\right) \\
&= \sum_{i=1}^{n} 2\beta_i^2 f_i(1-f_i) + 2\sum_{i<j} \beta_i\beta_j r_{ij}\sqrt{2f_i(1-f_i)}\sqrt{2f_j(1-f_j)}
\end{aligned}$$

donde r_{ij} es la correlación entre los SNPs i y j. LDAK se encarga de ajustar los valores de β para compensar el LD entre variantes próximas, de modo que es de esperar que, dado un par de variantes (i, j) correlacionadas, si el β_i es relativamente grande, el β_j será anulado o llevado a valores muy cercanos a cero. En esos casos, el término correspondiente de la covarianza se anula. Esto se aplica para todos los pares correlacionados, de modo que

$$\operatorname{var}(S) \approx \sum_{i=1}^{n} 2\beta_i^2 f_i(1-f_i)$$

Este es el valor que utilizamos.

A.1.4. Nota sobre el test *t* de diferencia de medias

Para encontrar el conjunto de 14 pares de fenotipos anticorrelacionados realizamos un test *t* de diferencia de medias de riesgo absoluto entre dos muestras independientes: los individuos "seleccionados" (qPRS subumbral) y los "no seleccionados" (qPRS supraumbral). Sin embargo, la diferencia que nos interesa mostrar como significativa es la del riesgo en los seleccionados *respecto de la población general*. Mostramos aquí que el test es equivalente.

Llamemos X al riesgo absoluto, $u \in (0, 1)$ a la proporción de individuos seleccionados, de modo que $1 - u \neq 0$ es la proporción de no seleccionados. Sean $\bar{X}_u$ y $\bar{X}_v$ el riesgo en los dos subgrupos respectivamente. Se cumple que el promedio general es

$$\bar{X} = u\bar{X}_u + (1-u)\bar{X}_v$$

Queremos mostrar que el promedio de ambos subgrupos difiere si y sólo si el promedio de los seleccionados difiere del promedio general:

$$\bar{X}_u \neq \bar{X}_v \;\Leftrightarrow\; \bar{X}_u \neq \bar{X}$$

Veamos:

($\Rightarrow$) Supongamos que ambos subgrupos difieren, $\bar{X}_u \neq \bar{X}_v$. Entonces:

$$\begin{aligned} \bar{X}_u &\neq \bar{X}_v \\ (1-u)\bar{X}_u &\neq (1-u)\bar{X}_v \\ u\bar{X}_u + (1-u)\bar{X}_u &\neq u\bar{X}_u + (1-u)\bar{X}_v \\ \bar{X}_u &\neq \bar{X} \end{aligned}$$

($\Leftarrow$) Supongamos que la media de los seleccionados difiere de la media general, $\bar{X}_u \neq \bar{X}$. Entonces:

$$\begin{aligned} \bar{X}_u &\neq \bar{X} \\ \bar{X}_u &\neq u\bar{X}_u + (1-u)\bar{X}_v \\ \bar{X}_u - u\bar{X}_u &\neq (1-u)\bar{X}_v \\ (1-u)\bar{X}_u &\neq (1-u)\bar{X}_v \\ \bar{X}_u &\neq \bar{X}_v \end{aligned}$$

Adaptado de stats.stackexchange (acceso: 10 abril 2022).

A.1.5. Correlaciones genéticas según la regresión de LD *score*

En la **Tabla** A.1 se listan valores de correlación genética r_g entre los fenotipos de cada uno de los pares anticorrelacionados, calculado en [132] con el método de regresión de LD *score* [77, 143], junto al error estándar y el P valor de la estimación. Nótese que en todos los casos la estimación no es significativa, es decir, no puede afirmarse con confianza estadística que la correlación sea distinta de cero. Recordemos, sin embargo, que estas correlaciones genéticas no toman en cuenta la conversión del PRS a valores de riesgo absoluto. En el caso del par H, no hay datos disponibles.

Par	r_g (err)	P valor
A	0.38 (0.26)	0.146
B	-0.24 (0.47)	0.603
C	0.29 (0.19)	0.131
D	-0.01 (0.15)	0.946
E	0.57 (0.46)	0.211
F	0.38 (0.26)	0.146
G	0.21 (0.39)	0.592
H	–	–
I	0.15 (0.52)	0.777
J	-0.24 (0.47)	0.603
K	0.29 (0.19)	0.131
L	0.19 (0.17)	0.243
M	0.18 (0.18)	0.312
N	-0.01 (0.15)	0.946

Tabla A.1: Correlaciones genéticas entre fenotipos de cada par calculadas con *LD score regression*, tomadas de [132]. err: error estándar.

A.1.6. La distribución de los qPRS midparentales

En los tests del capítulo 2, nos referimos en numerosas ocasiones al qPRS midparental de un embrión como el qPRS promedio de sus padres. Es decir, si el padre tiene un qPRS = 0.25 (i.e. su puntaje supera al 25 % de los puntajes de la población) y su madre tiene un qPRS = 0.30, el embrión tiene un qPRS midparental de 0.275.

Ahora bien, debemos tener cierto cuidado en la interpretación de este valor de qPRS como cuantil. Lo que quiere decir ese 0.275 es que en la distribución de *puntajes de individuos*, el PRS promedio de los padres supera al PRS del 27.5 % de los individuos. Lo que ese 0.275 de qPRS midparental *no* quiere decir es que el 27.5 % *de las parejas* tengan un qPRS midparental menor. Es decir, el cuantil está en términos de individuos, no de parejas.

Aquí debemos enfatizar que la distribución de puntajes individuales es muy diferente de la distribución de promedios del puntaje en pares de individuos. A partir de las 40804 parejas posibles entre individuos del dataset 1KGP-EUR, observamos que en los fenotipos de pares anticorrelacionados la proporción de parejas cuyo PRS midparental está en el primer quintil es de alrededor del 8 % –y no del 20 %, como erróneamente podría sugerir nuestra nomenclatura de quintiles $Q_1, \ldots, Q_5$. Se ve esto en la **Figura** A.3, que incluye los quintiles midparentales de todas las parejas posibles del dataset.

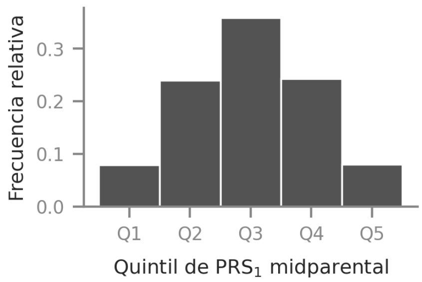


Figura A.3: Frecuencia relativa de parejas por quintil de PRS$_1$ midparental.

La razón de este fenómeno es que es más difícil que una pareja tenga un puntaje midparental en los quintiles extremos Q_1 o Q_5, pues en general es necesario para ello que *ambos padres* pertenezcan a ese quintil, algo mucho menos probable a que uno solo pertenezca al quintil extremo.

A.2. Figuras suplementarias

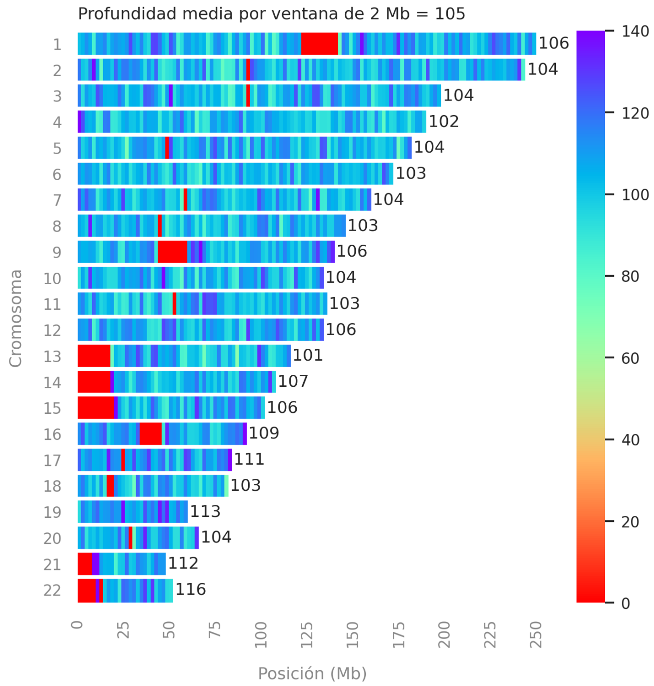


Figura A.4: Profundidad del Panel 100K SNPs en los 22 autosomas, en ventanas de 2 Mb, calculado en base a los genotipos de la muestra S1. Al lado de cada cromosoma se muestra el promedio de profundidad media en ese cromosoma particular.

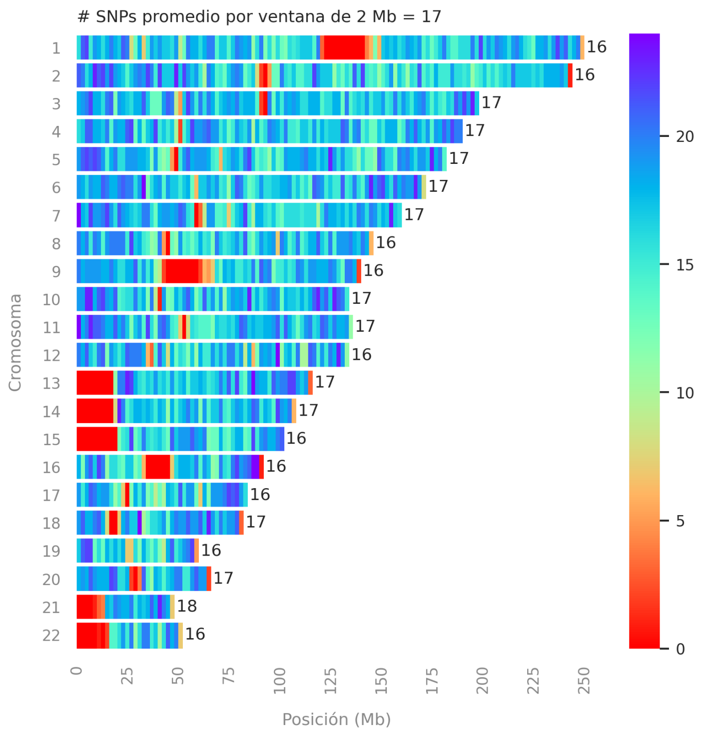


Figura A.5: Distribución de los SNPs del Panel 100K SNPs en los 22 autosomas, en ventanas de 2 Mb. Al lado de cada cromosoma se muestra el promedio de SNPs por ventana en ese cromosoma particular, excluyendo regiones de centrómeros y telómeros.

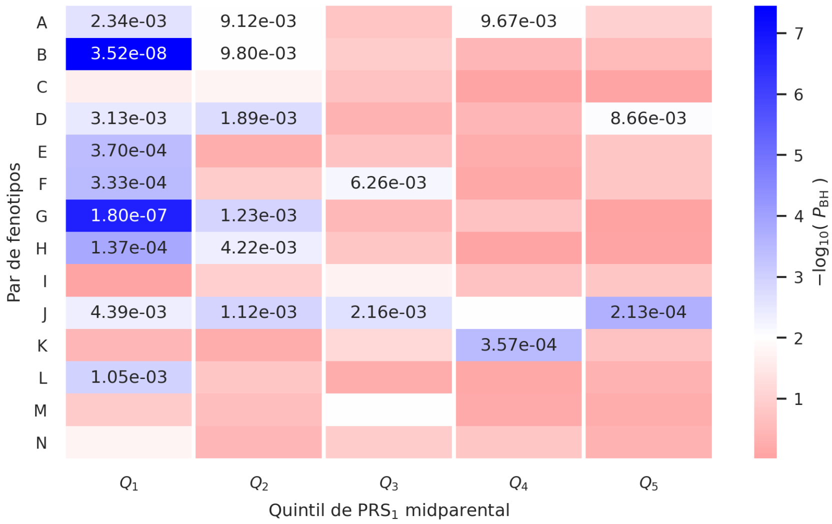


Figura A.6: P valores de los tests de Wilcoxon con la estrategia PBR. Se incluye el valor sólo en los casos nominalmente significativos con $\alpha = 0.01$.

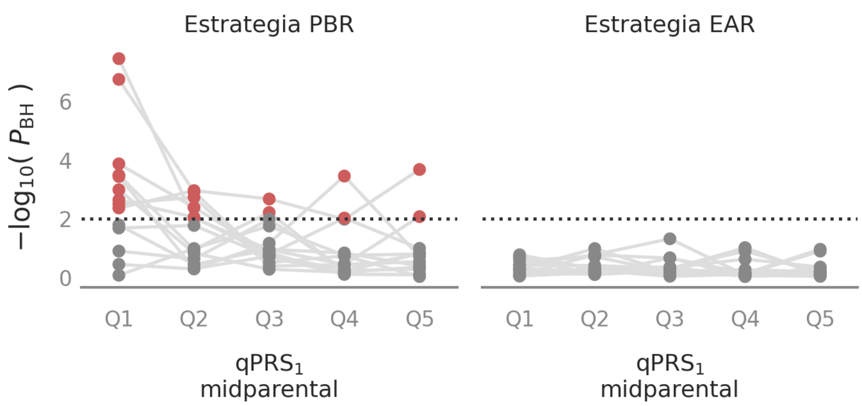


Figura A.7: $-\log_{10}(P_{BH})$ de los tests de Wilcoxon con las estrategias PBR y EAR. La línea punteada corresponde a $P_{BH} = 0.01$.

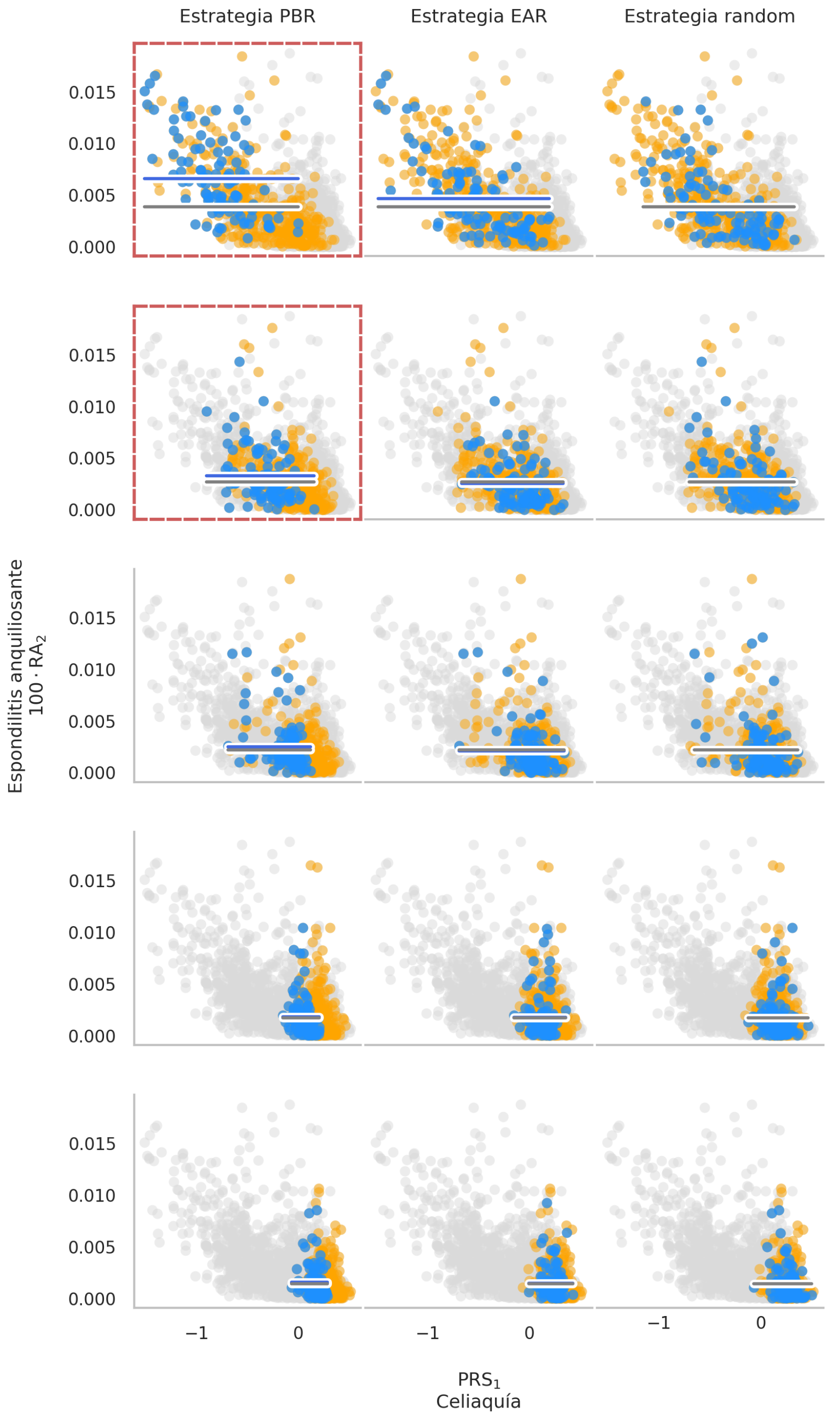


Figura A.8: Embriones del par de fenotipos B. Ver la expicación en el epígrafe de la **Figura** 2.35.

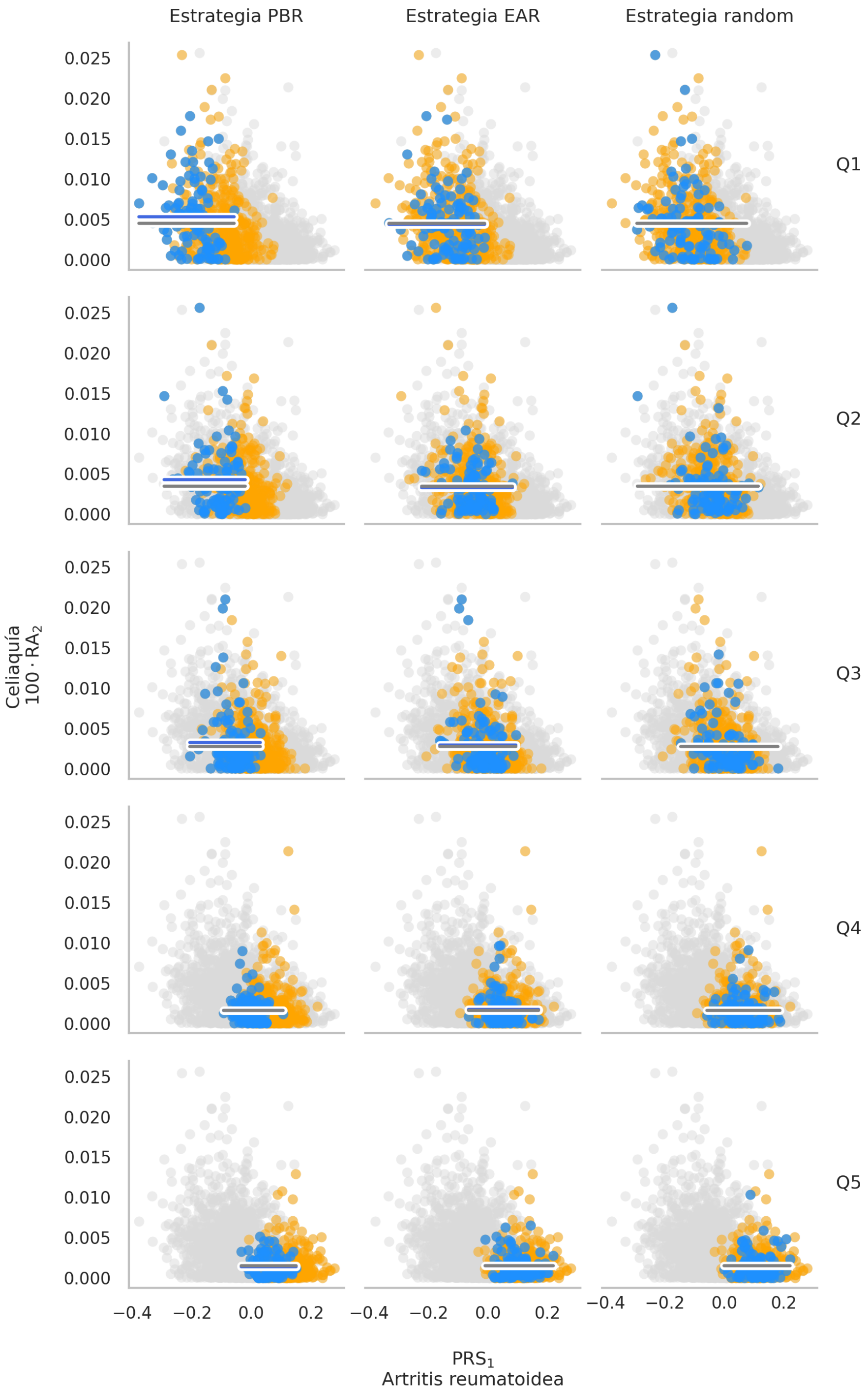


Figura A.9: Embriones del par de fenotipos C. Ver la expicación en el epígrafe de la **Figura** 2.35.

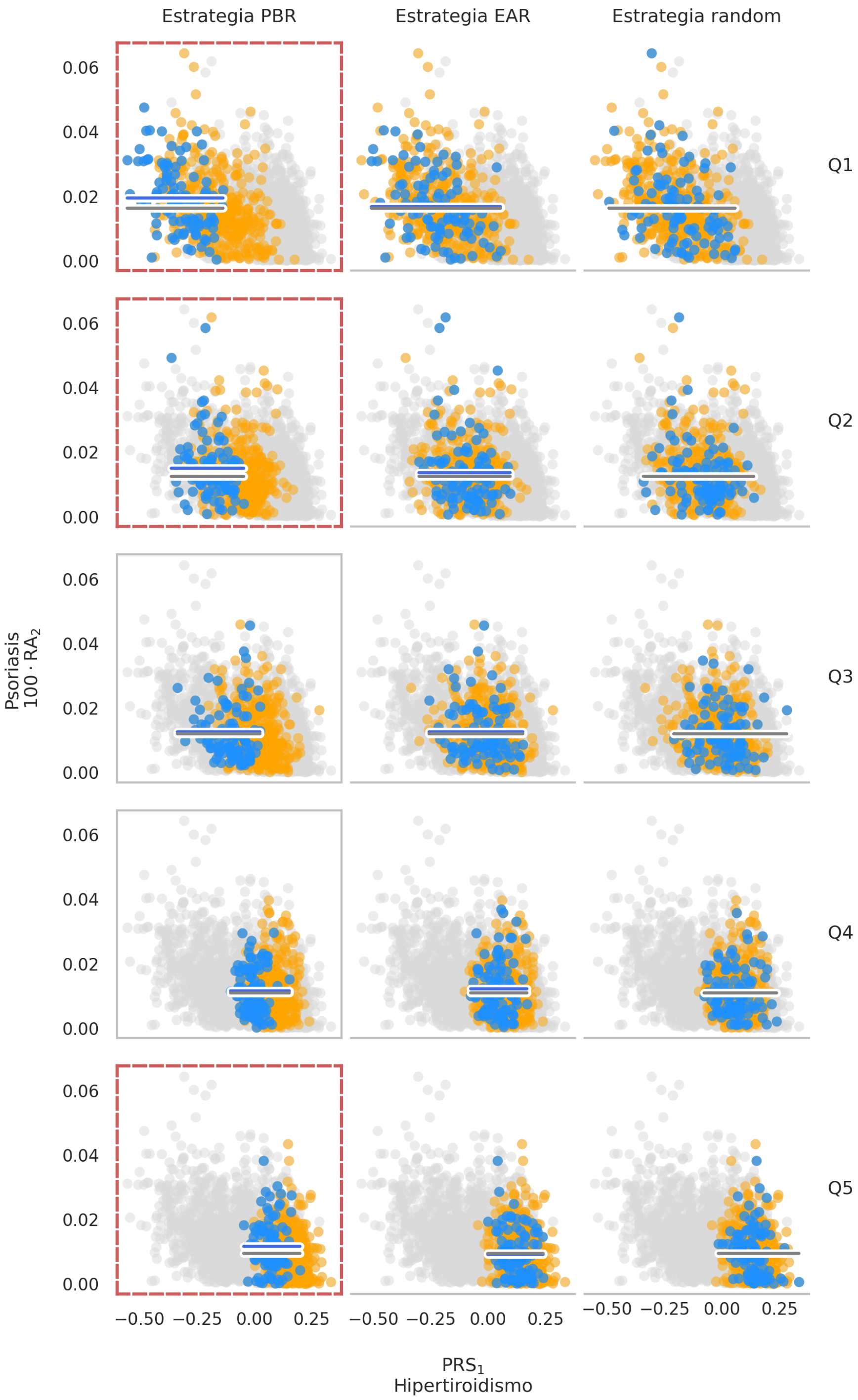


Figura A.10: Embriones del par de fenotipos D. Ver la expicación en el epígrafe de la **Figura** 2.35.

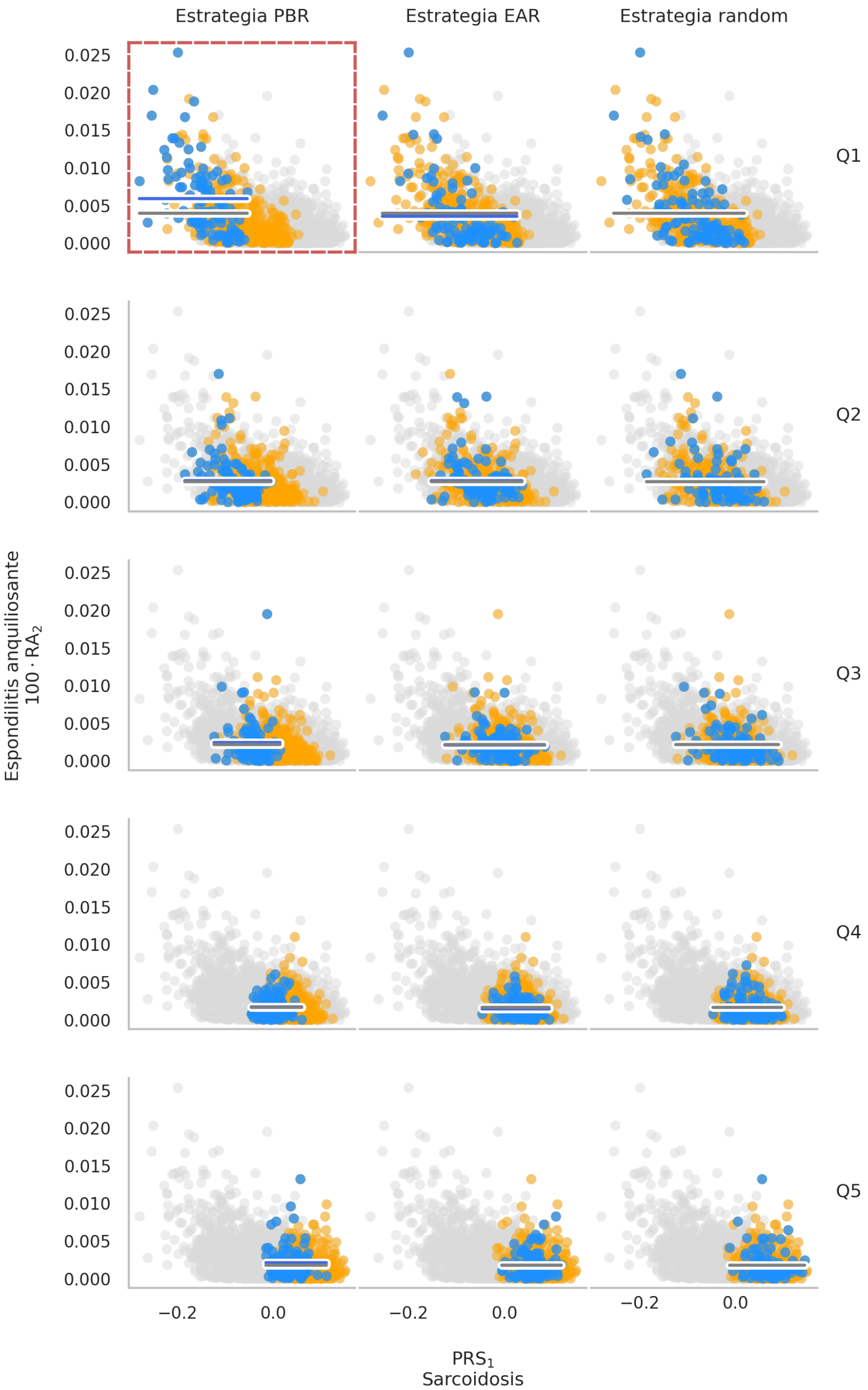


Figura A.11: Embriones del par de fenotipos E. Ver la expicación en el epígrafe de la **Figura** 2.35.

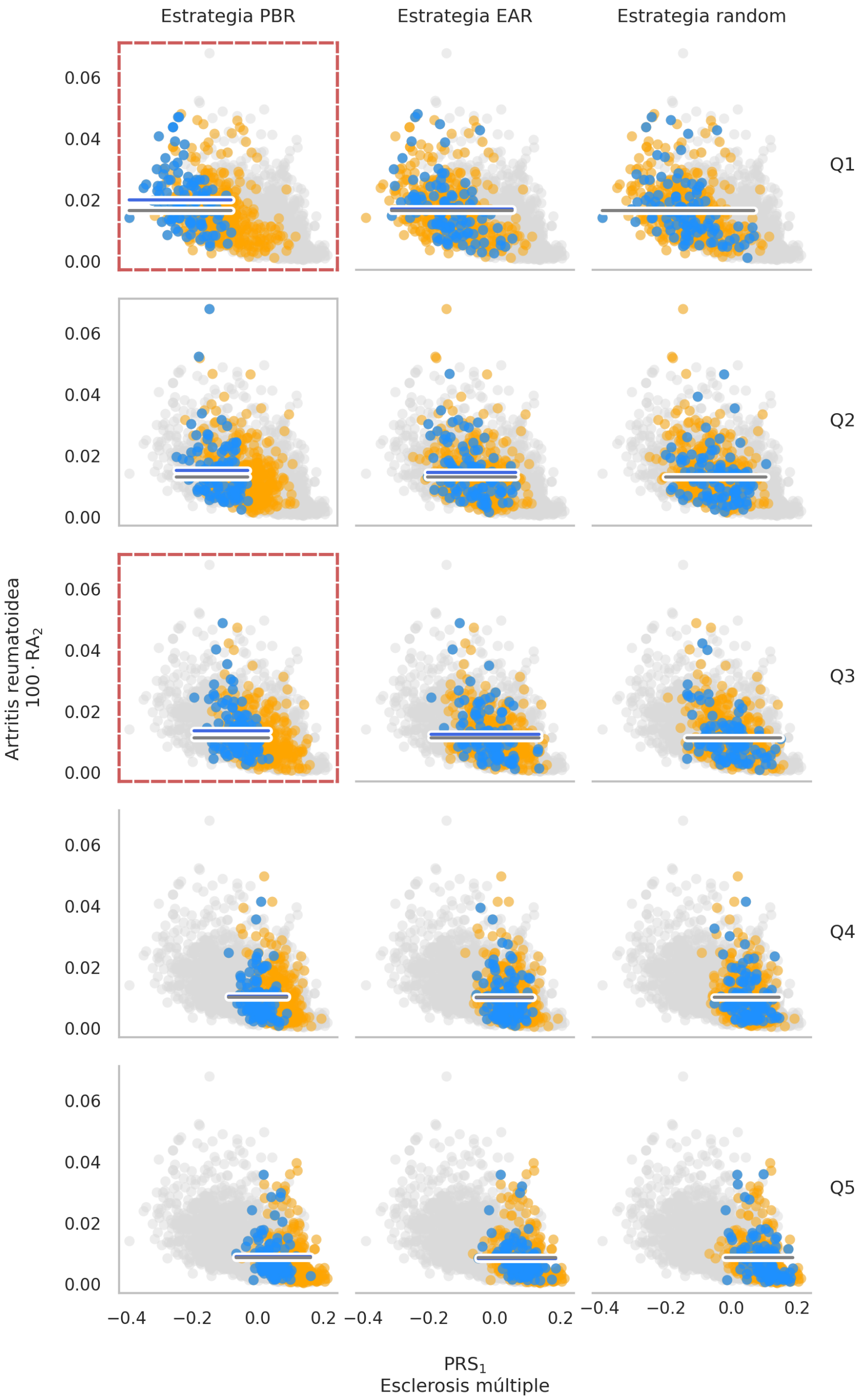


Figura A.12: Embriones del par de fenotipos F. Ver la expicación en el epígrafe de la **Figura** 2.35.

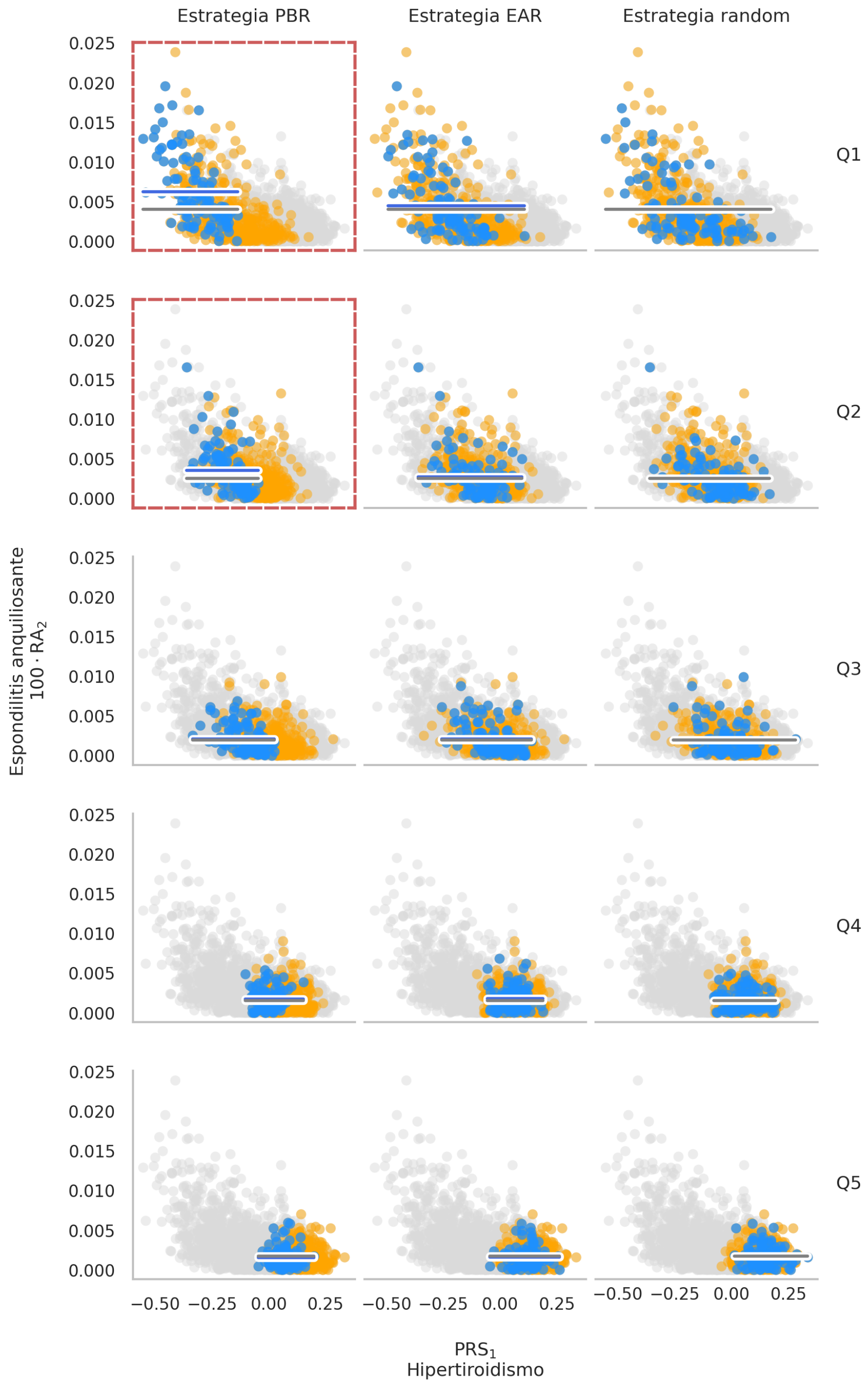


Figura A.13: Embriones del par de fenotipos G. Ver la expicación en el epígrafe de la **Figura** 2.35.

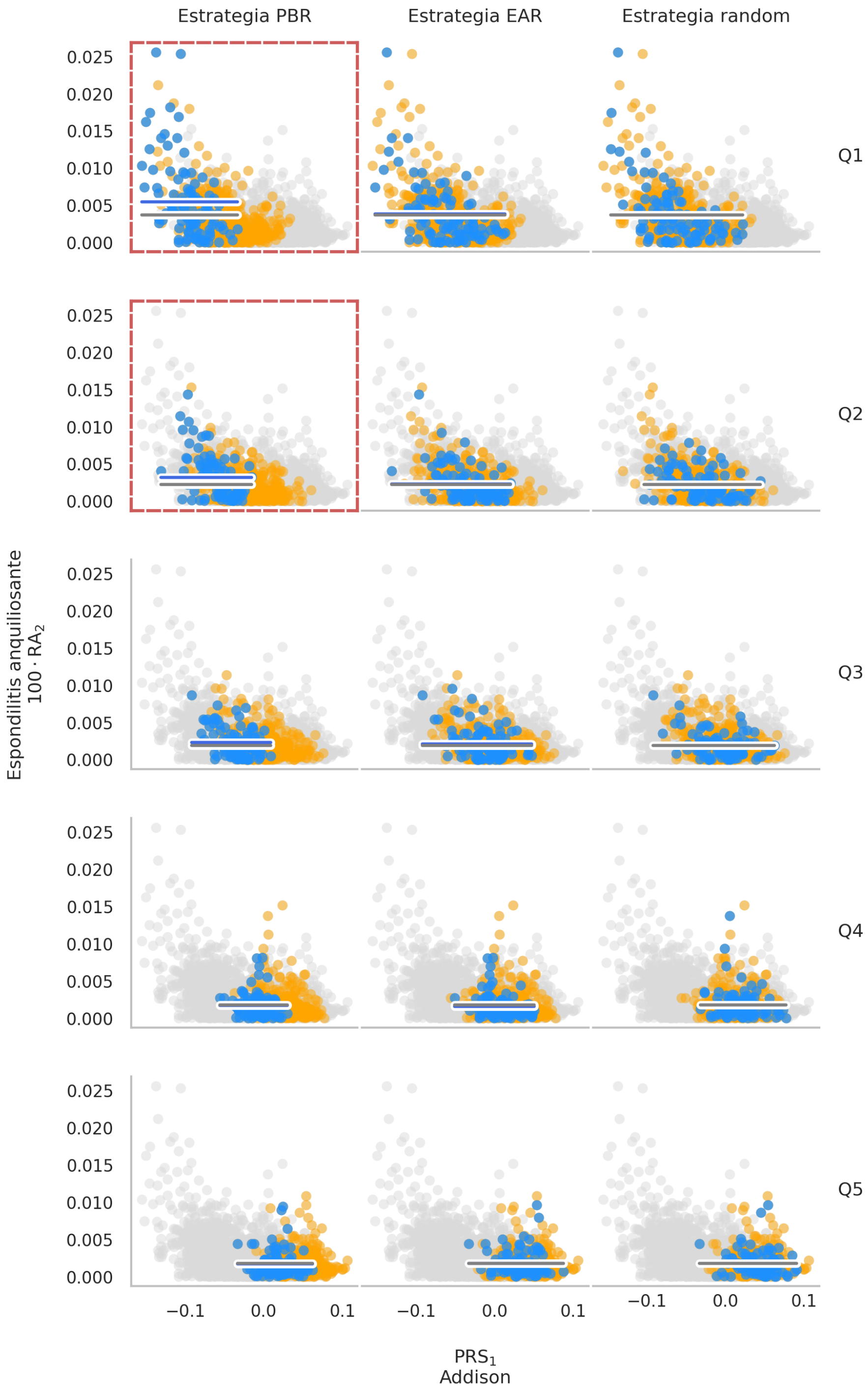


Figura A.14: Embriones del par de fenotipos H. Ver la expicación en el epígrafe de la **Figura** 2.35.

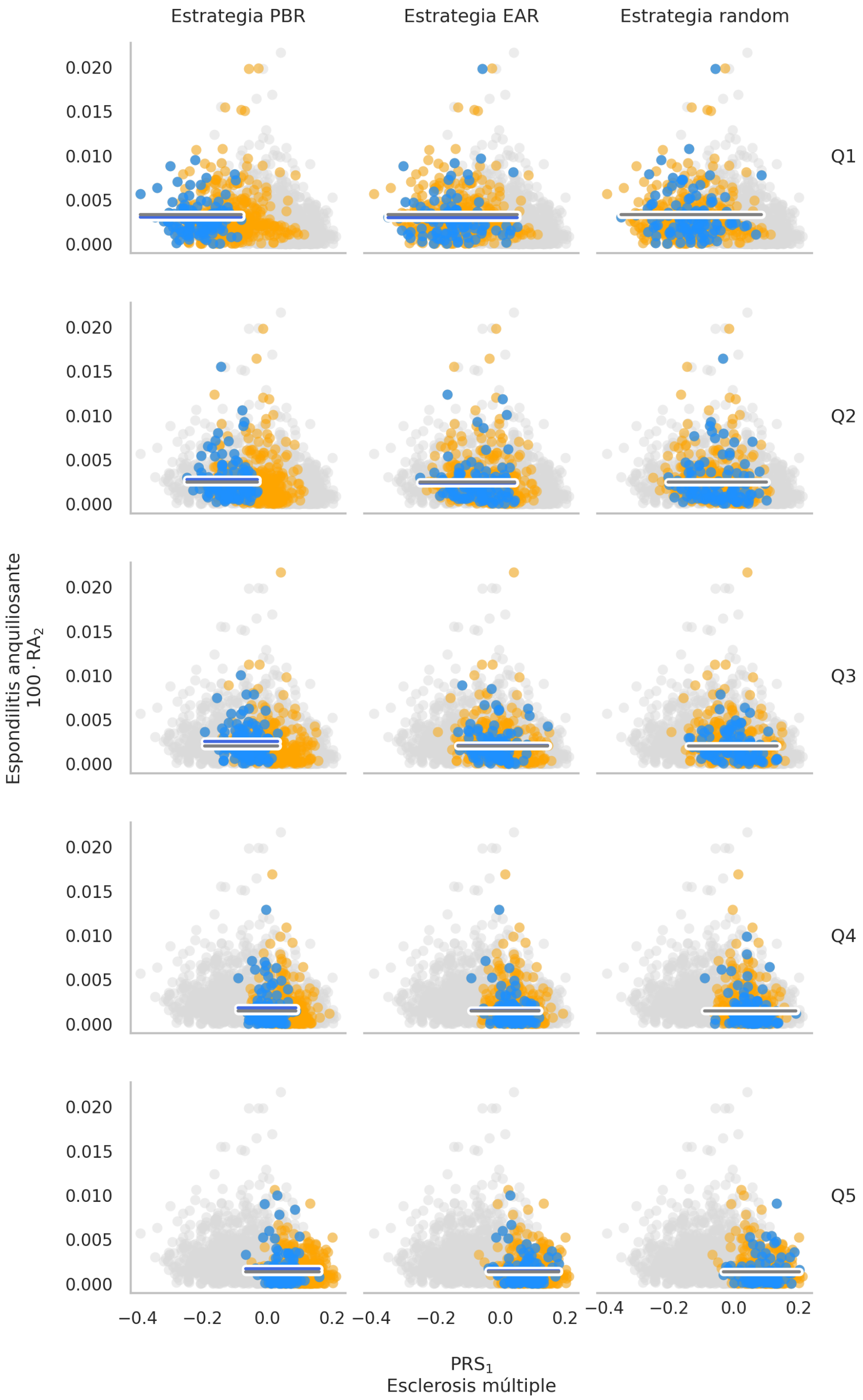


Figura A.15: Embriones del par de fenotipos I. Ver la expicación en el epígrafe de la **Figura** 2.35.

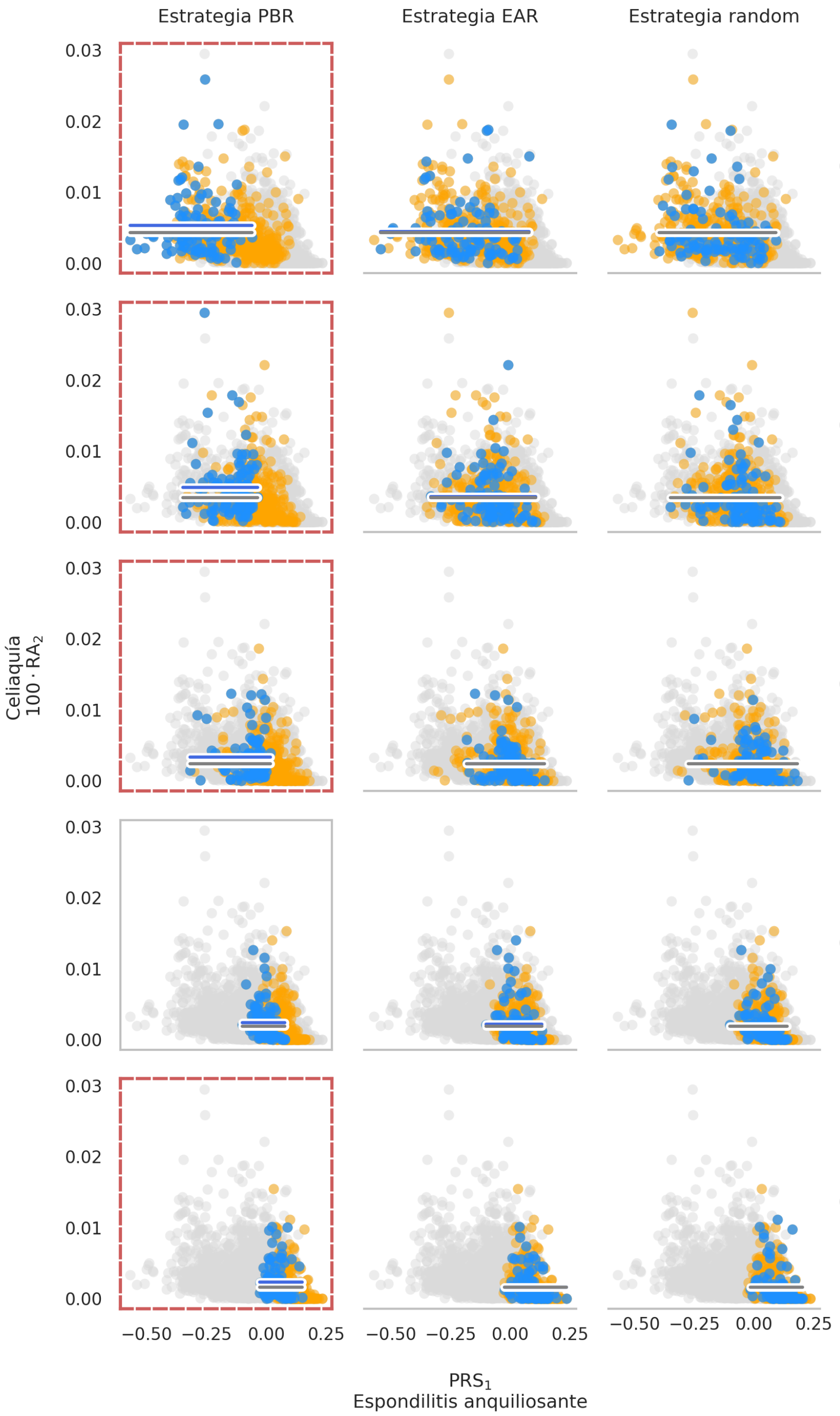


Figura A.16: Embriones del par de fenotipos J. Ver la expicación en el epígrafe de la **Figura** 2.35.

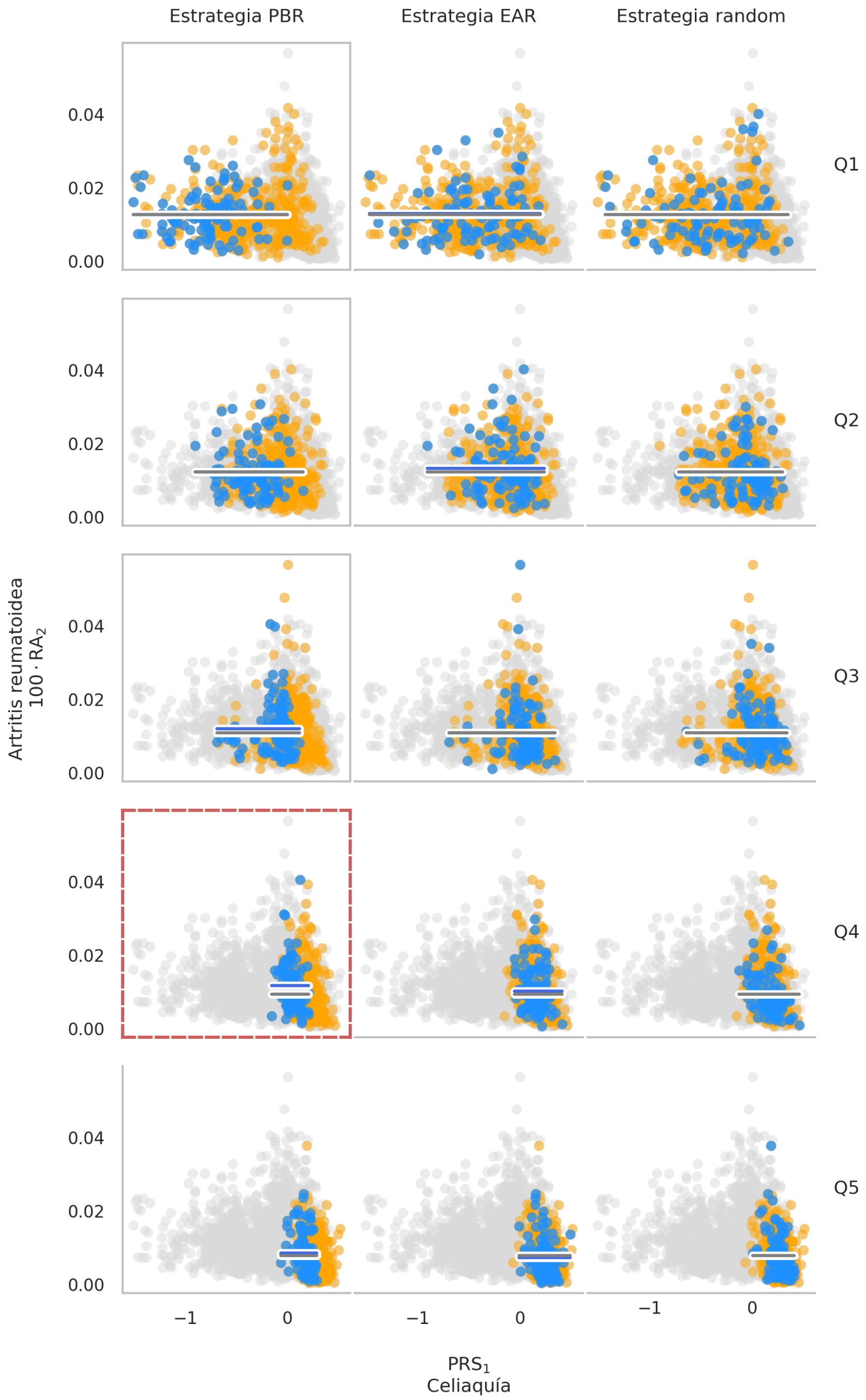


Figura A.17: Embriones del par de fenotipos K. Ver la expicación en el epígrafe de la **Figura** 2.35.

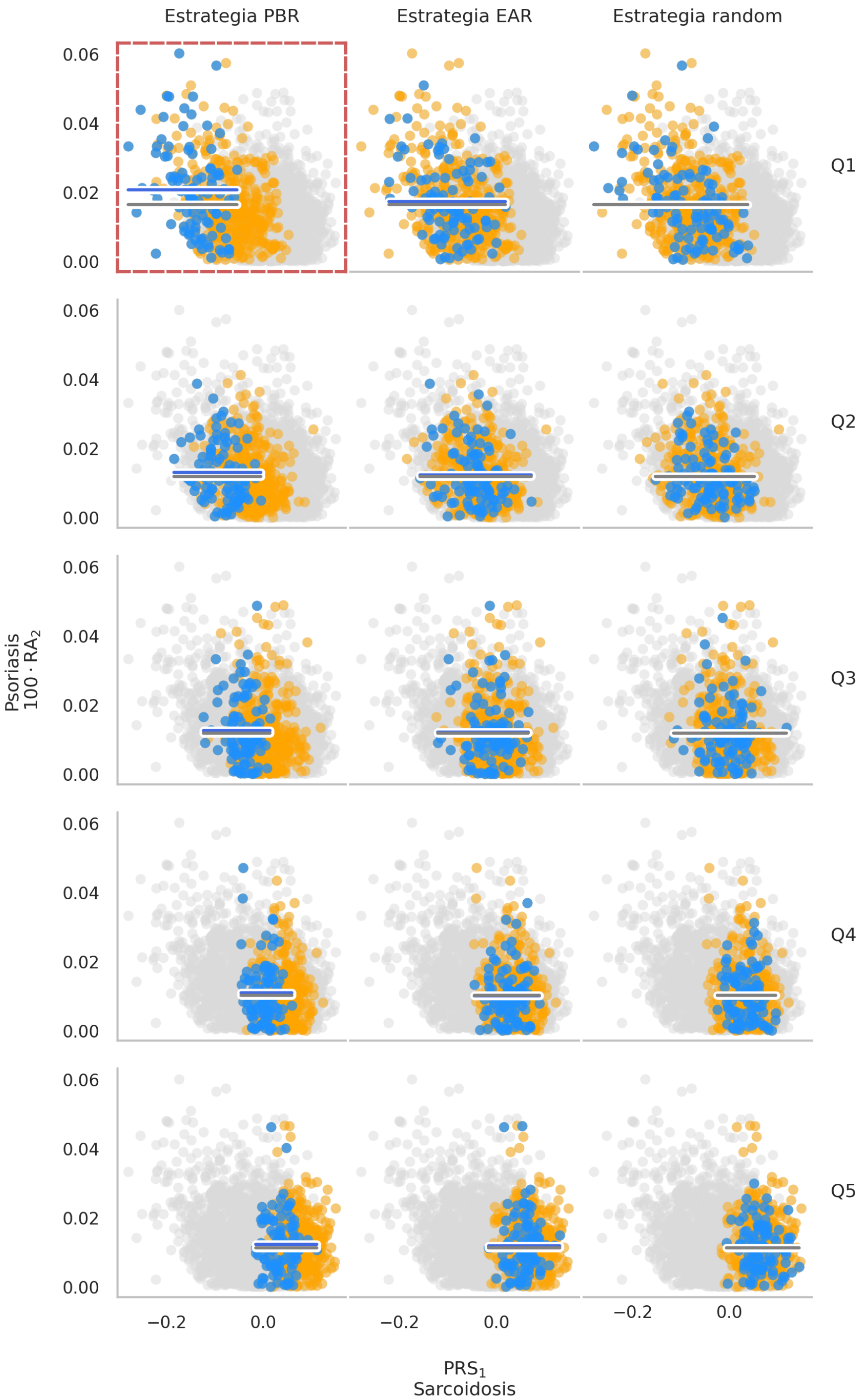


Figura A.18: Embriones del par de fenotipos L. Ver la expicación en el epígrafe de la **Figura** 2.35.

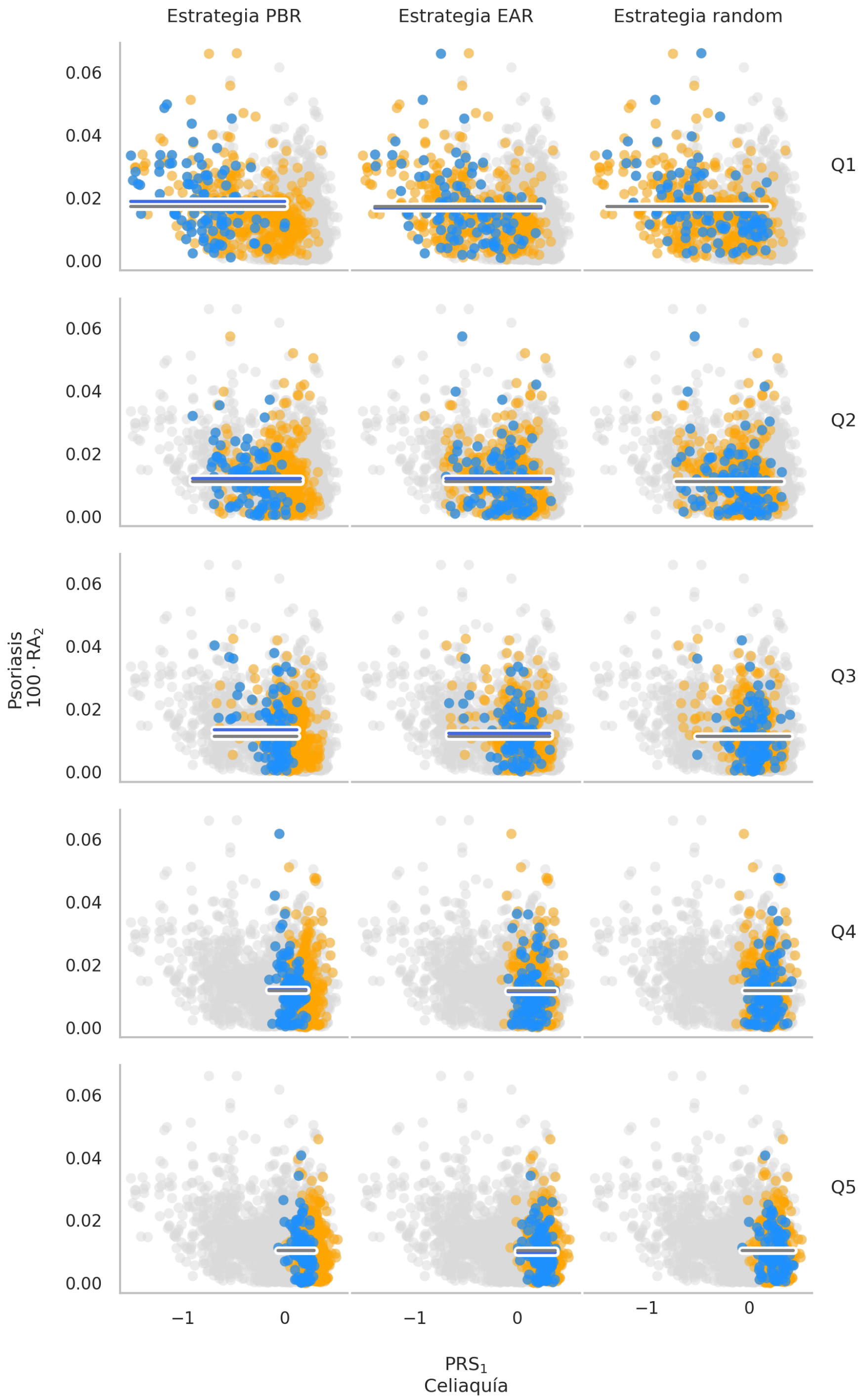


Figura A.19: Embriones del par de fenotipos M. Ver la expicación en el epígrafe de la **Figura** 2.35.

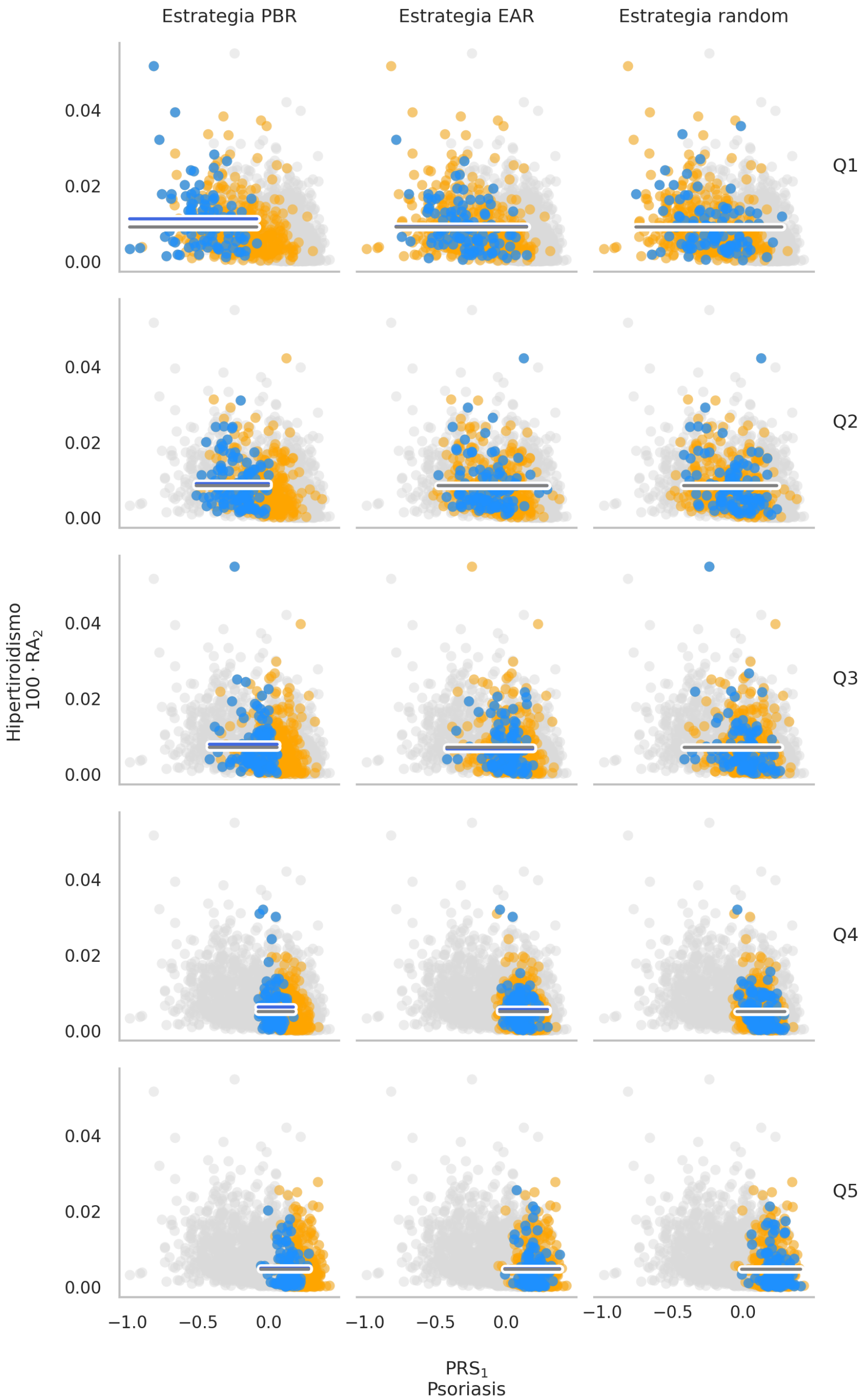


Figura A.20: Embriones del par de fenotipos N. Ver la expicación en el epígrafe de la **Figura** 2.35.

Glosario

Todas las definiciones dadas a continuación son para el uso particular de esta tesis y no pretenden agotar el sentido de los términos, que podrían recibir definiciones más generales en otros contextos.

1KGP-EUR dataset de genotipos de 404 individuos de ancestría europea del Proyecto 1000 Genomas. 75
aCGH *array Comparative Genomic Hibridization* 4
aneuploidía condición en la que el número de cromosomas no es un múltiplo exacto del número haploide de la especie (23 en humanos). Se suele usar el sufijo ”sómico” para estos casos: nulisómico ($2n-2$), monosómico ($2n-1$), trisómico ($2n+1$). 2, 12
aneuploidía sexual aneuploidías de los cromosomas sexuales, conjunto de condiciones en las que el número de cromosomas sexuales es anormal. 54
autosoma cualquiera de los cromosomas no sexuales. 12
BAC conteo del alelo B 23
bache región mayor a 500 Kb donde no hay SNPs del panel analizado. 17
BAF frecuencia del alelo B 5, 9
biopsia de trofectodermo técnica embriológica utilizada para recolectar entre 2 y 10 células del trofectodermo. 14
blastocisto estadio del embrión del mamífero en el momento de su implantación en la pared del útero. En humanos, corresponde al día 5-7 del desarrollo del cigoto. 13
cariotipo descripción del conjunto de los cromosomas presentes en una célula, organismo o especie. 12
casos y controles terminología utilizada en los GWAS para referirse a individuos enfermos y sanos, respectivamente. 67
CMH complejo mayor de histocompatibilidad 74
CNV variación del número de copias 5
copy number estadístico que indica el número de copias en una región del genoma, derivado de la fluorescencia en aCGH y SNP *arrays*, o del conteo de lecturas normalizado en NGS. 8
diploide situación en la que un organismo tiene dos copias de cada cromosoma, una materna y una paterna. Se designa como $2n$ ($2n = 46$ en humanos). 12
EAR Excluir alto riesgo 72
EMV estimador de máxima verosimilitud 42
enfermedad monogénica enfermedad causada por variantes en un solo gen, típicamente reconocida por ocurrir en múltiples miembros de una familia. 2
euploidía en sentido estricto, se dice de cualquier número de cromosomas que sea un múltiplo exacto del número haploide de la especie. En la práctica, en organismos diploides se suele reservar la expresión para el número correspondiente a $2n$ (la diploidía, $2n = 46$ en humanos), diferenciándolo de la triploidía, tetraploidía, etc. Priorizaremos este último uso. 12
FDR tasa de descubrimientos falsos 81
FISH hibridización fluorescente *in situ* 3
FIV fertilización *in vitro* 1
GBS estrategia de secuenciación NGS de alta profundidad, dirigida a regiones específicas del genoma que contienen un SNP y que permite genotipar. 10
genoma de referencia alguna de las versiones de la secuencia completa del genoma humano. 6
GWAS estudio de asociación de genoma completo 66
haploide situación en la que un organismo tiene una copia de cada cromosoma, usualmente designado como n ($n = 23$ en humanos). Los espermatozoides y óvulos normales son haploides. 12

heredabilidad proporción de la varianza fenotípica correspondiente a la varianza genética aditiva. 74
HWE equilibrio de Hary-Weinberg. Una población en HWE tiene frecuencias genotípicas constantes, que para un polimorfismo bialélico se definen en función de p, la frecuencia del alelo A, y q, la frecuencia del alelo B: p^2 (AA), $2pq$ (AB), q^2 (BB). 16
IRA el riesgo absoluto menos la prevalencia. 78
LD desequilibrio de ligamiento 67
lectura secuencia de nucleótidos generada por NGS. 6
$\log_2 R$ estadístico para detectar cambios en el número de copias de una región genómica. 4
LOH pérdida de heterocigosis 8, 28
MAD *median absolute deviation* 51
MAF frecuencia del alelo menor 16
marcador genético segmento de ADN con ubicación conocida. En esta tesis lo utilizamos como una palabra más general para referirnos a los SNPs. 66
MCI masa celular interna. Estructura celular ubicada en un polo del blastocisto. Da origen al feto. 13
medicina de precisión el uso de herramientas de diagnóstico, de prevención y tratamientos personalizados según la necesidad del paciente, basándose en sus características genéticas, entre otros aspectos. [83] 66
midparental promedio de los valores de la madre y del padre de un embrión. Por ejemplo, el qPRS midparental es el promedio del qPRS materno y el qPRS paterno. 83
modelo de umbral de propensión modelo que postula una variable continua subyacente a un fenotipo con clases discretas, como "sano" vs. "enfermo". El cambio de categoría ocurre cuando el valor de la variable sobrepasa un umbral fijo. 75
monosomía condición de un organismo diploide en la que un cromosoma del par está ausente. Un embrión con monosomía tiene un cromosoma menos que el número diploide normal: $2n - 1$. 12
mosaicismo fenómeno en el que un tejido u organismo tiene dos o más linajes celulares de diferente constitución cromosómica, pero ambas derivadas del mismo cigoto. Al organismo así afectado se le dice "mosaico". 13
  mosaicismo de nivel m mosaicismo con una proporción m de células muestreadas aneuploides y $1-m$ células muestreadas euploides. 29
NGS secuenciación de próxima generación 6
NNH inversa del IRA. *Number needed to harm* en inglés. 79
***non-targeted* NGS** estrategia de secuenciación NGS de baja profundidad, genera lecturas espaciadas a lo largo del genoma. 6
panel una colección de regiones genómicas de interés, usualmente definidas en format `bed`. Cada región es una tupla de (cromosoma, posición de inicio, posición final, [nombre]) 16
Panel 100K panel con *targets* a intervalos regulares de 100 Kb, que no necesariamente contienen SNPs. 16
Panel 100K SNPs resultado de conservar únicamente los intervalos del Panel 100K Full que contienen SNPs con MAF > 0.40. 16
PBR Priorizar bajo riesgo 72
PGD diagnóstico genético preimplantacional 2
PGS *screening* genético preimplantacional 2
PGT test genético preimplantacional 2
  PGT-A PGT de aneuploidías 2
  PGT-M PGT de enfermedades monogénicas 2
  PGT-P PGT de riesgo de enfermedades poligénicas 2
  PGT-SR PGT de rearreglos estructurales 2
pila lecturas de NGS alineadas por su secuencia a la región del genoma de referencia a la que pertenecen. 8
pleiotropía fenómeno en el que un único gen o SNP es responsable de efectos fenotípicos diferentes y en apariencia no relacionados entre sí. 72
poliploidía condición en la que un organismo o célula tiene más de dos complementos cromosómicos por célula. 4
prevalencia proporción de individuos de la población que tiene una enfermedad. Se suele simbolizar con la letra K. 74

profundidad número de veces que un mismo nucleótido es leído con NGS, o el promedio con que son leídos los nucleótidos de una región del genoma. 6
PRS puntaje de riesgo poligénico 2, 66
RA riesgo absoluto 77
SNP *array* tecnología de genotipado de SNPs bialélicos mediante comparación de fluorescencias. 5
SNP bialélico SNP con dos alelos observados en la población, que convencionalmente se denominan alelo A y alelo B. En este trabajo, suponemos que el alelo A es el alelo que está en el genoma de referencia. 5, 16
summary statistics conjunto de los efectos β y P valores de asocaciación calculados en un GWAS. 67
target regiones del genoma definidas en un panel, que serán amplificadas y secuenciadas. 8, 16
***targeted* NGS** estrategia de secuenciación NGS de alta profundidad, dirigida a regiones específicas del genoma, los *targets*. 8
TE trofectodermo. Capa de células que rodea a la masa celular interna en el blastocisto. Da origen a la placenta. 13
triploidía condición en la que un organismo tiene un número cromosómico correspondiente a tres veces el complemento haploide ($3n$), de modo que hay tres copias de cada cromosoma en su genoma. 12
trisomía condición de un organismo diploide en la que existen tres copias de un cromosoma en lugar de dos. Se lo representa como: $2n + 1$. 12
UKB-GWAS dataset de *summary statistics* de 1345 enfermedades realizadas por el Neale Lab sobre la población de UK BioBank. 74
UPD disomía uniparental. Condición en la que las dos copias de un cromosoma se derivan de una de las copias del mismo progenitor, de modo que no se observan genotipos heterocigotas en la región. Se contrasta con la condición normal de heterodisomía, en la que ambas copias vienen de progenitores diferentes. 4, 5, 8
ventana región acotada dentro de un cromosoma, usualmente de varias megabases de longitud. Por metonimia, decimos ventana de datos para referirnos al conjunto de datos pertenecientes a los SNPs que están en esa región. 37
ventana aneuploide ventana con al menos un 75 % de sus SNPs dentro de una región aneuploide. 39
ventana euploide ventana sin ningún SNP en regiones aneuploides. 39

Bibliografía

[1] M. C. Chang. «Fertilization of Rabbit Obva in vitro». En: *Nature* (1959) (vid. págs. ix, 1).

[2] Craig Niederberger y col. «Forty years of IVF». En: *Fertility and Sterility* 110.2 (2018), 185-324.e5. DOI: 10.1016/j.fertnstert.2018.06.005 (vid. págs. 1, 4, 5).

[3] David K Gardner y col. «Diagnosis of human preimplantation embryo viability». En: *Human Reproduction Update* 21.6 (2015), págs. 727-747. DOI: 10.1093/humupd/dmu064 (vid. págs. 1, 4, 14).

[4] Romualdo Sciorio, Luca Tramontano y James Catt. «Preimplantation genetic diagnosis (PGD) and genetic testing for aneuploidy (PGT-A): status and future challenges». En: *Gynecological Endocrinology* 36.1 (2020), págs. 6-11. DOI: 10.1080/09513590.2019.1641194 (vid. pág. 2).

[5] Andreas G. Schmutzler. «Theory and practice of preimplantation genetic screening (PGS)». En: *European Journal of Medical Genetics* 62.8 (2019), pág. 103670. DOI: 10.1016/j.ejmg.2019.103670 (vid. pág. 2).

[6] Fernando Zegers-Hochschild y col. «The International Glossary on Infertility and Fertility Care, 2017». En: *Fertility and Sterility* 108.3 (2017), págs. 393-406. DOI: 10.1016/j.fertnstert.2017.06.005 (vid. pág. 2).

[7] Nathan R. Treff y col. «Validation of concurrent preimplantation genetic testing for polygenic and monogenic disorders, structural rearrangements, and whole and segmental chromosome aneuploidy with a single universal platform». En: *European Journal of Medical Genetics* 62.8 (2019), pág. 103647. DOI: 10.1016/j.ejmg.2019.04.004 (vid. págs. 2, 5).

[8] Nathan R. Treff y col. «Utility and First Clinical Application of Screening Embryos for Polygenic Disease Risk Reduction». En: *Frontiers in Endocrinology* 10.December (2019), págs. 1-6. DOI: 10.3389/fendo.2019.00845 (vid. págs. 2, 71).

[9] Nathan R. Treff y col. «Preimplantation genetic testing for polygenic disease risk». En: *Reproduction* 160.5 (2020), A13-A17. DOI: 10.1530/REP-20-0071 (vid. pág. 2).

[10] Ehud Karavani y col. «Screening Human Embryos for Polygenic Traits Has Limited Utility». En: *Cell* 179.6 (2019), 1424-1435.e8. DOI: 10.1016/j.cell.2019.10.033 (vid. págs. 2, 71, 99).

[11] Patrick Turley y col. «Problems with Using Polygenic Scores to Select Embryos». En: *New England Journal of Medicine* 385.1 (2021), págs. 78-86. DOI: 10.1056/NEJMsr2105065 (vid. págs. 2, 69, 71, 99).

[12] Christian P. Schaaf, Joanna Wiszniewska y Arthur L. Beaudet. «Copy number and SNP arrays in clinical diagnostics». En: *Annual Review of Genomics and Human Genetics* 12 (2011), págs. 25-51. DOI: 10.1146/annurev-genom-092010-110715 (vid. págs. 3, 5).

[13] Stylianos E. Antonarakis y col. «Chromosome 21 and Down syndrome: From genomics to pathophysiology». En: *Nature Reviews Genetics* 5.10 (2004), págs. 725-738. DOI: 10.1038/nrg1448 (vid. pág. 3).

[14] Lauri D. Black y Jill M. Fischer. «Genetic Counseling for Preimplantation Genetic Testing». En: *Preimplantation Genetic Testing. Recent Advances in Reproductive Medicine*. 2020. Cap. 2, págs. 11-23. DOI: 10.1201/9780429445972-2 (vid. págs. 3, 14).

[15] Andrea Victor y col. «Preimplantation Genetic Testing for Aneuploidies: Where We Are and Where We're Going». En: *Preimplantation Genetic Testing. Recent Advances in Reproductive Medicine*. 2020. Cap. 3, págs. 25-48. DOI: 10.1201/9780429445972-3 (vid. págs. 3-5, 13).

[16] Paul R. Brezina, Raymond Anchan y William G. Kearns. «Preimplantation genetic testing for aneuploidy: what technology should you use and what are the differences?» En: *Journal of Assisted Reproduction and Genetics* 33.7 (2016), págs. 823-832. doi: 10.1007/s10815-016-0740-2 (vid. pág. 3).

[17] Nathan R. Treff y Richard T. Scott. «Four-hour quantitative real-time polymerase chain reaction-based comprehensive chromosome screening and accumulating evidence of accuracy, safety, predictive value, and clinical efficacy». En: *Fertility and Sterility* 99.4 (2013), págs. 1049-1053. doi: 10.1016/j.fertnstert.2012.11.007 (vid. pág. 3).

[18] Hsin Fu Chen, Ming Chen y Hong Nerng Ho. «An overview of the current and emerging platforms for preimplantation genetic testing for aneuploidies (PGT-A) in in vitro fertilization programs». En: *Taiwanese Journal of Obstetrics and Gynecology* 59.4 (2020), págs. 489-495. doi: 10.1016/j.tjog.2020.05.004 (vid. pág. 4).

[19] Francesco Fiorentino y col. «Application of next-generation sequencing technology for comprehensive aneuploidy screening of blastocysts in clinical preimplantation genetic screening cycles». En: *Human Reproduction* 29.12 (2014), págs. 2802-2813. doi: 10.1093/humrep/deu277 (vid. págs. 4, 6, 8).

[20] Elias M Dahdouh, Jacques Balayla y Juan Antonio García-Velasco. «Comprehensive chromosome screening improves embryo selection: a meta-analysis». En: *Fertility and Sterility* 104.6 (2015), págs. 1503-1512. doi: 10.1016/j.fertnstert.2015.08.038 (vid. pág. 4).

[21] Daniel Pinkel y col. «High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays». En: *Nature Genetics* 20.2 (1998), págs. 207-211. doi: 10.1038/2524 (vid. pág. 4).

[22] Anne Kallioniemi y col. «Comparative Genomic Hybridization for Molecular Cytogenetic Analysis of Solid Tumors». En: *Science* 258.5083 (1992), págs. 818-821. doi: 10.1126/science.1359641 (vid. pág. 4).

[23] Bassem A Bejjani y Lisa G Shaffer. «Application of array-based comparative genomic hybridization to clinical diagnostics.» En: *The Journal of molecular diagnostics : JMD* 8.5 (2006), págs. 528-533. doi: 10.2353/jmoldx.2006.060029 (vid. pág. 4).

[24] Uwe Heinrich y col. «Array comparative genetic hybridisation in clinical diagnostics: Principles and applications». En: *LaboratoriumsMedizin* 33.5 (2009), págs. 255-266. doi: 10.1515/JLM.2009.045 (vid. pág. 4).

[25] F. Fiorentino y col. «PGD for reciprocal and Robertsonian translocations using array comparative genomic hybridization». En: *Human Reproduction* 26.7 (2011), págs. 1925-1935. doi: 10.1093/humrep/der082 (vid. págs. 4, 5).

[26] Allen Kung y col. «Validation of next-generation sequencing for comprehensive chromosome screening of embryos». En: *Reproductive BioMedicine Online* 31.6 (2015), págs. 760-769. doi: 10.1016/j.rbmo.2015.09.002 (vid. págs. 4, 8).

[27] Francesca Scionti y col. «The cytoscan HD array in the diagnosis of neurodevelopmental disorders». En: *High-Throughput* 7.3 (2018), págs. 1-12. doi: 10.3390/ht7030028 (vid. pág. 5).

[28] Derek J. Nancarrow y col. «SiDCoN: A tool to aid scoring of DNA copy number changes in SNP chip data». En: *PLoS ONE* 2.10 (2007), págs. 1-8. doi: 10.1371/journal.pone.0001093 (vid. pág. 5).

[29] Can Alkan, Bradley P Coe y Evan E Eichler. «Genome structural variation discovery and genotyping». En: *Nature Reviews Genetics* 12.5 (2011), págs. 363-376. doi: 10.1038/nrg2958. arXiv: NIHMS150003 (vid. pág. 5).

[30] Nathan R. Treff y col. «Accurate single cell 24 chromosome aneuploidy screening using whole genome amplification and single nucleotide polymorphism microarrays». En: *Fertility and Sterility* 94.6 (2010), págs. 2017-2021. doi: 10.1016/j.fertnstert.2010.01.052 (vid. pág. 5).

[31] Cristina Gutiérrez-Mateo y col. «Validation of microarray comparative genomic hybridization for comprehensive chromosome analysis of embryos». En: *Fertility and Sterility* 95.3 (2011), págs. 953-958. doi: 10.1016/j.fertnstert.2010.09.010 (vid. pág. 5).

[32] Barton E. Slatko, Andrew F. Gardner y Frederick M. Ausubel. «Overview of Next-Generation Sequencing Technologies». En: *Current Protocols in Molecular Biology* 122.1 (2018), págs. 1-11. DOI: 10.1002/cpmb.59 (vid. pág. 6).

[33] Elaine R Mardis. «A decade's perspective on DNA sequencing technology.» En: *Nature* 470.7333 (2011), págs. 198-203. DOI: 10.1038/nature09796 (vid. pág. 6).

[34] W. Richard McCombie, John D. McPherson y Elaine R. Mardis. «Next-generation sequencing technologies». En: *Cold Spring Harbor Perspectives in Medicine* 9.11 (2019). DOI: 10.1101/cshperspect.a036798 (vid. pág. 6).

[35] E S Lander y col. «Initial sequencing and analysis of the human genome». En: *Nature* 409.6822 (2001), págs. 860-921. DOI: 10.1038/35057062. arXiv: 11237011 (vid. págs. 6, 66).

[36] Sara Goodwin, John D. McPherson y W. Richard McCombie. «Coming of age: Ten years of next-generation sequencing technologies». En: *Nature Reviews Genetics* 17.6 (2016), págs. 333-351. DOI: 10.1038/nrg.2016.49 (vid. pág. 7).

[37] Zhihong Yang y col. «Randomized comparison of next-generation sequencing and array comparative genomic hybridization for preimplantation genetic screening: A pilot study». En: *BMC Medical Genomics* 8.1 (2015), págs. 1-13. DOI: 10.1186/s12920-015-0110-4 (vid. pág. 8).

[38] Dagan Wells y col. «Clinical utilisation of a rapid low-pass whole genome sequencing technique for the diagnosis of aneuploidy in human embryos prior to implantation». En: *Journal of Medical Genetics* 51.8 (2014), págs. 553-562. DOI: 10.1136/jmedgenet-2014-102497 (vid. pág. 8).

[39] R S Zimmerman y col. «Preclinical validation of a targeted next generation sequencing-based comprehensive chromosome screening methodology in human blastocysts». En: *MHR: Basic science of reproductive medicine* 24.1 (2018), págs. 37-45. DOI: 10.1093/molehr/gax060 (vid. págs. 9, 11, 19, 62, 63).

[40] Diego Marin y col. «Validation of a targeted next generation sequencing-based comprehensive chromosome screening platform for detection of triploidy in human blastocysts». En: *Reproductive BioMedicine Online* 36.4 (2018), págs. 388-395. DOI: 10.1016/j.rbmo.2017.12.015 (vid. págs. 10, 11, 15, 62).

[41] Ashley W. Tiegs y col. «A multicenter, prospective, blinded, nonselection study evaluating the predictive value of an aneuploid diagnosis using a targeted next-generation sequencing–based preimplantation genetic testing for aneuploidy assay and impact of biopsy». En: *Fertility and Sterility* 115.3 (2021), págs. 627-637. DOI: 10.1016/j.fertnstert.2020.07.052 (vid. págs. 9, 11).

[42] Jiangfeng He y col. «Genotyping-by-sequencing (GBS), An ultimate marker-assisted selection (MAS) tool to accelerate plant breeding». En: *Frontiers in Plant Science* 5.SEP (2014), págs. 1-8. DOI: 10.3389/fpls.2014.00484 (vid. pág. 10).

[43] Nan Wang y col. «Applications of genotyping-by-sequencing (GBS) in maize genetics and breeding». En: *Scientific Reports* 10.1 (2020), págs. 1-12. DOI: 10.1038/s41598-020-73321-8 (vid. pág. 10).

[44] Pierluigi Strippoli y col. *Genetics and genomics of Down syndrome*. Vol. 56. Elsevier Ltd, 2019, págs. 1-39. DOI: 10.1016/bs.irrdd.2019.06.001 (vid. pág. 12).

[45] Terry Hassold y Patricia Hunt. «To err (meiotically) is human: The genesis of human aneuploidy». En: *Nature Reviews Genetics* 2.4 (2001), págs. 280-291. DOI: 10.1038/35066065 (vid. págs. 12, 13).

[46] Tyl H. Taylor y col. «The origin, mechanisms, incidence and clinical consequences of chromosomal mosaicism in humans». En: *Human Reproduction Update* 20.4 (2014), págs. 571-581. DOI: 10.1093/humupd/dmu016 (vid. págs. 12, 13).

[47] Maurizio Poli y Antonio Capalbo. «Mosaicism Mechanisms in Preimplantation Embryos». En: *Preimplantation Genetic Testing. Recent Advances in Reproductive Medicine*. 2020. Cap. 8, págs. 111-125. DOI: 10.1201/9780429445972-8 (vid. págs. 13, 30).

[48] Elpida Fragouli y col. «The origin and impact of embryonic aneuploidy». En: *Human Genetics* 132.9 (2013), págs. 1001-1013. DOI: 10.1007/s00439-013-1309-0 (vid. págs. 13, 14).

[49] Norbert Gleicher y col. «A single trophectoderm biopsy at blastocyst stage is mathematically unable to determine embryo ploidy accurately enough for clinical use». En: *Reproductive Biology and Endocrinology* 15.1 (2017), págs. 1-8. doi: 10.1186/s12958-017-0251-8 (vid. pág. 13).

[50] Mina Popovic, Felicitas Azpiroz y Susana M. Chuva de Sousa Lopes. «Engineered models of the human embryo». En: *Nature Biotechnology* 39.8 (2021), págs. 918-920. doi: 10.1038/s41587-021-01004-4 (vid. pág. 13).

[51] Peter Braude y col. «Preimplantation genetic diagnosis». En: *Fertility and Sterility* 82.SUPPL. 1 (2002), págs. 120-122. doi: 10.1016/j.fertnstert.2004.05.035 (vid. pág. 13).

[52] Nathan R. Treff y Rebekah S. Zimmerman. «Advances in Preimplantation Genetic Testing for Monogenic Disease and Aneuploidy». En: *Annual Review of Genomics and Human Genetics* 18.1 (2017), págs. 189-200. doi: 10.1146/annurev-genom-091416-035508 (vid. pág. 14).

[53] Jason M Franasiak y col. «The nature of aneuploidy with increasing age of the female partner: a review of 15,169 consecutive trophectoderm biopsies evaluated with comprehensive chromosomal screening». En: *Fertility and Sterility* 101.3 (2014), 656-663.e1. doi: 10.1016/j.fertnstert.2013.11.004 (vid. pág. 14).

[54] Preimplantation Genetic Diagnosis International Society. *PGDIS Position Statement on the Transfer of Mosaic Embryos 2021*. 2021 (vid. pág. 14).

[55] Manuel Viotti y col. «Using outcome data from one thousand mosaic embryo transfers to formulate an embryo ranking system for clinical use». En: *Fertility and Sterility* 115.5 (2021), págs. 1212-1224. doi: 10.1016/j.fertnstert.2020.11.041 (vid. págs. 14, 63).

[56] Kassie J. Hyde y Danny J. Schust. «Genetic considerations in recurrent pregnancy loss». En: *Cold Spring Harbor Perspectives in Medicine* 5.3 (2015). doi: 10.1101/cshperspect.a023119 (vid. pág. 14).

[57] Li Meng, Baoli Yin y Cuilian Zhang. «Validation of Preimplantation Genetic Tests for Aneuploidy With Cell-Free Dna From Spent Culture Media (Scm): Concordance Assessment and Implication». En: *Fertility and Sterility* 114.3 (2020), e420-e421. doi: 10.1016/j.fertnstert.2020.08.1224 (vid. pág. 14).

[58] Zhanhui Ou y col. «Re-analysis of whole blastocysts after trophectoderm biopsy indicated chromosome aneuploidy». En: *Human Genomics* 14.1 (2020), págs. 3-10. doi: 10.1186/s40246-019-0253-z (vid. pág. 14).

[59] Santiago Munné y col. «Preimplantation genetic testing for aneuploidy versus morphology as selection criteria for single frozen-thawed embryo transfer in good-prognosis patients: a multicenter randomized clinical trial». En: *Fertility and Sterility* 112.6 (2019), 1071-1079.e7. doi: 10.1016/j.fertnstert.2019.07.1346 (vid. pág. 15).

[60] Stefano Colella y col. «QuantiSNP: An objective bayes hidden-markov model to detect and accurately map copy number variation using SNP genotyping data». En: *Nucleic Acids Research* 35.6 (2007), págs. 2013-2025. doi: 10.1093/nar/gkm076 (vid. págs. 15, 24).

[61] Guillaume Assié y col. «SNP Arrays in Heterogeneous Tissue: Highly Accurate Collection of Both Germline and Somatic Genetic Information from Unpaired Single Tumor Samples». En: *American Journal of Human Genetics* 82.4 (2008), págs. 903-915. doi: 10.1016/j.ajhg.2008.01.012 (vid. págs. 15, 23).

[62] The 1000 Genomes Project Consortium. «A global reference for human genetic variation». En: *Nature* 526.7571 (2015). Ed. por Toomas Kivisild, págs. 68-74. doi: 10.1038/nature15393. arXiv: 15334406 (vid. págs. 16, 65, 66, 74).

[63] Nicholas Stoler y Anton Nekrutenko. «Sequencing error profiles of Illumina sequencing instruments». En: *NAR Genomics and Bioinformatics* 3.1 (2021), págs. 1-9. doi: 10.1093/nargab/lqab019 (vid. pág. 22).

[64] John W. Davey y col. «Genome-wide genetic marker discovery and genotyping using next-generation sequencing». En: *Nature Reviews Genetics* 12.7 (2011), págs. 499-510. doi: 10.1038/nrg3012 (vid. pág. 22).

[65] B. S. Everitt. «An introduction to finite mixture distributions». En: *Statistical Methods in Medical Research* 5.2 (1996), págs. 107-127. doi: 10.1177/096228029600500202 (vid. pág. 27).

[66] Geoffrey McLachlan y David Peel. *Finite Mixture Models*. 2000. DOI: 10.1198/tech.2002.s651 (vid. pág. 27).

[67] Jannie van Echten-Arends y col. «Chromosomal mosaicism in human preimplantation embryos: A systematic review». En: *Human Reproduction Update* 17.5 (2011), págs. 620-627. DOI: 10.1093/humupd/dmr014 (vid. pág. 29).

[68] Daniel A. Peiffer y col. «High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping». En: *Genome Research* 16.9 (2006), págs. 1136-1148. DOI: 10.1101/gr.5402306 (vid. pág. 30).

[69] Ming Lin y col. «dChipSNP: Significance curve and clustering of SNP-array-based loss-of-heterozygosity data». En: *Bioinformatics* 20.8 (2004), págs. 1233-1240. DOI: 10.1093/bioinformatics/bth069 (vid. pág. 39).

[70] Nathan R Treff y Rebekah S Zimmerman. «Advances in Preimplantation Genetic Testing for Monogenic Disease and Aneuploidy». En: *Annual Review of Genomics and Human Genetics* 18.1 (2017), págs. 189-200. DOI: 10.1146/annurev-genom-091416-035508 (vid. pág. 40).

[71] Ndeye Aicha Gueye y col. «Uniparental disomy in the human blastocyst is exceedingly rare». En: *Fertility and Sterility* 101.1 (2014), págs. 232-236. DOI: 10.1016/j.fertnstert.2013.08.051 (vid. pág. 40).

[72] Graciela Boente y Víctor Yohai. *Notas de Estadística* (vid. págs. 43, 60).

[73] David Skuse, Frida Printzlau y Jeanne Wolstencroft. *Sex chromosome aneuploidies*. 1.ª ed. Vol. 147. Elsevier B.V., 2018, págs. 355-376. DOI: 10.1016/B978-0-444-63233-3.00024-5 (vid. pág. 55).

[74] CoGEN. «COGEN Position Statement on Chromosomal Mosaicism Detected in Preimplantation Blastocyst Biopsies». En: (2018), págs. 2016-2019 (vid. pág. 63).

[75] Leo Breiman. «Statistical modeling: The two cultures». En: *Statistical Science* 16.3 (2001), págs. 199-215. DOI: 10.1214/ss/1009213726 (vid. pág. 63).

[76] Chris Cotsapas y col. «Pervasive sharing of genetic effects in autoimmune disease». En: *PLoS Genetics* 7.8 (2011). DOI: 10.1371/journal.pgen.1002254 (vid. págs. 65, 74).

[77] Brendan Bulik-Sullivan y col. «An atlas of genetic correlations across human diseases and traits». En: *Nature Genetics* 47.11 (2015), págs. 1236-1241. DOI: 10.1038/ng.3406 (vid. págs. 65, 73, 82, 107).

[78] Joseph K. Pickrell y col. «Detection and interpretation of shared genetic influences on 42 human traits». En: *Nature Genetics* 48.7 (2016), págs. 709-717. DOI: 10.1038/ng.3570 (vid. págs. 65, 73, 101).

[79] Kyoko Watanabe y col. «A global overview of pleiotropy and genetic architecture in complex traits». En: *Nature Genetics* 51.9 (2019), págs. 1339-1348. DOI: 10.1038/s41588-019-0481-0 (vid. págs. 65, 72).

[80] *Neale Lab UKB GWAS results, round 2* (vid. págs. 65, 74).

[81] Jiawen Chen y col. «Gamete simulation improves polygenic transmission disequilibrium analysis». En: *bioRxiv* (2020), pág. 2020.10.26.355602 (vid. págs. 65, 83).

[82] Qianqian Zhang y col. «Improved genetic prediction of complex traits from individual-level data or summary statistics». En: *Nature Communications* 12.1 (2021), pág. 4192. DOI: 10.1038/s41467-021-24485-y (vid. págs. 65, 69, 75, 103, 104).

[83] Ramya Ramaswami, Ronald Bayer y Sandro Galea. «Precision Medicine from a Public Health Perspective». En: *Annual Review of Public Health* 39 (2018), págs. 153-168. DOI: 10.1146/annurev-publhealth-040617-014158 (vid. págs. 66, 126).

[84] Masahira Hattori. «Finishing the euchromatic sequence of the human genome». En: *Nature* 431.7011 (2004), págs. 931-945. DOI: 10.1038/nature03001 (vid. pág. 66).

[85] The International y Hapmap Consortium. «The International HapMap Project.» En: *Nature* 426.6968 (2003), págs. 789-796. DOI: 10.1038/nature02168. arXiv: 1302.2710v1 (vid. pág. 66).

[86] Cathie Sudlow y col. «UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age». En: *PLOS Medicine* 12.3 (2015), e1001779. DOI: 10.1371/journal.pmed.1001779 (vid. págs. 66, 74).

[87] Danielle Welter y col. «The NHGRI GWAS Catalog, a curated resource of SNP-trait associations». En: *Nucleic Acids Research* 42.D1 (2014), págs. D1001-D1006. DOI: 10.1093/nar/gkt1229 (vid. pág. 66).

[88] *All of Us: Research Program Overview (Acceso: 2-mayo-2022)*. 2021 (vid. pág. 66).

[89] Inke R. König y col. «What is precision medicine?» En: *European Respiratory Journal* 50.4 (2017), págs. 1-12. DOI: 10.1183/13993003.00391-2017 (vid. pág. 66).

[90] Geoffrey S. Ginsburg y Kathryn A. Phillips. «Precision medicine: From science to value». En: *Health Affairs* 37.5 (2018), págs. 694-701. DOI: 10.1377/hlthaff.2017.1624 (vid. pág. 66).

[91] Polygenic Risk Score Task Force of the International Common Disease Alliance. «Responsible use of polygenic risk scores in the clinic: potential benefits, risks and gaps.» En: *Nature medicine* 27.November (2021). DOI: 10.1038/s41591-021-01549-6 (vid. pág. 66).

[92] Karolina Kauppi y col. «Effects of polygenic risk for Alzheimer's disease on rate of cognitive decline in normal aging». En: *Translational Psychiatry* 10.1 (2020), pág. 250. DOI: 10.1038/s41398-020-00934-y (vid. pág. 66).

[93] Gad Abraham y col. «Genomic risk score offers predictive performance comparable to clinical risk factors for ischaemic stroke». En: *Nature Communications* 10.1 (2019), pág. 5819. DOI: 10.1038/s41467-019-13848-1 (vid. pág. 66).

[94] Tatiane Yanes y col. «Clinical applications of polygenic breast cancer risk: a critical review and perspectives of an emerging field». En: *Breast Cancer Research* 22.1 (2020), pág. 21. DOI: 10.1186/s13058-020-01260-3 (vid. pág. 66).

[95] Felipe Padilla-Martínez y col. «Systematic Review of Polygenic Risk Scores for Type 1 and Type 2 Diabetes». En: *International Journal of Molecular Sciences* 21.5 (2020), pág. 1703. DOI: 10.3390/ijms21051703 (vid. pág. 66).

[96] Peter M. Visscher y col. «10 Years of GWAS Discovery: Biology, Function, and Translation». En: *American Journal of Human Genetics* 101.1 (2017), págs. 5-22. DOI: 10.1016/j.ajhg.2017.06.005 (vid. págs. 66, 67).

[97] Robert Plomin, Claire M. A. Haworth y Oliver S. P. Davis. «Common disorders are quantitative traits». En: *Nature Reviews Genetics* 10.12 (2009), págs. 872-878. DOI: 10.1038/nrg2670 (vid. pág. 66).

[98] Cathryn M. Lewis y Evangelos Vassos. «Polygenic risk scores: from research tools to clinical instruments». En: *Genome Medicine* 12.1 (2020), pág. 44. DOI: 10.1186/s13073-020-00742-5 (vid. págs. 66, 68).

[99] Teri A. Manolio. «Genomewide Association Studies and Assessment of the Risk of Disease». En: *New England Journal of Medicine* 363.2 (2010), págs. 166-176. DOI: 10.1056/nejmra0905980 (vid. pág. 67).

[100] Alice B. Popejoy y Stephanie M. Fullerton. «Genomics is failing on diversity». En: *Nature* 538.7624 (2016), págs. 161-164. DOI: 10.1038/538161a. arXiv: 15334406 (vid. pág. 67).

[101] Deepti Gurdasani y col. «Genomics of disease risk in globally diverse populations». En: *Nature Reviews Genetics* (2019). DOI: 10.1038/s41576-019-0144-0 (vid. pág. 67).

[102] Melinda C. Mills y Charles Rahal. «A scientometric review of genome-wide association studies». En: *Communications Biology* 2.1 (2019). DOI: 10.1038/s42003-018-0261-x (vid. págs. 67, 68).

[103] Christopher S. Carlson y col. «Generalization and Dilution of Association Results from European GWAS in Populations of Non-European Ancestry: The PAGE Study». En: *PLoS Biology* 11.9 (2013). DOI: 10.1371/journal.pbio.1001661 (vid. pág. 67).

[104] Urko M. Marigorta y Arcadi Navarro. «High Trans-ethnic Replicability of GWAS Results Implies Common Causal Variants». En: *PLoS Genetics* 9.6 (2013). Ed. por Scott M. Williams, e1003566. DOI: 10.1371/journal.pgen.1003566 (vid. pág. 67).

[105] Naomi R Wray y Michael E Goddard. «Multi-locus models of genetic risk of disease». En: *Genome Medicine* 2.2 (2010), pág. 10. DOI: 10.1186/gm131 (vid. págs. 68, 75).

[106] Nilanjan Chatterjee, Jianxin Shi y Montserrat García-Closas. «Developing and evaluating polygenic risk prediction models for stratified disease prevention». En: *Nature Reviews Genetics* 17.7 (2016), págs. 392-406. DOI: 10.1038/nrg.2016.27 (vid. pág. 68).

[107] Ali Torkamani, Nathan E Wineinger y Eric J Topol. «The personal and clinical utility of polygenic risk scores». En: *Nature Reviews Genetics* (2018). DOI: 10.1038/s41576-018-0018-x (vid. pág. 68).

[108] Judit Kumuthini y col. «The clinical utility of polygenic risk scores in genomic medicine practices: a systematic review». En: *Human Genetics* 0123456789 (2022). DOI: 10.1007/s00439-022-02452-x (vid. págs. 68, 71).

[109] International Schizophrenia Consortium y col. «Common polygenic variation contributes to risk of schizophrenia and bipolar disorder (Supplementary Information)». En: *Nature* 460.7256 (2009), págs. 748-52. DOI: 10.1038/nature08185. arXiv: NIHMS150003 (vid. pág. 69).

[110] Bjarni J. Vilhjálmsson y col. «Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores». En: *American Journal of Human Genetics* 97.4 (2015), págs. 576-592. DOI: 10.1016/j.ajhg.2015.09.001 (vid. pág. 69).

[111] Timothy Shin Heng Mak y col. «Polygenic scores via penalized regression on summary statistics». En: *Genetic Epidemiology* 41.6 (2017), págs. 469-480. DOI: 10.1002/gepi.22050 (vid. pág. 69).

[112] Xiang Zhu y Matthew Stephens. «Bayesian large-scale multiple regression with summary statistics from genome-wide association studies». En: *Annals of Applied Statistics* 11.3 (2017), págs. 1561-1592. DOI: 10.1214/17-AOAS1046 (vid. pág. 69).

[113] Jiabo Wang y col. «Expanding the BLUP alphabet for genomic prediction adaptable to the genetic architectures of complex traits». En: *Heredity* 121.6 (2018), págs. 648-662. DOI: 10.1038/s41437-018-0075-0 (vid. pág. 69).

[114] Luke R. Lloyd-Jones y col. «Improved polygenic prediction by Bayesian multiple regression on summary statistics». En: *Nature Communications* 10.1 (2019). DOI: 10.1038/s41467-019-12653-0 (vid. pág. 69).

[115] Ali Torkamani y col. «A primer on deep learning in genomics». En: *Nature Genetics* 51.1 (2018), págs. 12-18. DOI: 10.1038/s41588-018-0295-5 (vid. pág. 71).

[116] Antonio Regalado. «Eugenics 2.0 : We're at the Dawn of Choosing Embryos by Health, Height, and More». En: *MIT Technology Review* (2017) (vid. pág. 71).

[117] Pete Shanks. «Polygenic Traits, Human Embryos, and Eugenic Dreams». En: (2019) (vid. pág. 71).

[118] Nathan R. Treff y col. «Preimplantation Genetic Testing for Polygenic Disease Relative Risk Reduction: Evaluation of Genomic Index Performance in 11,883 Adult Sibling Pairs». En: *Genes* 11.6 (2020), pág. 648. DOI: 10.3390/genes11060648 (vid. págs. 71, 99).

[119] L. Duncan y col. «Analysis of polygenic risk score usage and performance in diverse human populations». En: *Nature Communications* 10.1 (2019), pág. 3328. DOI: 10.1038/s41467-019-11112-0 (vid. pág. 71).

[120] Soke Yuen Yong y col. «Genetic architecture of complex traits and disease risk predictors». En: *Scientific Reports* 10.1 (2020), págs. 1-14. DOI: 10.1038/s41598-020-68881-8 (vid. pág. 71).

[121] J. D. Schulman y R. G. Edwards. «Preimplantation diagnosis is disease control, not eugenics». En: *Human Reproduction* 11.3 (1996), págs. 463-464. DOI: 10.1093/HUMREP/11.3.463 (vid. pág. 71).

[122] Todd Lencz y col. «Utility of polygenic embryo screening for disease depends on the selection strategy». En: *eLife* 10 (2021). DOI: 10.7554/eLife.64716 (vid. págs. 71, 72, 75, 78).

[123] Annalise B. Paaby y Matthew V. Rockman. «The many faces of pleiotropy». En: *Trends in Genetics* 29.2 (2013), págs. 66-73. DOI: 10.1016/j.tig.2012.10.010 (vid. pág. 72).

[124] Zhi Wang, Ben Yang Liao y Jianzhi Zhang. «Genomic patterns of pleiotropy and the evolution of complexity». En: *Proceedings of the National Academy of Sciences of the United States of America* 107.42 (2010), págs. 18034-18039. DOI: 10.1073/pnas.1004666107 (vid. pág. 72).

[125] Miguel dos Santos, Melanie Ghoul y Stuart A. West. «Pleiotropy, cooperation, and the social evolution of genetic architecture». En: *PLoS Biology* 16.10 (2018), págs. 1-25. DOI: 10.1371/journal.pbio.2006671 (vid. pág. 72).

[126] Peter M. Visscher y Jian Yang. «A plethora of pleiotropy across complex traits». En: *Nature Genetics* 48.7 (2016), págs. 707-708. DOI: 10.1038/ng.3604 (vid. pág. 72).

[127] Huwenbo Shi, Gleb Kichaev y Bogdan Pasaniuc. «Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data». En: *American Journal of Human Genetics* 99.1 (2016), págs. 139-153. DOI: 10.1016/j.ajhg.2016.05.013 (vid. pág. 72).

[128] Wouter van Rheenen y col. «Genetic correlations of polygenic disease traits: from theory to practice». En: *Nature Reviews Genetics* 20.10 (2019), págs. 567-581. DOI: 10.1038/s41576-019-0137-z (vid. págs. 72, 73).

[129] Günter P Wagner y Jianzhi Zhang. «The pleiotropic structure of the genotype–phenotype map: the evolvability of complex organisms». En: *Nature Reviews Genetics* 12.3 (2011), págs. 204-213. DOI: 10.1038/nrg2949 (vid. págs. 72, 73, 100).

[130] Huwenbo Shi y col. «Local Genetic Correlation Gives Insights into the Shared Genetic Architecture of Complex Traits». En: *The American Journal of Human Genetics* 101.5 (2017), págs. 737-751. DOI: 10.1016/j.ajhg.2017.09.022 (vid. pág. 73).

[131] Zheng Ning, Yudi Pawitan y Xia Shen. «High-definition likelihood inference of genetic correlations across human complex traits». En: *Nature Genetics* 52.8 (2020), págs. 859-864. DOI: 10.1038/s41588-020-0653-y (vid. pág. 73).

[132] *Neale UKB Genetic Correlations* (vid. págs. 73, 82, 87, 100, 107).

[133] M. Yu Zakharova y col. «The contribution of major histocompatibility complex class II genes to an association with autoimmune diseases». En: *Acta Naturae* 11.4 (2019), págs. 4-12. DOI: 10.32607/20758251-2019-11-4-4-12 (vid. pág. 74).

[134] Minal Caliskan, Christopher D. Brown y Joseph C. Maranville. «A catalog of GWAS fine-mapping efforts in autoimmune disease». En: *American Journal of Human Genetics* 108.4 (2021), págs. 549-563. DOI: 10.1016/j.ajhg.2021.03.009 (vid. pág. 74).

[135] *Neale Lab UKB Heritability* (vid. págs. 74, 77, 92, 100).

[136] Christopher C. Chang y col. «Second-generation PLINK: rising to the challenge of larger and richer datasets». En: *GigaScience* 4.1 (2015), pág. 7. DOI: 10.1186/s13742-015-0047-8. arXiv: 1410.4803 (vid. pág. 74).

[137] Douglas S Falconer y Trudy F C Mackay. «Chapter 18. Threshold Characters». En: *Introduction to quantitative genetics*. 1996. Cap. 18 (vid. págs. 75, 76).

[138] Margaux L.A. Hujoel y col. «Liability threshold modeling of case–control status and family history of disease increases association power». En: *Nature Genetics* 52.5 (2020), págs. 541-547. DOI: 10.1038/s41588-020-0613-6 (vid. pág. 75).

[139] Peter M. Visscher y Naomi R. Wray. «Concepts and Misconceptions about the Polygenic Additive Model Applied to Disease». En: *Human Heredity* 80.4 (2016), págs. 165-170. DOI: 10.1159/000446931 (vid. págs. 75, 76, 100).

[140] Yoav Benjamini y Yosef Hochberg. «Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing». En: *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1 (1995), págs. 289-300 (vid. pág. 80).

[141] Soke Yuen Yong y col. «Genetic architecture of complex traits and disease risk predictors». En: *Scientific Reports* 10.1 (2020), págs. 1-36. DOI: 10.1038/s41598-020-68881-8 (vid. pág. 82).

[142] Zijie Zhao y col. «PUMAS: fine-tuning polygenic risk scores with GWAS summary statistics». En: *Genome Biology* 22.1 (2021), pág. 257. DOI: 10.1186/s13059-021-02479-9 (vid. pág. 103).

[143] Brendan Bulik-Sullivan y col. «LD score regression distinguishes confounding from polygenicity in genome-wide association studies». En: *Nature Genetics* 47.3 (2015), págs. 291-295. DOI: 10.1038/ng.3211 (vid. pág. 107).

Fin

Esta tesis se terminó de escribir el día 20 de julio del año 2022 de nuestra era, en el barrio del Abasto, en la Ciudad Autónoma de Buenos Aires, en la República Argentina.