UNIVERSIDAD DE MURCIA

ESCUELA INTERNACIONAL DE DOCTORADO

Herramientas Bioinformáticas para el Diagnóstico Genético Preimplantacional mediante Secuenciación Masiva

Dª Natalia Castejón Fernández

2021

Universidad de Murcia

Bioarray S.L.

HERRAMIENTAS BIOINFORMÁTICAS PARA EL DIAGNOSTICO GENÉTICO PREIMPLANTACIONAL MEDIANTE SECUENCIACIÓN MASIVA

"Esta tesis doctoral está sometida a procesos de protección o transferencia de tecnología o de conocimiento, por lo que algunos contenidos están inhibidos en la publicación en los repositorios institucionales.
Autorizado por la Comisión General de Doctorado de la Universidad de Murcia con fecha 24 de febrero de 2021"

Tesis Doctoral

Natalia Castejón Fernández

2021

La presente tesis ha sido desarrollada bajo el marco de la *Ayuda para la formación de doctores en empresas, "Doctorados industriales"*, englobada en el *Programa Estatal de Promoción del Talento y su Empleabilidad en I+D+I* otorgada por el, entonces *Ministerio de Economía, Industria y Competitividad*, actual *Ministerio de Ciencia, Innovación y Universidades* y cofinanciada por el *Banco Europeo de Inversiones*, a la empresa Bioarray SL con CIF:B54363049 para la contratación de Natalia Castejón Fernández como doctorando bajo el proyecto específico DI-14-06922 titulado "Herramientas Bioinformáticas para el Diagnóstico Genético Preimplantacional mediante técnicas de Secuenciación Masiva".

"Esta tesis doctoral está sometida a procesos de protección o transferencia de tecnología o de conocimiento, por lo que los siguientes contenidos están inhibidos en la publicación en los repositorios institucionales.

Autorizado por la Comisión General de Doctorado de la Universidad de Murcia con fecha 24 de febrero de 2021"

TABLAS SOMETIDAS A CONFIDENCIALIDAD

FIGURAS SOMETIDAS A CONFIDENCIALIDAD

CONTENIDOS SOMETIDOS A CONFIDENCIALIDAD

Consentimientos

UNIVERSIDAD DE
MURCIA

D. Fernando Soler Pardo, Catedrático de Universidad del Área de Bioquímica y Biología Molecular yPresidente de la Comisión Académica del Programa de Doctorado en Biología Molecular y Biotecnología,

INFORMA:

Que vista la solicitud de autorización de presentación de tesis doctoral de D. Natalia Castejón Fernnández, titulada "Herramientas Bioinformáticas para el Diagnóstico Genético Preimplantacional mediante Secuenciación Masiva", realizada bajo la inmediata dirección y supervisión de D. Jesualdo Tomás Fernández Breis y Luis Alcaraz Más, y evaluado el expediente completo, la Comisión Académica del Programa de Doctorado, en sesión celebrada el día 26 de febrero de 2021, y de conformidad con lo establecido en el artículo 21 del "Reglamento por el que se regulan las enseñanzas oficiales de doctorado de la Universidad Murcia", resolvió la autorización de presentación de la tesis doctoral.

Asimismo, le envía el informe de la Comisión de Rama de Conocimiento de Ciencias sobre la propuesta de expertos que pueden formar parte del tribunal que ha de juzgarla, junto con los preceptivos informes de idoneidad.

Murcia, a 26 de febrero de 2021

SOLER PARDO FERNANDO - 27435221Q

Firmado digitalmente por SOLER PARDO FERNANDO - 27435221Q
Fecha: 2021.02.26 11:48:58 +01'00'

Fdo.: Fernando Soler Pardo

COMISIÓN GENERAL DE DOCTORADO. UNIVERSIDAD DE MURCIA

UNIVERSIDAD DE
MURCIA

D. Jesualdo Tomás Fernández Breis, Catedrático de Universidad del Área de Lenguajes y Sistemas Informáticos en el Departamento de Informática y Sistemas, AUTORIZA:

La presentación de la Tesis Doctoral titulada "Herramientas Bioinformáticas para el Diagnóstico Genético Preimplantacional mediante Secuenciación Masiva ", realizada por Dª. Natalia Castejón Fernández, bajo mi inmediata dirección y supervisión, y que presenta para la obtención del grado de Doctor por la Universidad de Murcia.

En Murcia, a 8 de diciembre de 2020

BIOARRAY
Diagnóstico Genético

Parque Científico y Empresarial de la UMH. Edificio Quorum III
Avenida de la Universidad s/n. 03202 Elche (Alicante - Spain)
Tlf: +34 96 668 25 00 Fax:+34 96 668 25 01
info@bioarray.es www.bioarray.es

Elche, a 15 de diciembre de 2020

D. Luis Antonio Alcaraz Mas, Director Técnico de Bioarray, S.L.U., y Director de tesis de Natalia Casetjón Fernández,

AUTORIZA:

La presentación de la tesis doctoral de D. Natalia Castejón Fernández, titulada "Herramientas Bioinformáticas para el Diagnóstico Genético Preimplantacional mediante Secuenciación Masiva", realizada bajo mi dirección y desarrollada en el marco del a empresa.

Firmado digitalmente por ALCARAZ MAS LUIS ANTONIO - 48368441Q
Fecha: 2020.12.15 12:16:47 +01'00'

Dr. Luis A. Alcaraz Mas
Director Técnico.
Bioarray

UNIVERSIDAD DE
MURCIA

D. Jesualdo Tomás Fernández Breis, Catedrático de Universidad del Área de Informática y Sistemas, y Director de tesis de Natalia Castejón Fernández,

INFORMA:

La tesis doctoral de D. Natalia Castejón Fernández, titulada "Herramientas Bioinformáticas para el Diagnóstico Genético Preimplantacional mediante Secuenciación Masiva", realizada bajo la inmediata dirección y supervisión de D. Jesualdo Tomás Fernández Breis y Luis Alcaraz Más, se ha realizado en el seno de la empresa Bioarray, estando Natalia contratada con financiación de la empresa y del Programa de Doctorados Industriales (DI-14-06922) financiado por el Ministerio de Economía y Competitividad, por lo que considera adecuada su solicitud de lectura y exposición de tesis en circunstancias excepcionales de confidencialidad.

Los apartados sometidos a confidencialidad son los siguientes:

3.1.1 MiNtagSNP
3.1.1.1 SPA: Algoritmo de predicción de SNPs
3.1.1.2 SSA: Algoritmo de selección de SNPs
3.1.1.3 Especificaciones del usuario
3.1.1.4 Tiempo de ejecución

4.1 Valores óptimos para MAxP y HETrate
4.2 Validación In silico
4.2.1 Población
4.2.2 Paneles diseñados
4.2.3 Tiempo de ejecución
4.2.4 Informatividad
4.2.5 Imputación
4.3 Implementación
4.3.1 Informatividad
4.3.2 Imputación

COMISIÓN GENERAL DE DOCTORADO. UNIVERSIDAD DE MURCIA

Tablas asociadas a estos apartados

Tabla 11: Distribución demográfica de la población escogida.
Tabla 12: Polimorfismos registrados en la base de datos 1000GenomesDB
Tabla 13: Nº de tagSNPs diseñados por cada algoritmo en cada región analizada.
Tabla 14: Resumen de la reducción y tiempo empleados por MiNtagSNP
Tabla 15: Porcentaje de informatividad arrojadapor cada algoritmo.
Tabla 16: Precisión de imputación para cada algoritmo en cada región.
Tabla 17: Datos relativos al análisis de la zona de 14 Mb.
Tabla 18: Tabla resumen de casos in vitro.
Tabla 19: Resumen de las muestras con mayor y menor porcentaje de polimorfismos

Figuras asociadas a estos apartados

Figura 40: Relación r2 entre tagSNPs de los paneles diseñados por cada algoritmo.
Figura 41: Relación D’ entre tagSNPs de los paneles diseñados por cada algoritmo.

Murcia, a 12 de Febrero de 2021

Fdo.: Jesualdo Tomás Fernández Breis

BIOARRAY
Diagnóstico Genético

Parque Científico y Empresarial de la UMH. Edificio Quorum III
Avenida de la Universidad s/n. 03202 Elche (Alicante - Spain)
Tlf: +34 96 668 25 00 Fax:+34 96 668 25 01
info@bioarray.es www.bioarray.es

Elche, a 15 de diciembre de 2020

D. Luis Antonio Alcaraz Mas, Director Técnico de Bioarray, S.L.U., y Director de tesis de Natalia Casetjón Fernández,

INFORMA:

La tesis doctoral de D. Natalia Castejón Fernández, titulada "Herramientas Bioinformáticas para el Diagnóstico Genético Preimplantacional mediante Secuenciación Masiva", realizada bajo la inmediata dirección y supervisión de D. Luis A. Alcaraz y D. Jesualdo Tomás Fernández (Universidad de Murcia), se ha realizado íntegramente en Bioarray, estando Natalia contratada por la empresa, con ayuda del Programa Doctorado Industrial (DI-14-06922) financiado por el Ministerio de Economía y Competitividad. Algunas partes de la tesis contienen información sensible para empresa, por lo que considero adecuada su solicitud de lectura y exposición en circunstancias excepcionales de confidencilidad.

Firmado digitalmente por ALCARAZ MAS LUIS ANTONIO - 48368441Q
Fecha: 2020.12.15 10:29:07 +01'00'

Dr. Luis A. Alcaraz Mas
Director Técnico.
Bioarray

A mi madre,
quién me enseñó a leer con cartulinas
y a contar con vasos de agua.

A la memoria de mi padre,
de quién aprendí a dedicar siempre cinco
minutos más.

A mi hermano,
que me ayudó a reír también.

Gracias.

AGRADECIMIENTOS

En primer lugar al Dr. Luis Antonio Alcaraz Mas, y a Don Andrés Antón Salas, de la empresa Bioarray, por darme la oportunidad de realizar algo que, más que un título profesional, es para mí una meta personal.

A mi director, el Dr. Jesualdo Fernández Breis, de la Universidad de Murcia, por ser psicólogo en los primeros meses del proyecto, cuando aún no había alcanzado la madurez necesaria para afrontarlo, y por haber sido siempre enlace, punto de información y solución frente a la burocracia universitaria. Gracias.

A mis compañeros de Bioarray, por el apoyo en los momentos de pánico. Por la experiencia y la paciencia. Por las bromas. Nunca olvidaré esta etapa.

Al Dr. Alberto Ruiz García y al Dr. Francisco Javier Abellán García. No solo por la oportunidad de impartir clases en la Universidad de Murcia y conocer el mundo docente, sino por valorar y agradecer cada opinión y cada idea, considerándome una más. Pero, sobre todo, por esos ratitos de charla con nuestro "me lo como". Gracias por hacerme sentir tan cómoda, valorada e integrada.

A mi supervisor de estancia en *Aarhus Universitet*, el Dr. Thomas Mailund, por darme la oportunidad de colaborar en su proyecto y disfrutar de una experiencia tan enriquecedora. Sheiliza, Maria, Svend, Doug, Andders, Kate, Jenna, Enza, Luca y Pere, el mejor resumen de una estancia fabulosa en Dinamarca. Un país que me conquistó como nunca habría esperado. "T*akfor at repræsentere glæde og håb. Tak fordi du var min største støtte under denne vanskelige proces. Tak fordi du fortsætter med at støtte mig på afstand. Vi ses snart, venner".*

A mis padres, que me han enseñado a trabajar con tesón y luchar por lo que se desea. Valores y principios que me han permitido llevar este proyecto hasta el final y poder presentarlo con orgullo. Por creer en mi incondicionalmente y soportar mis peores momentos con paciencia y cariño. Sin vosotros, sin vuestras enseñanzas, no habría sido capaz de lograrlo. Pero ante todo, porque aunque ha sido duro estar distanciados, siempre hayáis preferido tenerme lejos y feliz a tenerme cerca y hundida. Os quiero.

A ti, por seguir ahí. Por zeta. Por tratar de ir siempre *S by S*.

Y por último, a mí hermano. Simplemente por ser tú, lo mejor que tengo. Porque aunque estemos lejos siempre estamos cerca cuando hace falta.

Sinceramente, gracias a todos.

Natalia

Pd: Bueno… y a Tiza y Bruma, que son "*hygge*".

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ÍNDICE

RESUMEN

- **Epítome**

Aproximadamente el 99,9% de la secuencia de ADN es común para toda la especie humana. Si bien es cierto que la mayoría de las alteraciones de ese 0,1% carecen de importancia fenotípica y se denominan polimorfismos, otras pueden llegar a ser cruciales en el desarrollo de enfermedades; estas últimas son las que se denominan mutaciones.

A pesar de la gran complejidad que acarrean las alteraciones genéticas, tradicionalmente se clasifican en cinco grandes grupos: Cromosómicas, monogénicas, mitocondriales, multifactoriales y adquiridas. La técnica del diagnóstico genético preimplantacional (DGP) surge como alternativa al diagnóstico prenatal para parejas en riesgo de tener descendencia afecta por alteraciones cromosómicas y/o enfermedades monogénicas que desean evitar la necesidad de recurrir a terminaciones voluntarias del embarazo, para lo cual se someten a un ciclo de fertilización *in vitro* (IVF) en un proceso de reproducción asistida.

El diagnóstico genético preimplantacional de aneuploidías (DGP-A) se define como el procedimiento que analiza si los embriones presentan anomalías cromosómicas antes de ser transferidos al útero materno. Para poder aplicar la técnica, los embriones son biopsiados para extraer una/unas pocas células que serán analizadas para conocer el estado de ploidía. Si son diagnosticadas como euploides, entonces el embrión del que proceden es catalogado como "normal" y puede ser transferido al útero materno para iniciar la gestación. En los últimos años, la técnica preferida para DGP-A es la secuenciación masiva (NGS). El protocolo de elaboración de la librería es ligeramente diferente con respecto a otras técnicas de diagnóstico genético, pues los fragmentos de ADN a secuenciar no se originan por fragmentación del material original sino por amplificación a través de cebadores aleatorios que se unen al ADN original. A veces, durante el proceso se producen sesgos que generan dos posibles fuentes de artefactos: la primera, debida a la hibridación de cebadores aleatorios en ADN amplificado en lugar de ADN genómico; la segunda debida a la amplificación de la librería tras la ligación del adaptador. En ambos casos, se trata de duplicados de PCR, que pueden enmascarar los resultados. La llegada de la secuenciación masiva para DGP-A puso de manifiesto un fenómeno que había pasado inadvertido, el mosaicismo, definido como la presencia de dos o más líneas celulares distintas en un embrión. El porcentaje de aneuploidía presente, los cromosomas y el tipo celular afetados determinarán, en gran medida, la viabilidad de un embrión. Cabe destacar que la bibliografía recoge que estas anomalías producen bajos ratios de implantación en procesos

de reproducción asistida. Además, algunos estudios afirman que las tasas de éxito en transferencias de embriones con bajo porcentaje de mosaicismo son en realidad falsos positivos introducidos durante el proceso de biopsia, amplificación y/o análisis y, por ello, esto embriones presentan tasas de implantación equivalentes a las de un embrión normal. Por su parte, otras corrientes acusan que las técnicas presentan falsos negativos, alegando la existencia de casos donde embriones determinados como euploides engendraron bebés mosaico[4]. En cualquier caso, todos parecen coincider en que la determinación del porcentaje exacto de aneuplodía es un elemento crítico para determinar la probabilidad de implantación de un embrión mosiaico. Esto provoca que el desarrollo de un método preciso de determinación del mosaicismo, controlando niveles de sensibilidad y especificidad, sea una tarea esencial que ha generado gran interés en el mundo del DGP.

La presente tesis desarrolla un algoritmo destinado al filtrado de duplicados y artefactos de PCR denominado MiNFilterDups y un algoritmo que permite la detección de porcentajes bajos de aneuploidía y determinación del nivel de mosaicismo, el MiNmos. Para su validación se generaron varios conjuntos de archivos a partir de los datos de embriones reales de pacientes que se habían sometido a un proceso de reproducción asistida.

El DGP también puede ser aplicado en la determinación de embriones libres de enfermedades monogénicas cuando los padres son portadores mediante la aplicación del diagnóstico molecular, lo que se conoce como DGP-M. Las técnicas son muy variadas y han ido evolucionando, pero todas ellas deben enfrentarse al fenómeno ADO (*Allele Drop-Out)* o amplificación preferencial de un alelo frente al otro. Este efecto provoca que, al secuenciar, un locus heterocigoto, éste aparece como homocigoto debido a que uno de los dos alelos no es amplificado. Esto genera cierta incertidumbre respecto al resultado emitido ya que si al analizar un embrión éste aparece como homocigoto y no se detecta la alteración en estudio, cabe la posibilidad de que el alelo mutado no haya sido amplificado, constituyendo un falso negativo que puede desencadenar la transferencia de un embrión enfermo. Para evitar este problema, generalmente no se estudia únicamente la mutación en los embriones, sino que también se analizan varios polimorfismos adjacentes al locus mutado. Mediante un estudio de ligamiento, se determina si estos polimorfismos cosegregan con el alelo mutado o con el silvestre. Al estudio de la mutación en los embriones se le conoce como estudio directo; al estudio de polimorfismos, estudio indirecto.

Tradicionalmente, el estudio indirecto se ha realizado mediante el análisis de STRs (short tandem repetas). Sin embargo, la ventaja del empleo de SNPs consiste en la gran densidad que presentan a lo largo de todo el genoma, lo que permite que los embriones

portadores puedan ser descartados si presentan los polimorfismos asociados al cromosoma afecto. Secuenciar todos los SNPs de una región es algo redundante y costoso; afortunadamente, se puede reducir el número de SNPs empleando estrategias de selección de tagSNPs (polimorfismos que representan a otros) basadas en el desequilibrio de ligamiento. Existen multitud de parámetros que permiten calcular si un SNP puede ser o no útil en la predicción de otros, pero en el caso concreto de la selección de polimorfismos útiles para DGP-M se requiere la consideración de más factores que la correlación entre los mismos, pues la mayoría de técnicas de selección obtienen paneles que no son útiles a la hora de distinguir fenómenos de recombinación. El presente estudio desarrolla MiNtagSNP, un algoritmo de selección de tagSNPs útiles en DGP-M, que permite, en combinación con el fasado, determinar los polimorfismos presentes en cada embrión y descartar aquellos que porten la alteración, no solo por su detección directa, sino por mostrar los polimorfismos que cosegregan con la alteración en los parentales. La validación *in silico* se realizó empleando datos simulados a partir de las principales bases de datos de polimorfismos. Para la validación *in vitro*, se comprobó la concordancia de los resultados con respecto a los métodos tradicionales. Finalmente, la metodología se implementó en el laboratorio, y se realizó un seguimiento de los casos en un periodo de tiempo.

Así, podemos concluir que el principal objetivo de los algoritmos desarrollados en el marco de esta tesis persigue el diseño de un método de análisis rápido y eficaz que permita aunar los procesos de análisis de DGP-A y DGP-M mediante secuenciación por NGS a partir de una biopsia única.

Abstract

Humans share about 99% of the DNA sequence. Polymorphisms are the most common alterations of the remaining 0.1% and they lack phenotypic importance, but the variants called mutations are crucial in the development of diseases.

Despite the great complexity involved in genetic alterations, there are traditionally five large groups: chromosomal, monogenic, mitochondrial, multifactorial, and acquired alterations. The Preimplantation Genetic Diagnosis (PGD) technique arises as an alternative to prenatal diagnosis for couples at risk of having offsprings affected by chromosomal alterations and / or monogenic diseases, who wish to avoid the need to resort to voluntary terminations of pregnancy, so they decided to undergo an *in vitro* fertilization (IVF) cycle in an assisted reproduction process.

Preimplantation Genetic Testing for Aneuploidy (PGT-A) is defined as the analysis of embryos for chromosomal abnormalities before being transferred to the mother's uterus. In order to apply the technique, the embryos are biopsied to extract one/a few cells that will be analysed to know the state of ploidy. If cells are diagnosed as euploid, then the embryo from which they come is classified as "normal" and can be transferred to the mother's uterus to initiate gestation. In recent years, the preferred technique for PGT-A is mass sequencing (NGS). The library elaboration protocol is slightly different with respect to other genetic diagnostic techniques, since the DNA fragments to be sequenced do not originate by fragmentation of the original material but by amplification through random primers that bind to the original DNA. Sometimes, biases occur during the process that generate two possible sources of artifacts: the first, due to the hybridization of random primers in amplified DNA instead of genomic DNA; the second due to amplification of the library after adapter ligation. In both cases, they are PCR duplicates, which can mask the results. The advent of massive sequencing for DGP-A revealed a phenomenon that had gone unnoticed, mosaicism, defined as the presence of two or more different cell lines in an embryo. The percentage of aneuploidy present, the chromosomes and the cell type affected will largely determine the viability of an embryo. It should be noted that the bibliography shows that these anomalies produce low implantation rates in assisted reproduction processes. Furthermore, some studies affirm that the success rates in embryo transfers with a low percentage of mosaicism are actually false positives introduced during the biopsy, amplification and / or analysis process and, therefore, these embryos have implantation rates equivalent to those of a normal embryo. On the other hand, other currents accuse that the techniques present false negatives, alleging the existence of cases where embryos

determined as euploids generated mosaic babies. In any case, all seem to agree that determining the exact percentage of aneuploidy is a critical element in determining the probability of implantation of a mosaic embryo. This makes the development of a precise method for determining mosaicism, controlling levels of sensitivity and specificity, an essential task that has generated great interest in the world of PGD.

We present an algorithm called MiNFilterDups developed for filtering duplicates and PCR artifacts and a second algorithm called MiNmos to detect low percentages of aneuploidy and determine the level of mosaicism. For validation, several sets of files were generated from embryos of patients who had undergone an assisted reproduction process.

Molecular diagnosis can also be applied in the determination of monogenic diseases when the parents are carriers, this technique is known as PGT-M (*Preimplantation Genetic Testing to Monogenic Diseases)*. Several techniques have been evolved, but all of them must face the ADO (*Allele Drop-Out*) phenomenon, known as the preferential amplification of one allele over the other. This effect causes that, when sequencing, a heterozygous locus appears as homozygous because one of the two alleles is not amplified. This phenomenon generates some uncertainty regarding the result issued since if an embryo appears as homozygous and the alteration under study is not detected, it is possible that the mutated allele has not been amplified, constituting a false negative that can trigger the transfer of a diseased embryo. To avoid this problem, generally not only the mutation is studied, but several polymorphisms adjacent to the mutated locus are also analysed. Through a linkage study, it is determined whether these polymorphisms cosegregate with the mutated allele or with the wild one. The study of the mutation is known as direct study; the study of associated polymorphisms, indirect study.

Traditionally, the indirect study has been carried out through the analysis of STRs (short tandem repeats). However, the advantage of using SNPs is the high density that they present throughout the entire genome, which allows embryos to be discarded if the polymorphisms associated with the affected chromosome are present. Sequencing all SNPs in a region is redundant and expensive; Fortunately, the number of SNPs can be reduced by tagSNP strategies (polymorphisms that represent others) based on linkage disequilibrium. There are many parameters that allow us to calculate whether or not one SNP may be useful in predicting others, but in the specific case of selecting useful polymorphisms for PGT-M, consideration of more factors than the correlation between them is required to obtain panels that are useful when it comes to distinguishing recombination phenomena. We present MiNtagSNP, an algorithm for the selection of tagSNPs useful in PGT-M, which allows

to determine the polymorphisms present in each embryo and to discard those that carry the alteration, not only by direct detection, but also for showing the polymorphisms that cosegregate with the alteration in the parents. *In silico* validation was performed using simulated data from the main polymorphism databases. For the *in vitro* validation, results were checked for agreement with traditional methods. Finally, the methodology was implemented in the laboratory, and the cases were followed up over a period of time.

Thus, we can conclude that the main objective of the algorithms presented within the framework of this thesis pursues the design of a fast and efficient analysis method that allows combining the analysis processes of PGT-A and PGT-M by NGS techniques from of a single biopsy.

I. Introducción

Capítulo 1: Conceptos generales

La información genética está codificada por 4 moléculas básicas denominadas nucleótidos. Estos 4 tipos de nucleótidos son la adenina (A), la timina (T), la guanina (G) y la citosina (C), los cuales se ensartan por pares formando los filamentos de ADN. El ADN humano presenta, aproximadamente, 2500 millones de pares de bases (pb), que se organizan en 22 pares de cromosomas autosómicos más un par sexual; en total 46 cromosomas. Las mujeres presentan dos cromosomas sexuales iguales denominados cromosoma X mientras los varones presentan un cromosoma X y uno más pequeño llamado cromosoma Y. Los cromosomas miembros del mismo par se denominan homólogos y, normalmente, cada uno es heredado de un progenitor gracias a que en las células de la línea germinal una célula diploide (46 cromosomas) se divide en un proceso denominado meiosis que origina los gametos (células haploides de 23 cromosomas).

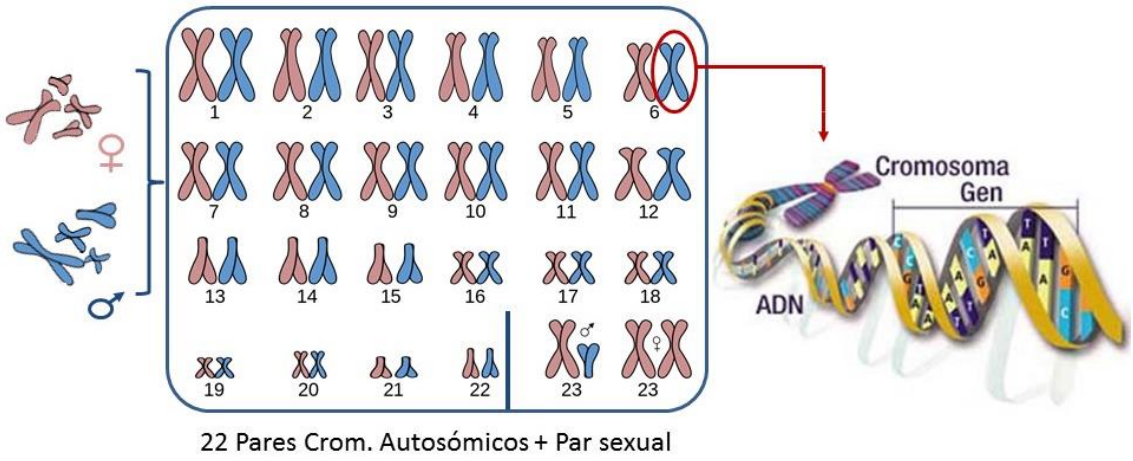

Figura 1: Esquema de la composición del ADN y los cromosomas de un humano. Imagen modificada a partir de ilustraciones de libre distribución de A. Geremia y Linen Tale.

Cuando este gameto haploide se une al gameto haploide del otro progenitor se produce un zigoto diploide que seguirá dividiéndose en un proceso llamado mitosis que dará lugar al feto. En la Figura 1 se puede ver un esquema de la localización de los genes y nucleótidos en la especie humana. La Figura 2 muestra la imagen de un cariotipo normal realizado a una mujer en nuestro laboratorio; en él pueden verse los 23 pares de cromosomas homólogos y la ausencia de cromosoma Y.

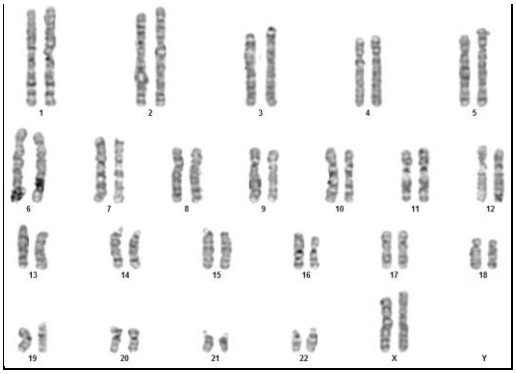

Figura 2: Cariotipo normal 46 XX.

Los genes son pequeños fragmentos de ese ADN que codifican proteínas con funciones específicas dentro del organismo, desde dar lugar al color de los ojos hasta metabolizar los nutrientes básicos en el cerebro. Al poseer un cromosoma de cada progenitor también se posee un alelo de cada uno. La combinación alélica heredada es el genotipo. La expresión de los alelos dará lugar al fenotipo, es decir, al conjunto de caracteres que son visibles como resultado de la interacción del genotipo.

Aproximadamente el 99,9% de la secuencia de ADN es común para toda la especie humana[5,6]. Las alteraciones del patrón del ADN en ese 0,1%, es decir, de la secuencia de nucleótidos respecto a la secuencia considerada como "normal", pueden provocar cambios en la funcionalidad de los genes. Si bien es cierto que algunas alteraciones carecen de importancia fenotípica y se denominan polimorfismos[7], otras pueden ser cruciales en el desarrollo, llegando a ser incluso letales; estas últimas se denominan mutaciones[8]. Una mutación se considera un polimorfismo cuando la frecuencia de su alelo en la población es superior al 1%[6]. Los polimorfismos son responsables de la diversidad existente dentro de la misma especie[9], pero el correcto funcionamiento de todos los genes esenciales es algo básico para el correcto desarrollo del feto y posterior adulto.

1.1 Origen y clasificación de las alteraciones genéticas

Las alteraciones genéticas, tradicionalmente se clasifican en cinco grandes grupos: cromosómicas, monogénicas, mitocondriales, multifactoriales y adquiridas[10].

Las anomalías cromosómicas aparecen en un lado del espectro, ya que presentan manifestaciones fundamentalmente prenatales, siendo responsables de la mayor parte de las pérdidas fetales espontáneas[11]. En el otro extremo encontramos las enfermedades de

origen multifactorial, que presentan una muy baja manifestación prenatal[6] y, desde el punto de vista preimplantacional, difícilmente podrían llegar a abordarse mediante DGP-M debido a la dificultad de su estudio, desconocimiento de los genes implicados y las posibles implicaciones éticas y sociológicas que pueden representar. Las alteraciones monogénicas y mitocondriales estarían en un punto intermedio, siendo las últimas un caso algo peculiar, pues son heredadas exclusivamente por vía materna[12]. Por último, encontramos las alteraciones adquiridas, ampliamente relacionadas con el cáncer, como un grupo a parte debido a que no están presentes en el momento de la concepción sino que se originan por efecto del propio envejecimiento celular y la exposición a factores ambientales tóxicos y de riesgo[13].

1.1.1 Alteraciones cromosómicas

Las alteraciones cromosómicas se clasifican en alteraciones numéricas, que afectan al número total de cromosomas del individuo, y alteraciones estructurales, que afectan la estructura interna de los cromosomas manteniendo el número global del individuo[14]. Se nombran empleando la nomenclatura consenso ISCN, del inglés *International System for Human Cytogenetic Nomenclature*[15].

Su aparición es debida a distintos fenómenos que afectan a los gametos, tales como reordenamientos estructurales en los progenitores, retrasos durante la migración de las meiosis que conllevan pérdidas de cromosomas durante la anafase debidas al cierre prematuro de la pared nuclear dejando fuera un cromosoma que será degradado por las nucleasas del citoplasma, o fenómenos de no disyunción meiótica (proceso por el cual los cromosomas homólogos o las cromáticas de un cromosoma no se separan correctamente durante la meiosis gamética generando isodisomías, es decir, la tenencia de dos cromosomas procedentes de un mismo parental). También pueden producirse *de novo* en el individuo, es decir, aparecer espontáneamente en el zigoto o el feto durante la etapa de división celular[16].

En conjunto, las anomalías cromosómicas son mucho más frecuentes que todos los trastornos mendelianos monogénicos juntos, pues se estima que afectan a 1 de cada 150 recién nacidos vivos, de los que 2/3 sufrirán discapacidad mental o física a consecuencia de la alteración portada[6]. Estas alteraciones son, además, la primera causa de abortos espontáneos[17], por lo que tomando en conjunto ambos datos, se cree que al menos un 10-15% de las concepciones presentan este tipo de alteraciones en madres de edad entre 20 y

24 años[6]. El número de concepciones con alteraciones cromosómicas asciende progresivamente hasta un 51% cuando la madre tiene más de 35 años[6]; de las que el 95% no llega a término[17,18]. La bibliografía recoge que las anomalías cromosómicas están presentes en más del 60% de los abortos espontáneos ocurridos durante el primer trimestre[19], entre un 15-25% de los casos registrados en el segundo trimestre y en un 5-7% de los mortinatos[20–22]. Estas alteraciones provienen, principalmente, de los ovocitos de adultos en edad reproductiva[6].

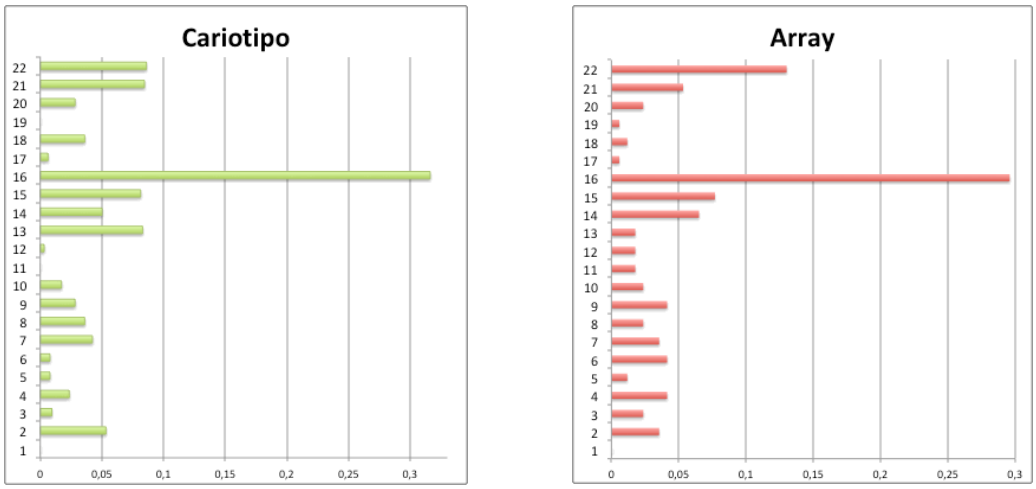

Figura 3: Frecuencia de aparición de las distintas trisomías en abortos espontáneos analizados con técnicas de cariotipado o array.

Podemos observar en la Figura 3 que las frecuencias de aparición de trisomías en abortos espontáneos, derivadas de los estudios publicados por *Warburton et al*. [23] empleando técnicas de cariotipo convencional, coinciden con las frecuencias derivadas del estudio de *Shen et al*. mediante detección por array[24]. La figura evidencia que el cromosoma que aparece con mayor frecuencia alterado es el cromosma 16, siendo en la mayoría de casos el responsable de los abortos espontáneos. Por el contrario, las trisomías viables (aquellas que afectan a los cromosmas 13, 18 y 21) son las más infrecuentes.

1.1.1.1 Alteraciones cromosómicas numéricas

Definimos las anomalías cromosómicas como aquellos casos en que la dosis cromosómica presenta ganancias o pérdidas de uno o más cromosomas respecto a la dosis considerada como normal para el individuo[25]. Podemos dividirlas en poliploidías y aneuploidías en función del número de cromosomas implicados. Así, en las primeras, las

células contienen una o más dosis cromosómicas extra, siempre en múltiplos de 23 para la especie humana, siendo la más común la triploidía (69 cromosomas), totalmente incompatible con la vida[26], si bien en algunas raras ocasiones pueden desarrollarse durante el embarazo e, incluso, dar lugar a un reción nacido vivo que fallece a las pocas horas. Las poliploidías están presente en el 1-3% de todas las concepciones y representan el 15% del total de anomalías cromosómicas observadas en abortos[27]. Normalmente, estas anomalías ocurren de manera espontánea por la fecundación de un óvulo con dos espermatozoides[28], aunque la bibliografía recoge algún estudio que afirma la existencia de cierta predisposición genética a sufrir este fenómeno en la descendencia[27]. La Figura 4 muestra un ejemplo de cariotipo de un feto triploide realizado a un aborto espontáneo en la semana 9+4 (9 semanas y 4 días) de desarrollo embrionario en nuestro laboratorio; puede observarse la presencia de un cromosoma Y.

Por su parte, las aneuploidías son alteraciones donde la dosis cromosómica del individuo presenta un número de cromosomas que no es múltiplo de la dosis haplotípica normal para dicho individuo[6]. La causa más común es la no disyunción meiótica durante la formación de los gametos, tanto en cromosomas autosómicos como sexuales[29]. Según afecte a la pérdida o la ganancia de material estaremos ante una monosomía (pérdida de uno de los cromosomas homólogos), nulisomía (pérdida del par de cromosomas homólogos) o una trisomía (ganancia de un cromosoma en uno de los pares homólogos)[30].

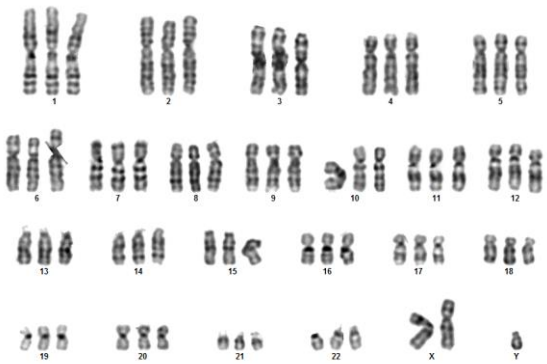

Figura 4: Cariotipo 69 XXY en un aborto espontáneo. Ejemplo de triploidía.

Generalmente, la pérdida o ganancia de material genético de esta magnitud es un fenómeno incompatible con la vida, presente en el 5% de todas las concepciones[31] y mostrando tasas de aparición en abortos espontáneos que ascienden al 50% para fenómenos de trisomía (la más común es la trisomía del cromosoma 16) y al 20% para las monosomías[6]. Sin embargo existen ciertas aneuploidías compatibles con la vida, siendo las

más comunes las trisomías, presentes en el 0,3% de la población y causantes del 35% del total de abortos espontáneos[32].

Las trisomías compatibles con la vida en cromosomas autosómicos son fácilmente reconocibles debido al fenotipo característico que producen y afectan a los cromosomas 13, 18 y 21[33]. La trisomía del cromosoma 13 provoca el Síndrome de Patau, presente en 1 de cada 5000 recién nacidos vivos (RNV)[34]; la trisomía del cromosoma 18 se denomina Síndrome de Edwards y presenta una incidencia de 1 cada 7000 RNV y 1 cada 2600 si la madre presenta edad avanzada[35]. Ambas alteraciones cursan con una esperanza de vida muy baja, apenas superando el año de vida y, en muchos casos, los fetos portadores fallecen *in utero*[34,35]. Por último, la trisomía del cromosoma 21 cursa con el Síndrome de Down, presente en 1 de cada 10000 RNV y cuya incidencia aumenta a 1 cada 750 cuando la madre supera los 35 años[36].

La única monosomía compatible con la vida afecta a los cromosomas sexuales, concretamente al cromosoma X y cursa con el Síndrome de Turner, presente en el 1-2% de las concepciones, aunque el 99% terminan en aborto[6] por lo que se observa finalmente en 1 cada 2500 a 5000 mujeres RNV[37]. En algunos casos nos encontramos con monosomías parciales y el grado de afectación de la portadora depende del grado de inactivación aleatoria del cromosoma no afecto con respecto al afecto[38,39]. La Figura 5a y la Figura 5b muestran un ejemplo de un cariotipo de un paciente trisomía para el cromosoma 21 (a) y otro con monosomía para el X (b).

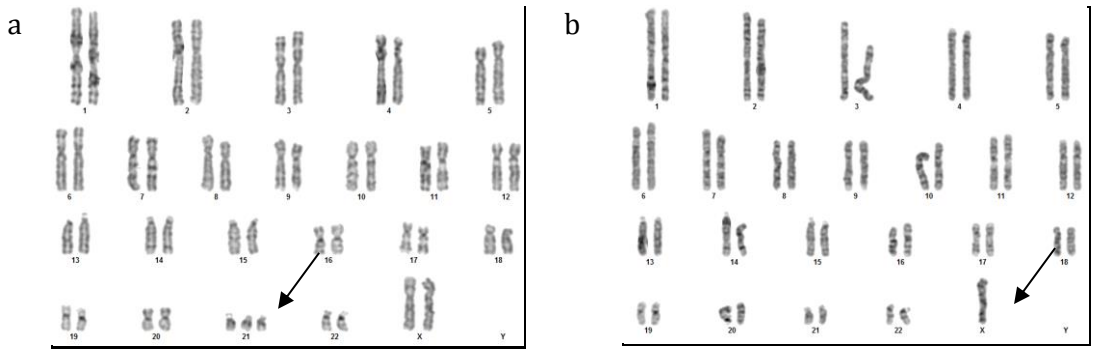

Figura 5: Ejemplo de aneuploidías numéricas a) Trisomía 47 XX +21 S. Down b) Monosomía 45 X0 S. Turner

Además de la monosomía del cromosma X, existen otras aneuploidías que afectan a los cromosomas sexuales. Este tipo de aneuploidías presentan un fenotipo más leve y tienen una incidencia de 1/400 varones RNV y 1/650 mujeres RNV[6], aunque suelen cursar con retraso mental y del desarrollo sexual del individuo entre otros síntomas. Entre ellos encontramos el Síndrome de la triple X o Súper Hembra (47 XXX), que es el más frecuente

de todos con incidencias de aparición de 1 cada 900 mujeres RNV[40]; el Síndrome de Klinefelter (47 XXY) registrado en 1 cada 500-1000 varones RNV[41,42]; y por último, el síndrome de la doble Y o Súper Hombre (XYY) con una incidencia de 1 cada 1000 varones RNV[43].

1.1.1.2 Alteraciones cromosómicas estructurales

Durante la meiosis se pueden producir fenómenos de rotura seguidos de reordenamientos del material cromosómico que generan alteraciones estructurales con tasas de aparición de 1/375 RNV[6]. Estas alteraciones pueden surgir *de novo* o ser heredadas de progenitores portadores. Si el reordenamiento se produce sin pérdida ni ganancia de material genético serán alteraciones equilibradas, mientras que si hay variación de la cantidad de material genético serán alteraciones no equilibradas[44]. Con tasas del 0,4% en muestras prenatales y el 0,2% en neonatos[45,46], las anomalías cromosómicas estructurales equilibradas no suelen mostrar una clínica fácilmente detectable salvo que los puntos de rotura afecten algún gen funcional importante[44], sin embargo, el riesgo aparece en la descendencia, ya que los portadores presentan una probabilidad de 1/2 de generar gametos con alteraciones no equilibrada[6]. La consecuencia más frecuente de esto son abortos recurrentes, pero también es común tener descendencia gravemente afecta, con retraso metal y rasgos dismórficos entre otras caracterstícas fenotípicas. Entre las alteraciones cromosómicas estructurales más frecuentes encontrarmos traslocaciones, deleciones, duplicaciones e inversiones, pero existen también cromosomas en anillo, e isocromosomas, entre otras.

Las traslocaciones se producen a consecuencia de la existencia de un punto de rotura y el intercambio del material genético seccionado[47]. Si la rotura se produce en dos cromosomas que intercambian mutuamente el material, la traslocación se denomina traslocación recíproca y los cromosomas pasan a ser llamados cromosomas derivados[47]. La incidencia poblacional de este tipo de traslocaciones es del 0,14%[48]. Cuando los puntos de rotura se encuentran muy cerca o dentro del centrómero de dos cromosomas acrocéntricos, se produce la fusión de los brazos largos de dichos cromosomas y pérdida de los cortos en una fenómeno denominado traslocación robertsoniana[49] cuya incidencia asciende al 12%[48]. Cabe destacar que la pérdida del material genético situado en los brazos cortos de estos cromosomas acrocéntricos no presenta significado clínico ya que en esa región no presentan genes esenciales, sino pseudogenes y duplicaciones de otros genes localizados en

otras regiones, por lo que suele considerarse un reordenamiento balanceado a pesar de que el número de cromosomas final sea 45[6]. La Figura 6 muestra un ejemplo de estos dos tipos de traslocaciones.

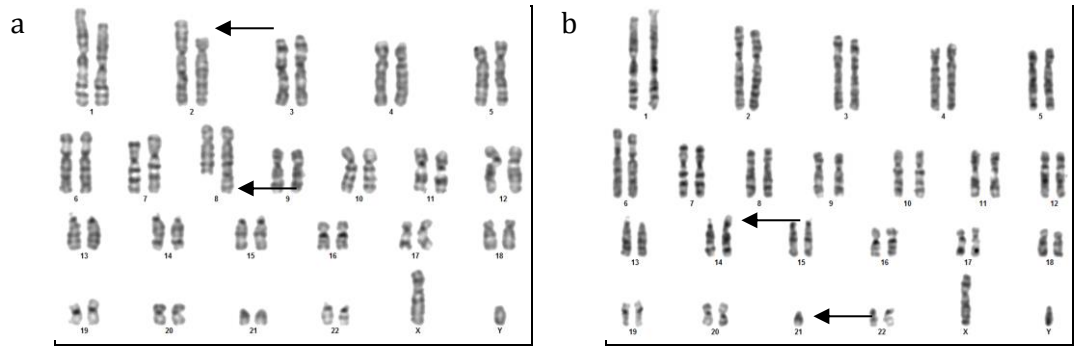

Figura 6: Ejemplo de traslocaciones a) Recíproca entre los cromosomas 2 y 8 b) Robertsoniana entre los cromosomas 14 y 21.

A pesar de que los portadores de este tipo de alteraciones suelen ser fenotípicamente normales, tienen un gran riesgo de engendrar embriones afectos de patología genética o, directamente, no viables debido a los gametos desbalanceados que se pueden producir durante la meiosis; particularmente, un estudio afirma que las traslocaciones presentan un riesgo de descendencia anómala entre el 1 y el 20% en función del tamaño del segmento implicado[50], pero ascendería hasta cifras teóricas del 50% para una translocación balanceada.

Las deleciones y duplicaciones son, respectivamente, pérdida y ganancia de material genético en un cromosoma, provocando una monosomía o trisomía génica del segmento afectado, que sufre haploinsuficiencia o trisomía parcial[51]. Si bien en la mayoría de las ocasiones son alteraciones pequeñas y que no causan patología genética alguna (consideradas incluso como polimorfismos, por su elevada freuencia en la población), en ocasiones nos encontramos con alteraciones de gran tamaño (detectables mediante cariotipo), con una frecuencia de 1 de cada 7000 RNV[6]. Estas últimas suelen proceder del reordenamiento meiótico de un portador de traslocación y presentarse juntas a causa del reordenamiento, de manera que un segmento del cromosoma aparece duplicado mientras otro segmento presentará una deleción[51]. La significancia clínica dependerá del número y función de genes afectados.

Los isocromosomas son cromosomas que han perdido el brazo corto y cuyo brazo largo se ha duplicado[52]. La causa más común de formación la encontramos durante la división meiótica, donde el centrómero se divide según el plano transversal en vez del seguir el plano vertical; como consecuencia de este suceso uno de los brazos del cromosoma

original se pierde y resulta en un cromosoma cuyos brazos son idénticos entre sí pero en sentido inverso[52].

Las inversiones son fenómenos originados dentro de un mismo cromosoma tras una doble rotura a la que sigue un giro de 180 grados del fragmento sustraído y una reinserción en sentido inverso[53]. Se clasifican de acuerdo a la localización de los puntos de rotura con respecto al centrómero. Así, si la región invertida contiene el centrómero, cada punto de rotura se sitúa en uno de los brazos y la inversión se denomina pericéntrica. En caso contrario los puntos de rotura se presentan en el mismo brazo cromosómico, y la inversión es denominada paracéntrica[53]. Nuevamente, al no existir pérdida ni ganancia de material, los portadores serán indistinguibles clínicamente, sin embargo existe una tasa de riesgo de descendencia afecta que asciende al 5-10% dependiendo del tamaño de la región invertida[6]. También debemos señalar que algunas deleciones no generan problema alguno, como el caso de la inversión pericéntrica del cromosoma 9, frecuente en el 1% de la población y que es considerada un polimorfismo benigno sin significancia clínica reseñable[54].

Las inserciones se producen debido a la inserción de un segmento cromosómico en otra parte del genoma. La significancia clínica dependerá de si hay ganancia de material genético o interrupción de genes y, en caso de haber variaciones en la cantidad de ADN, del tamaño y los genes afectados[55].

Por último, los cromosomas en anillo se originan debidos a la rotura de los dos extremos de un cromosomas, la deleción de los segmentos terminales y la unión de la porción central[56]. Son fenómenos poco frecuentes que generan muchos problema en la meiosis celular y la consecuencia fenotípica depende del tamaño del segmento delecionado, aunque generalmente se trata de alteraciones graves. Además el riesgo de transmisión asciende al 40%[6].

Así, aunque por definición algunas de estas alteraciones no presentan una significancia clínica relevante, en términos reproductivos la segregación anormal que experimentan los cromosomas involucrados en este tipo de anomalías durante la meiosis provocan altas tasas de infertilidad y abortos espontáneos así como altas probabilidades de engendrar descendencia afecta por anomalías congénitas. De hecho, se ha observado una prevalencia de enfermedades congénitas 25 veces superior en parejas subfértiles con alteraciones cromosómicas estructurales con respecto al resto de la población.[57]

1.1.2 Alteraciones monogénicas

Las alteraciones monogénicas se definen como cambios en la secuencia del ADN que no afectan a la estructura de los cromosomas[58]. Actualmente, el *Online Mendelian Inheritance in Man: OMIM* describe más de 23000 variantes monogénicas distintas[59], aunque no todas producen enfermedad ya que son consideradas polimorfismos benignos que contribuyen a la diversidad genética. Alrededor del 1% de RNV presenta alguna de las variantes clasificadas como patogénicas[60]. Actualmente, ClinVar recoge más de 100.000 mutaciones clasificadas como patogénicas o probablemente patogénicas[61].

Las alteraciones monogénicas más frecuentes se deben a sustituciones, deleciones, inserciones y a la repetición de secuencias[58]. Cuando las tres primeras afectan a un solo nucleótido se denominan mutaciones puntuales. Estos cambios en el ADN pueden ser originados por errores en los mecanismos de replicación y reparación del ADN o bien por fenómenos ambientales. La consecuencia fenotípica es diferente según el nucleótido afectado y el cambio producido[62]. Estos cambios pueden clasificarse en:

- Mutación por cambio de sentido: En muchos casos se traduce en un cambio en un aminoácido de una proteína.
- Mutación sin sentido: Puede aparecer un codón de terminación que interrumpe prematuramente la formación de una proteína.
- Splicing: Se produce una modificación de los puntos de corte y empalme (conocido normalmente por su término en inglés).
- Inserciones y deleciones de un nucleótido dando lugar a cambios en el marco de lectura y, con ello, a proteínas anómalas.

Asimismo, la variación de la funcionalidad será diferente según la alteración se produzca en la región promotora del gen (influenciando la actividad transcripcional del gen), en los intrones (modulando la estabilidad de la proteína), en los sitios de *splicing* (modulando la eliminación de intrones y unión de exones) o en regiones intragénicas[63,64].

1.1.2.1 Patrón de herencia.

Una característica importante que influye en que las alteraciones monogénicas cursen o no con una clínica compatible con patología se basa en el patrón de herencia presentado[65].

- Autosómicas dominantes: aparecen en 1 de cada 200 individuos[6]. En este caso, el gen con la mutación se sitúa en uno de los 22 cromosomas autosómicos y tan solo necesita portar una dosis del alelo afecto para manifestar la clínica característica de la enfermedad. Normalmente se presenta en todas las generaciones de una familia y los padres de un afecto serán siempre afectos a excepción de casos con penetrancia incompleta. La penetrancia es el porcentaje de individuos con un genotipo específico que cursan el fenotipo esperado. Será completa si todo individuo con el genotipo presenta dicho fenotipo, de lo contrario se denominará penetrancia incompleta[66]. Los hijos de un afecto tendrán un 50% de probabilidad de padecer la enfermedad[65].
- Autosómicas recesivas: son modificaciones de la secuencia de genes que también se sitúan en uno de los 22 cromosomas alterados, pero en este caso es necesario portar las dos dosis del alelo afecto para manifestar la clínica característica. Por tanto, ambos parentales deben portar dicho alelo, bien en homocigosis, siendo afectos, o en heterocigosis, siendo sanos. Los descendientes de dos heterocigotos portadores sanos tendrán un riesgo del 25% de padecer la enfermedad al heredar los dos alelos mutados, un 50% de probabilidad de ser portador de alguno de los alelos mutados pero portar otro sano y un 25% de portar ambos alelos sanos. A veces los alelos afectados no son exactamente iguales, pero afectan al mismo gen por complementación generando un descendiente afecto por heterocigosis compuesta[58].
- Recesivas ligadas al cromosoma X: se producen por mutación de un gen localizado en la parte diferencial del cromosoma X con respecto al Y. Todos los varones portadores presentarán la afectación debido a la hemicigosis del cromosoma X respecto al Y (solo hay una dosis alélica del cromosoma, en este caso causante de enfermedad, por lo que no existe una segunda dosis alélica sana que pueda contrarrestar el efecto). En el caso de las mujeres solo las portadoras de ambos alelos afectos manifestarán la enfermedad, aunque las portadoras de una única dosis podrían manifestar una clínica en menor grado debido a la inactivación aleatoria del cromosoma X[38]. El 100% de la descendencia de un varón afecto será sana, ya que las hijas heredarán el alelo sano por parte de madre y los hijos nunca heredan el cromosoma X del padre. Por el contrario, las hijas de una mujer portadora tendrán un 50% de

probabilidades de ser portadoras sanas y un 50% de no portar el alelo, mientras que los hijos serán enfermos en el 50% de los casos y sanos en el otro 50% por no portar el alelo [58].

- Dominantes ligadas al cromosoma X: se localizan en la misma región descrita para las recesivas y tan solo es necesaria una dosis del alelo para manifestar la clínica, por lo que las mujeres pueden ser afectas. Sin embargo, en varones es letal en muchos casos.. La descendencia de una madre afecta se verá afectada en el 50% de los casos y sana en el otro 50%; por su parte, las hijas de un varón enfermo serán siempre enfermas, mientras que los hijos serán todos sanos[58].

También podemos encontrar otros patrones de herencia menos comunes como la pseudodominancia, que es la dominancia aparente de un alelo recesivo debido a la deleción del alelo en el otro cromosoma[67]. Este fenómeno es similar a la hemicigosis sufrida por los varones en los cromosomas sexuales. La codominancia es definida por aquellos casos en que el fenotipo viene determinado por la expresión conjunta de ambos alelos[68]. Y por último las enfermedades ligadas al sexo, que afectan a genes autosómicos en los que el alelo patogénico se comporta como dominante o recesivo según el portador sea varón o mujer.

1.1.2.2 Clasificación de alteraciones monogénicas

Las principales alteraciones monogénicas, patogénicas o no, son los polimorfismos de secuencia repetida (VNTR, del inglés *Variable Tandem Number Repeat*)[69,70] y los polimorfismos de nucleótido simple (SNPs, del inglés *Single Nucleotide Polimorphism*).

Los VNTR presentan un número variable de repeticiones en tándem. Los minisatélites corresponden a la repetición de unas pocas decenas de nucleótidos[6], mientras que los microsatélites (STRs del inglés *Short Tandem Repeats*) corresponden a la repetición de entre 2 y 5 nucleótidos[70]. La ventaja de este tipo de polimorfismos es que cada loci puede presentar muchos alelos distintos (tantos como repeticiones presente) con frecuencias muy similares entre sí por lo que la probabilidad de que un individuo sea heterocigoto es muy elevada[71]; además se distribuyen a lo largo de todo el genoma.

Por su parte, los SNPs son modificaciones de un único nucleótido en la secuencia de ADN, que se mantienen y heredan[72]. Según la localización se clasifican en cSNPs (situados en regiones codificantes), rSNPs (regiones reguladoras) y gSNPs (regiones intergénicas)[6].

Los SNPs presentan una menor tasa de mutación, lo que los convierte en las dianas perfectas para ser empleados en estudios poblacionales[73]. Otra ventaja respecto a los polimorfismos de repetición es que los SNPs aparecen en alta densidad repartidos por todo el genoma[73]. Actualmente hay más de 10 millones de SNPs descritos en las bases de datos[74] y se cree que existen aproximadamente 50 millones de SNPs comunes, es decir, SNPs cuyo alelo de menor frecuencia presenta una frecuencia superior al 1%[75].

1.2 El diagnóstico genético preimplantacional

1.2.1 Definición

Podríamos establecer el origen de la genética médica en 1902, momento en que Alfred Baring Garrod reconoce en "*The incidence of Alkaptonuria: a study in chemical individuality*"[76] que las leyes de Mendel, descritas en 1865, pueden ser empleadas para explicar la presencia de alcaptonuria en una familia[76]. Desde entonces y tomando como base los avances del descubrimiento del ADN por Rosalind Franklin y Maurice Wilkins y su posterior descripción por James Watson y Francis Crick, la técnica ha ido evolucionando hasta convertirse en una auténtica especialidad.

Tras el nacimiento el 25 de Julio de 1978 del primer niño fruto de la reproducción asistida[77], la bibliografía recoge el rápido desarrollo de las técnicas de selección de embriones aptos para engendrar individuos sanos [77] durante la década de los 80, hasta que en 1990 se evidencia la primera aplicación exitosa del diagnóstico genético preimplantacional (DGP) en humanos a través de la selección de los cromosomas sexuales[78].

La técnica del DGP surge entonces como alternativa al diagnóstico prenatal para parejas en riesgo de tener descendencia afecta por alteraciones cromosómicas y/o enfermedades monogénicas[2] que desean evitar la necesidad de recurrir a terminaciones voluntarias del embarazo, para lo cual se someten a un proceso de reproducción asistida en un ciclo de fertilización *in vitro* (IVF)[79]. Un ciclo de IVF con DGP comúnmente comprende las siguientes etapas, que podemos ver reflejadas en la Figura 7:

- Estimulación ovárica.
- Aspiración de los folículos ováricos y recuperación de los oocitos.
- Fecundación con espermatozoides previamente capacitados.
- Cultivo de los oocitos y biopsia.
- Realización de los análisis genéticos procedentes de acuerdo a la clínica presentada por la pareja.

- Transferencia de los embriones aptos y vitrificación de aquellos aptos que no hayan sido transferidos.

Así pues, solo aquellos embriones que, tras la realización de los análisis genéticos pertinentes, son diagnosticados como no portadores de alteración genética serán calificados como potencialmente transferibles y podrán engendrar un bebé sano[2]. Cabe destacar que el DGP se diferencia de las técnicas de diagnóstico genético aplicadas a otros campos en 2 características principales. En primer lugar, el tiempo de respuesta debe ser mucho menor, pues en muchos casos los resultados deben ser obtenidos en menos de 24 horas para permitir la transferencia de los embriones dentro del mismo ciclo, lo que se conoce como transferencia en fresco[80]. En segundo lugar, cada pareja produce entre 6 y 10 embriones en promedio que deben ser procesados y analizados, lo que incrementa el costo global y el tiempo de dedicación[81]. Estas dos características definitorias hacen del DGP un campo en constante crecimiento, promoviendo la búsqueda de nuevas técnicas que abaraten los costes y el tiempo necesario para la obtención de los resultados.

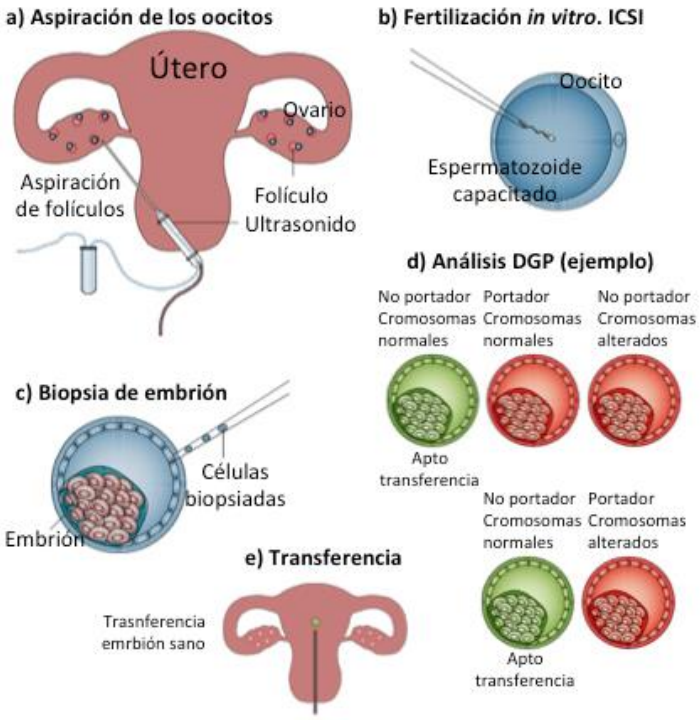

Figura 7: Esquema proceso de DGP. Imagen modificada a partir de [82]

Se estima que entre el 2 y el 3% de los recién nacidos presenta alguna alteración genética que ocasiona discapacidad, retraso mental, y/o muerte precoz[6]. Este porcentaje era antiguamente mucho mayor, pero, la aplicación de las técnicas de DGP ha logrado disminuir la tasa en los últimos años[82]. Como se ha mencionado anteriormente, sin tener en

cuenta la etiología genética de las pérdidas fetales y mortinatos, las alteraciones genéticas representan entre el 10 y el 30% de los ingresos hospitalarios pediátricos en países desarrollados ya que son responsables del 40-50% de la mortalidad infantil, el 50% de las cegueras y sorderas y más del 50% de los casos de retraso mental[6]. Sin embargo, el DGP sigue siendo una técnica costosa que no están al alcance de todos, por lo que el desarrollo de técnicas asequibles a la vez que rápidas y eficaces es un interés constante entre los que se dedican a este campo.

Finalmente, cabe destacar que una de las principales limitaciones de la aplicación del DGP, en comparación con otras técnicas de diagnóstico genético, consiste en la baja cantidad de ADN disponible, procedente de una o unas pocas células. Por ese motivo, generalmente, se requiere la utilización de técnicas de amplificación de genoma completo. Además, esto puede desembocar en la obtención de resultados erróneos[83] debido a dos factores principlamente: la presencia de mosaicismo y *allele drop-out* (o amplificación preferencial de un alelo). Las técnicas de DGP actuales deben ir encaminadas a evitar ambas fuentes de error.

1.2.1.1 Biopsia

La biopsia comprende el proceso de extracción de una o varias células que serán analizadas para determinar si el embrión del que procede es o no transferible[84]. La literatura describe tres procedimientos básicos de biopsia según el momento del desarrollo del embrión en que se realice.. Cada metodología de extracción tiene ventajas específicas y limitaciones críticas que deberán ser tenidas en cuenta a la hora de emplear un procedimiento u otro.

Para facilitar el entendimiento del problema, en la Figura 8 se muestra un esquema de la formación de los gametos femenino y masculino, la fecundación y la posterior formación del zigoto. A simple vista se puede observar que el producto de la gametogénesis masculina consiste en 4 espermatozoides, mientras que la gametogénesis femenina resulta en un solo óvulo, que completa su maduración y proceso de división meiótica una vez ha sido fecundado. Los subproductos formados durante la división del gameto femenino se denominan corpúsculos polares y no formarán parte del futuro cigoto.

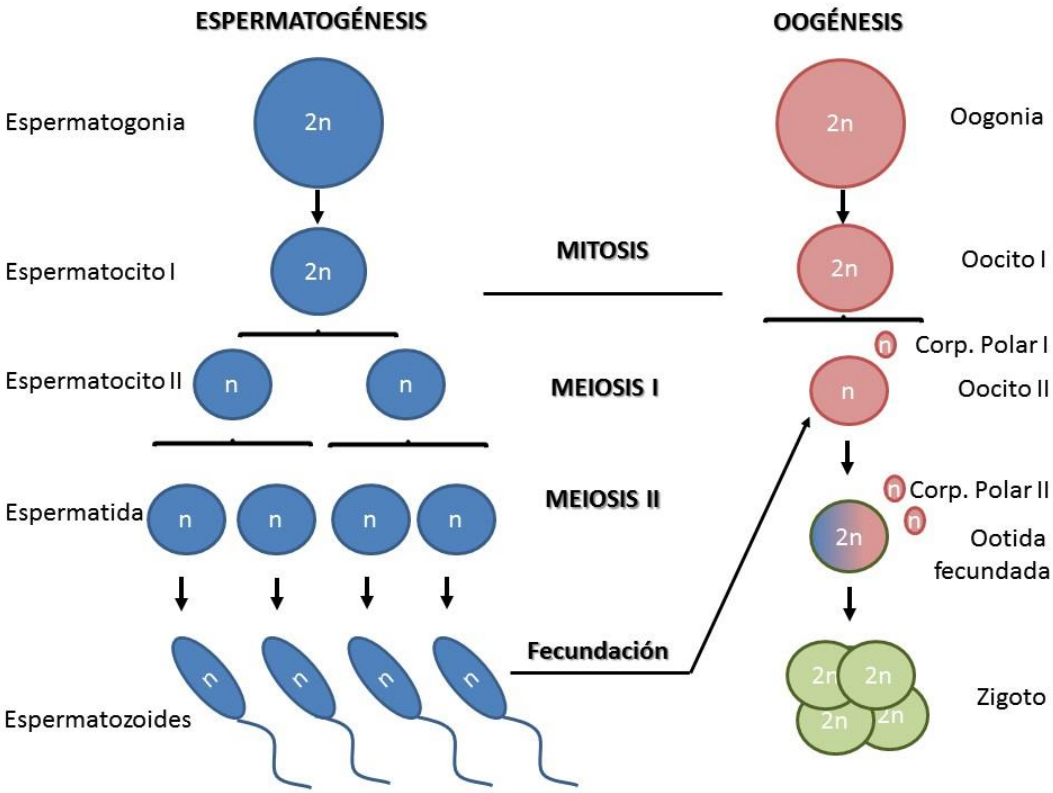

Figura 8: Esquema de la gametogénesis femenina y masculina, además de la formación del cigoto.

En primer lugar, la biopsia de corpúsculos polares es una opción cuando las mutaciones a evaluar son de origen materno[85]. A pesar de la evidente limitación que supone tener información solo del ADN materno, resulta interesante considerar la aplicación de la técnica en países donde las restricciones ético-políticas limitan la obtención y/o análisis de las células embrionarias[85]. Para su realización debemos tener en cuenta que los corpúsculos polares surgen como producto de la meiosis en fase I o fase II, por tanto, su extracción no interfiere en la capacidad reproductiva del oocito, convirtiendo esta técnica en la menos invasiva de las opciones disponibles[85]. Sin embargo tiene ciertas limitaciones. Como ya se ha mencionado, solo es útil en la detección de mutaciones procedentes de la línea materna[85]. Así, frente a enfermedades recesivas, un análisis realizado sobre biopsia de corpúsculo polar solo será capaz de discernir si el embrión porta el alelo sano o el causante de enfermedad materno. De esta manera, se descartarán embriones que porten el alelo mutado materno, sin considerar el paterno, lo que se traduce en que hay un 50% de posibilidades de desechar embriones válidos (aquellos que, portando el alelo mutado materno, resulten portadores del alelo sano paterno). Esta incertidumbre plantea serios dilemas morales[86]. Otra característica de la biopsia de corpúsculo polar es que requiere el análisis de ambos corpúsculos para tener un resultado de las alteraciones que puede tener el ovocito, lo que duplica el coste de la prueba frente a otras alternativas. Por último, un

hecho que motiva a decantarse por otras técnicas reside en que en el 45% de los corpúsculos se observa un fenómeno de no disyunción prematura de las cromátidas hermanas de los corpúsculos de fase I, provocando una sobreestimación de las aneuploidías del embrión y, con ello, el descarte de embriones potencialmente sanos[86]. Este hecho es especialmente importante en parejas de cuyos ciclos IVF se obtuvieron pocos embriones.

A partir de 1988, momento en que se registra la idea de que la extracción de una célula a partir de un embrión de 8 células no resulta perjudicial para el correcto desarrollo del mismo[78], una segunda opción que fue ganando adeptos consiste en realizar la biopsia de una blastómera en día 3 del desarrollo embrionario[87]. Una vez excluidos los embriones que en los exámenes morfológicos muestran baja calidad, se puede realizar una biopsia durante las primeras horas del día 3 del desarrollo embrionario, cuando tiene de 6 a 8 células. Para ello, se realiza un orificio en la zona pelúcida mediante ácido Tyrodes o con el empleo de un pulso láser, y se procede a la extracción por aspiración ("captura") de una de las blastómeras. En este caso, cualquier célula puede ser escogida para ser biopsiada, ya que el embrión aún no se ha diferenciado y puede continuar su desarrollo "reponiendo" la falta.

La principal dificultad de la técnica atañe a la obtención exitosa del material analizable, pues la blastómera debe ser extraída intacta y, además, contener tan solo una copia del material genético original, es decir que, a la hora de elegir la blastómera a capturar, se debe escoger aquella que presente claramente visible, un solo núcleo[88]. Además, el protocolo de fechas[88] debe ser rigurosamente respetado para evitar alterar el desarrollo del embrión y asegurar el éxito de la biopsia.

La ventaja de las técnicas de biopsia de células procedentes del embrión (de este procedimiento como del descrito a continuación) reside en que tanto el material materno como el paterno es analizado. Además, otra ventaja de los embriones biopsiados en día 3 estriba en que, si se realiza el DGP en 24 horas, estos pueden ser transferidos en día 5 del estadio embrionario sin la necesidad de hacer realizar un segundo ciclo de estimulación para la recepción de los embriones ni de vitrificarlos, lo que reduce los costes del proceso. Esto es lo que se conoce como transferencia en fresco. En caso de no desear transferir inmediatamente, los embriones pueden ser vitrificados y guardados para ser transferidos en un ciclo posterior[89]. Durante años la biopsia en día 3 fue el método escogido debido a estas ventajas y al hecho de que las blastómeras son sencillas de manejar, por lo que no se requiere ningún requerimiento especial.

Como contrapartida, la biopsia en día 3 se presta a la discusión de cuán fiel es el material genético de dicha célula al material del resto del cromosoma. Esta incógnita se basa en la existencia de fenómenos de mosaicismo, es decir, la presencia de 2 o más dotaciones cromosómicas diferentes en un organismo. Así, al analizar el material procedente de una biopsia en día 3 se corre el riesgo de que la célula analizada no represente al resto del embrión[90].

La tercera opción es la biopsia de trofoectodermo en día 5 ó 6 del desarrollo embrionario[91]. La principal ventaja de la biopsia de trofoectodermo es que, si se realiza correctamente, no compromete en absoluto el desarrollo embrionario[91]. Además, queda demostrado por la literatura en los numerosos casos registrados de mejora de las tasas de éxito en la implantación frente a otros procedimientos [91,92].

En este caso el embrión es cultivado hasta el día 5 del desarrollo evolutivo, cuando se presenta en fase de blastocisto. Una vez cultivados, se seleccionan los embriones biopsiables en función de su morfología. En general, se biopsian aquellos que presentan una morfología adecuada para su estadío, donde se pueda diferenciar claramente la capa del trofoectodermo de la masa celular interna. Cabe destacar que el trofoectodero formará posteriormente las estructuras placentarias, mientras que la masa celular interna formará el feto. Por tanto, la biopsia de trofoectodermo se considera menos agresiva que la biopsia en día 3 porque no se están manipulando las células que formarán al propio feto. Para la realización de la biopsia, en primer lugar se debe perforar la zona pelúcida mediante láser, dado que ahora se requiere mayor precisión al ser ésta más delgada que en día 5. Tras esto, se aspiran 4-5 células del trofoectodermo, ayudándonos de pulsos láser para separar las células. Además, las células biopsiadas deben estar suficientemente alejadas de la masa celular interna[93], para evitar cualquier posible daño ya que, en esta etapa, el embrión ya está suficientemente diferenciado y el daño causado por la sustracción de una célula podría ser irreparable.

Por tanto, las técnicas de biopsiado en día 5 requieren de personal altamente cualificado, equipamiento especializado como un láser, así como de un programa de vitrificación embrionario para conservarlos y que ofrezca garantía plena de que los embriones biopsiados sobrevivirán hasta el momento de mayor receptividad endometrial en la madre[94], ya que no es viable la transferencia más allá de esta fase embrionaria de desarrollo en día 6 y, normalmente, las clínicas no optan por la transferencia dentro del mismo ciclo ovárico debido al corto espacio de tiempo disponible para la obtención de resultados.

La ventaja de la biopsia en día 5 reside precisamente en que se sustraen varias células, entre 4 y 10 células[95], e incluso si se extrajese alguna más, el embrión continuaría siendo viable siempre que la masa celular hubiese quedado intacta. El análisis de un mayor número de células permite un diagnóstico más preciso, dado que aumenta la posibilidad de detectar un posible mosaicismo[77]; de hecho, el ratio mínimo de aneuploidía necesario para detectar una aneuploidía se establece en que, al menos, el 50% de las células biopsiadas presenten la misma aneuploidía, aunque hay estudios que afirman que es posible discernir cierto desplazamiento del perfil respecto a la normalidad cuando un 25% de las células lo presentan[77].

Sin embargo, sigue existiendo cierta controversia respecto al grado de identidad entre las células biopsiadas y el embrión; algunos autores aseguran que la concordancia es suficientemente alta[96], mientras otros niegan que exista tal relación[4,95].

Una vez realizada la biopsia se procede al análisis genético de las posibles alteraciones presentadas por los embriones. Los embriones cuyo resultado sea compatible con la normalidad serán clasificados como potencialmente transferibles y podrán engendrar un individuo sano.

1.2.2 Aspectos éticos

Debido a la controversia ética que genera la posibilidad de engendrar "bebés a medida", el DGP se encuentra actualmente bajo diferente regulación en función del país en el que se practique, hasta el punto de haber sido prohibido en algunas regiones[85]. Por ejemplo, ciertos países consideran "más aceptable" la idea de realizar una terminación voluntaria del embarazo en el segundo trimestre de desarrollo embrionario, tras la realización de una prueba prenatal, que la realización de una evaluación preimplantacional que conlleve la "elección" sobre el futuro bebé.

Dado que el DGP es realizado antes de la implantación del embrión, puede ser empleado, siempre que sea genéticamente posible, para seleccionar embriones libres de enfermedad al evitar la transferencia del alelo causal, de la misma manera que, potencialmente, podría escogerse el alelo adecuado al color de los ojos o el pelo. También permite, por ejemplo, que un futuro bebé sea compatible con un hermano enfermo. Por otro lado, el DGP puede asegurar que el bebé nacido sea escogido en función de los cromosomas sexuales de manera que se asegure que no sufrirá una enfermedad ligada al cromosoma X,

pero también podría seleccionarse para que pertenezca al género socialmente "más aceptado".

Una encuesta realizada en 2016 en la Universidad de Harvard informó que el 83% de los encuestados estaba en contra de la manipulación genética para mejorar el intelecto o las capacidades físicas en un futuro bebé[97]. Por otro lado, a pesar de que normalmente los padres escogen embriones libres de la alteración portada para evitar su transmisión al futuro bebé, según una encuesta realizada en 190 clínicas de Estados Unidos por la Universidad John Hopkins y publicada en *The New York Times*, en un sorprendente 3% de los casos los padres emplearon esta técnica para escoger un embrión con la misma alteración genética que ellos padecían (por ejemplo sordera)[98].

El alcance actual del DGP permite pues la selección de los embriones en tanto convenga a la situación de los progenitores, pero siempre debe evitarse la selección de "bebés a la carta". Cada país presenta su propia normativa en función de la consideración ética que conlleva[99], algo que debemos tener siempre en cuenta a la hora de diseñar un análisis DGP para evitar posibles abusos de la técnica. Además, es esencial realizar un consejo genético adecuado[100] teniendo siempre presente que el fin principal del consejo genético consiste en permitir que personas portadoras de alteraciones genéticas puedan reproducirse y vivir de la forma más normalizada posible[100].

En España, la regulación del DGP-M se establece a través de la ley 14/2006, de 26 de mayo, sobre técnicas de reproducción humana asistida. Dicha ley recoge la necesidad de contar con la autorización de la Comisión Nacional de Reproducción Asistida, órgano colegiado dependiente del Ministerio de Sanidad, Servicios Sociales e Igualdad, y que tiene encomendada la función de realizar informes preceptivos previos ante una serie de supuestos que contempla el artículo 20.4 de la citada Ley. A grandes rasgos, sólo se autoriza DGP-M para aquellas enfermedades que cumplan los siguientes requisitos:

1. Enfermedad hereditaria crónica, severa y/o progresiva.
2. Exista un informe genético previo donde se establezca claramente la causa de la patología.
3. Exista un riesgo genético conocido de transmisión.

Esta autorización debe solicitarse caso a caso, y es el comité quien decide si se puede realizar o no. De esta manera, se evitan posibles abusos y la utilización poco ética para las técnicas de DGP-M.

El DGP-A tiene una regulación algo más laxa, y tan sólo es necesario que haya causa médica justificada para que se pueda realizar, sin necesidad de solicitar autorización a la Comsión Nacional. Aunque mediante esta técnica es posible conocer el sexo del embrión, está prohibida la utilización de esta información para la selección de embriones e incluso facilitar a los pacientes la información sobre el sexo del embrión transferido.

1.2.3 DGP-A

A pesar de que actualmente se estima en más de 5 millones los bebés nacidos como resultado de la aplicación de técnicas de reproducción asisitida[101], el *U.S. Department of Health and Human Services and the Centers for Disease Control* afirmó recientemente que cerca del 51.9% de los procesos de reproducción asistida en los que no se emplea donante no producen una gestación exitosa[102,103]. Frecuentemente, las clínicas tratan de mejorar las tasas de implantación y embarazo a través de la implantación de más de un embrión al útero materno[104]. Este hecho provoca, evidentemente, el aumento del riesgo de embarazo múltiple a la vez que favorece la gestación de embarazos ectópicos, abortos espontáneos, partos prematuros, gestaciones de bajo peso y otras complicaciones de salud que se reflejan tanto en la madre como en los fetos gestados[105]. Sin embargo, cada vez son más las clínicas que están abandonando esta práctica para optar por la transferenica de un embrión único. Para que las tasas de embarazo por transferencia embrioniaria no se vean afectadas, una buena estrategia consiste en la selección de aquellos embriones que presentan mejor pronóstico con base en criterios como la morfología[106–108], la velocidad de división celular[109–111] y el análisis de alteraciones cromosómicas[104].

Como se ha explicado anteriormente, el cribado genético preimplantacional, antiguamente denominado PGS por sus siglas en inglés, *Preimplantation Genetic Screening* y actualmente conocido como diagnóstico genético preimplantacional de aneuploidías (DGP-A)[112], se define como el procedimiento que permite analizar si los embriones presentan anomalías cromosómicas numéricas antes de ser transferidos al útero materno[1–3]. Para poder aplicar la técnica los embriones son obtenidos a partir de pacientes que se someten a ciclos de fecudación *in vitro*; estos embriones son biopsiados para extraer una/unas pocas células que serán analizadas para conocer el estado de ploidía[77]. Si son diagnosticadas como euploides, entonces el embrión del que proceden es catalogado como "normal" y puede ser transferido al útero materno para iniciar la gestación[113].

El DGP-A fue introducido en la práctica clínica con la intención de mejorar las tasas de éxito en parejas con problemas de fertilidad, basándose en la idea de que la alta tasa de aneuploidías encontradas en las etapas tempranas de desarrollo de los embriones de estas parejas podían ser las responsables de las bajísimas tasas de éxito que este tipo de parejas experimentaban al exponerse a técnicas convencionales de reproducción asistida[77]. Así, existen datos que aseguran que la aplicación de estas técnicas permitió disminuir el tiempo necesario para engendrar descendencia en parejas subfértiles de los 4-6 años a menos de 4 meses[114], dependiendo de la indicación. Se considera una pareja subfertil a aquella pareja en edad reproductva que presenta dificultades para la concepción después de un año manteniendo relaciones sexuales habituales sin el uso de anticonceptivos[115].

Pero el DGP-A no solo se desarrolló con la intención de mejorar las tasas de implantación y embarazo de las parejas que se sometían a ciclos IVF, sino para disminuir las tasas de abortos espontáneos[3] y el riesgo de tener descendencia afecta por una cromosomopatía o requerir una terminación voluntaria del embarazo[78,116]. Se observó que la incidencia de abortos espontáneos en dichas parejas subfértiles pasó del 90% a ser inferior al 15%[114].

El caso de embarazos de madres de edad avanzada es otra de las indicaciones más comunes para la aplicación de los estudios de DGP-A, pues estudios recientes muestran que, a medida que aumenta la edad materna aumenta la aparición de aneuploidías en los embriones generados, pasando del 20 al 60% en mujeres de más de 35 años[117]. Pero no solo los factores maternos son importantes, pues se ha observado que la incidencia de espermatozoides aneuploides es mayor en varones con problemas de concepción[118], lo cual puede explicar las elevadas tasas de aborto y baja implantación que se observan en algunas parejas con problemas de fertilidad[118].

Otra de las razones más comunes del uso de los análisis de DGP-A son los abortos de repetición, definidos generalmente (algunos países varían esta consideración) como la ocurrencia de 3 o más abortos espontáneos con al menos 14 semanas de gestación de embriones resultado de concepciones naturales[119]. Se ha observado que entre el 28 y el 78% de los embriones analizados en parejas con abortos de repetición sufrían aneuploidías[120]. Esto puede deberse, en muchos casos, a que los pacientes presentan anomalías estructurales balanceadas que desconocen y tan solo se ponen de manifiesto en la siguiente generación[121]. No hay que confundir los abortos de repetición con los fallos de implantación repetidos, definidos como 3 o más ciclos de IVF fallidos tras la transferencia de 10 o más

embriones de calidad[122], aunque debemos señalar que la definición de este suceso tampoco ha sido consensuada por todos los países.

Por último, y como es lógico, la ocurrencia de fenómenos de aneuploidía en la descendencia previa es motivo más que suficiente para considerar el uso de un análisis de DGP-A para prevenir futuras ocurrencias[121].

Así, en la bibliografía científica podemos encontrar cientos de artículos que corroboran los beneficios de emplear estas técnicas de selección sobre los embriones con potencial de ser posteriormente transferidos[123–128]. Sin embargo, en la bibliografía también podemos encontrar detractores de esta técnica[129,130], que afirman que las conclusiones obtenidas en estudios previos no muestran relación entre la mejoras de las tasas y el uso del DGP-A. Estos autores contrarios argumentan entre otras, que el elevado coste de las pruebas no justifica su utilización, que la biopsia puede dañar al embrión, o que los resultados del DGP-A son inexactos debido al mosaicismo y a errores técnicos. De esta forma, la utilización del DGP-A contribuiría a un incremento en costes del tratamiento, y al descarte de embriones potecialmente viables, lo que finalmente se traduciría en una reducción de las tasas de éxito de un tratamiento de IVF si incluye el cribado de aneuploidías. Sin embargo, mucha de estas afirmaciones quedan ensombrecidadas si se aplica un tratamiento riguroso de los datos, así como una estratificación adecuada de los pacientes, una aplicación adecuada de las técnicas (principalmente de biopsia), y si se tiene en consideración no sólo el número parejas que finalizan un ciclo de IVF con un niño en casa sano, sino también el tiempo que han tardado en lograrlo, el número de transferencias necesarias y el número de abortos sufridos[131–133].

Como acabamos de comentar, los protocolos más comunes para realizar un análisis DGP-A por NGS suelen comenzar con una amplificación completa del genoma o WGA[134]. En muchos casos esta amplificación se realiza mediante PCR con oligos aleatorios y su objetivo es la generación de suficiente cantidad de material genético para llevar a cabo el análisis DGP-A posterior u otras técnicas de diagnóstico preimplantacional[82]. Posteriormente, el material amplificado es secuenciado a muy baja cobertura (0,01X aproximadamente según resultados experimentales no publicados)[135]. La cobertura se define como el número de veces que cada posición está representada en las lecturas producidas[136]. La secuenciación a tan baja cobertura supone un hecho diferencial con respecto a otros análisis genéticos basados en secuenciación masiva donde los rangos de cobertura oscilan entre los 30 y los 1000 X dependiendo de la técnica y objetivo a lograr[136]. Una vez realizada la secuenciación, las lecturas secuenciadas se alinean al genoma de referencia para construir un archivo BAM.

El archivo BAM es una versión comprimida con BGZF del archivo SAM, que da un buen nivel de compresión a la vez que provee un eficiente acceso aleatorizado al fichero para búsquedas indexadas. El archivo SAM es un archivo que contiene información sobre las lecturas y su alineamiento. Evidentemente, esta baja cobertura conlleva que las lecturas no son solapantes, a diferencia de lo que ocurre generalmente en secuenciación masiva pero, como se verá más adelante, para este tipo de análisis no necesitamos esa redundancia. Posteriormente, los archivos BAM serán analizados mediante algoritmos de detección de cambios de número de copia para diagnosticar la ploidía del embrión original y si es o no apto para ser transferido.

Un paso clave en el análisis bioinformático de estos resultados es el filtrado para eliminar lecturas aberrantes que puedan distorsionar el resultado emitido. Una de ellas es lo que tradicionalmente se define como duplicado de PCR, que son aquellas lecturas que presentan la misma posición inicial y final cuando son alineadas con el genoma referencia[137], y que han sido originadas a partir de la misma molécula original. De esta forma, la generación de un duplicado de PCR provoca que ciertos nucleótidos aparezcan sobrerepresentados en el archivo BAM final. Esta diferencia entre el material original y el producto de la librería contribuye a la generación de una dispersión o ruido que provoca que el modelo mostrado en el análisis DGP-A se desvíe del modelo real de ploidía de la muestra.

Como hemos comentado, se suele considerar duplicado de PCR a aquellas lecturas idénticas. Recientemente se ha demostrado que un gran número de duplicados de PCR surgen durante la fase de amplificación de la PCR [138–140], concretamante en el DGP-A esta amplificación ocurre en la fase de preparación del *template* o molde. Este paso consiste en la creación de múltiples copias del mismo amplicón, cercas la unas de las otras, para que durante la secuenciación se produzca una señal lo suficientemente fuerte como para ser detectada. El protocolo seguido para DGP-A durante esta tesis tiene la peculiaridad de que este paso se realiza mediante ampificación isoterma (IA por sus siglas en inglés, *Isothermal Amplification*). La IA es un proceso sencillo que se hace a temperatura constante controlada, lo que reduce considerablemente el tiempo necesario para completar la etapa de amplificación[141]. Esto resulta bastante ventajoso para el análisis DGP-A, pero genera errores[142]. Durante la etapa de IA, el ADN es amplificado en la superficie de una ISP (Ion Sphere Particule), es decir, una molécula de ADN se unirá a esta partícula, y durante la IA esta molécula se copiará por toda la superficie de la ISP. En este proceso, es crítico que una ISP esté recubierta de copias exactas de una única molécula de ADN. Para evitar que varias moléculas se unan a una misma ISP, la IA se realiza en una matriz 3D en la cual tiene lugar

una reacción de emulsión isotérmica[143]. El uso de esta matriz reduce, pero no elimina, el riesgo de contaminación[142], por lo que, a veces, una hebra del ADN original puede difundir en la mezcla y viajar de una ISP a otra.

El uso de estructuras cerradas reduce el riesgo de contaminación[142], pero provoca que, a veces, una hebra del ADN original hibride en dos pocillos, de manera que cuando la polimerasa comienza la elongación, se generan lecturas con distinto tamaño y punto de origen. También pueden originarse secuencias con distinto tamaño y punto de terminación si la reacción finaliza antes de alcanzar el final del fragmento[144].

Por último, debemos tener en cuenta que durante los primeros ciclos de amplificación predomina la formación de los fragmentos que se ceban más fácilmente, mientras que en los ciclos finales lo hacen aquellos que se elongan con mayor eficiencia[144]. Esto también puede generar desviaciones de la dispersión que hagan que el archivo BAM final no represente fielmente la muestra original.

Los duplicados de PCR producidos durante el proceso de amplificación del genoma y/o IA tienen una característica especial con respecto a los duplicados típicos de un experimento de NGS, y es que, al originarse a partir de cebadores aletarios, pueden no tener exactamente el mismo origen y fin, lo que los hace difícil de identificar. Este otro tipo de duplicados no exactamente iguales, contribuye a "deformar" el perfil del embrión aportando un ruido extra que puede enmascarar la verdadera ploidía y desembocar en un diagnóstico erróneo. Por ello, tanto los duplicados "clásicos" como los definidos en segundo lugar deben ser correctamente procesados y eliminados del archivo BAM antes de proceder al análisis de aneuploidías.

En este punto debemos destacar que la llegada de las nuevas tecnologías basadas en NGS para DGP-A pusieron de manifiesto un nuevo fenómeno: el mosaicismo[145]. Como ya comentamos, el mosaicismo se define como la presencia de dos o más genotipos en un mismo organismo que ha sido producido a partir de un zigoto simple[146]. Cuando este fenómeno cursa con anomalías cromosómicas el origen se debe a un error de la división celular en las etapas tempranas del desarrollo embrionario[146]. El porcentaje de aneuploidía presente, la naturaleza de la anomalía y los tejidos afectados por las mismas determinarán el grado de afección y la expresión clínica del portador[90], pero resulta muy complicado establecer el alcance fenotípico a nivel prenatal debido tanto a la complicación obvia para obtener muestra. A nivel preimplantacional, cabe destacar que estas anomalías pueden provocar menores tasas de implantación de los embriones transferidos en un proceso de

fertilización *in vitro*[147,148]. Además, a nivel postnatal el mosaicismo ha sido asociado con fenómenos de retraso cognitivo y desarrollo sexual anormal entre otras complicaciones[116,149,150] y puede afectar a cualquier tipo celular. Por todo esto, resulta esencial poder seleccionar y transferir embriones controlando este problema. Sin embargo, también existen diversos estudios que reportan nacimientos de niños sanos a partir de la transferencia de embriones mosaico[151], lo que podría apoyar la teoría de la existencia de un proceso conocido como rescate embrionario[152], por el cual los embriones con bajos porcentajes de aneuploidía (bajos porcentajes de mosaicismo) relegan esas células a diferenciarse en la formación de tejidos "menos importantes"[153] o sufren apoptosis[154], de manera que dichos embriones podrían ser hábiles para ser transferidos y resultar en el nacimiento de un bebé sano[153–155]. Debido a todo esto, en la actualidad existe un gran debate sobre qué hacer con los embriones mosaico: transferirlos y correr el riesgo de que se desarrolle un feto que puede acarrear problemas, o descartarlos con el peligro de estar descartando embriones con un claro potencial de implantación y de dar lugar a un niño sano. Así, recientemente la *Preimplantation Genetic Diagnosis International Society* (PGDIS)[156] publicó una serie de recomendaciones sobre como tratar estos embriones[157] . A grandes rasgos, este sistema trata de establecer una serie de riesgos para la transferencia de embriones mosaico, dependiendo del cromosoma afectado y el nivel de mosaicismo detectado.

Por otro lado podemos encontrar en la bibliografía numerosos estudios que contradicen estas tasas de éxito en la transferencia de embriones mosaico de bajo porcentaje, alegando que dichas tasas son debidas a falsos positivos de las técnicas de detección (embriones euploides diagnosticados erróneamente como mosaico)[158], apoyándose además en casos donde la transferencia de embriones diagnosticados como euploides dio como resultado descendencia mosaico (falsos negativos)[4].

Este hecho provoca que la identificación y determinación de los embriones mosaico, priorizando la transferencia de aquellos que sean 100% euploides, sea algo necesario. Así, el desarrollo de un método de determinación del porcentaje de aneuploidía, que controle los niveles de especificidad y sensibilidad, es una tarea esencial que podría permitir establecer con seguridad (si existe) el porcentaje umbral de mosaicismo para transferir un embrión con seguridad.

1.2.4 DGP-M

El DGP también puede ser aplicado para la determinación de embriones libres de enfermedades monogénicas mediante la aplicación del diagnóstico molecular; esto es lo que hoy en día se conoce como DGP-M[159].

Las técnicas de realización del DGP-M son muy variadas y han ido evolucionando a lo largo del tiempo, pero, al igual que sucedía en el DGP-A, suelen requerir una amplificación previa del ADN debido a que se trabaja con muy poco material de partida, normalmente el ADN de enre una y cuatro células dependiendo del tipo de biopsia realizada[95]. A diferencia del DGP-A, el método tradicional de amplificación por excelencia debido al gran tamaño de los fragmentos producidos y a la baja tasa de error de los mismos es el MDA (del inglés, *Multiple Displacemente Amplification*), que amplifica el material genético por medio del empleo de la polimerasa Phi29 empleando una temperatura constante[160].

Un problema recurrente en el DGP-M es el fenómmino llamado ADO (del inglés, *Allele Drop-Out*) o amplificación preferencial de un alelo frente al otro, de manera que un locus heterocigoto aparece como homocigoto[161,162] tras la secuenciación. Este fenómeno afecta por igual a todos los polimorfismos a lo largo del genoma con independencia de su frecuencia alélica, provocando fallos de diagnóstico debidos a la incapacidad de detectar los alelos realmente presentes[163]; tan solo será descartable el suceso de ADO cuando aparezcan polimorfismos en heterocigosis[164], pues es el único caso en que se puede asegurar que ambos alelos han sido detectados. Así, este fenómeno supone uno de los mayores retos a los que debe enfrentarse el DGP, pues si al realizar el análisis en un embrión este aparece como homocigoto y no se detecta la alteración en estudio cabe la posibilidad de que el alelo mutado no haya sido amplificado, constituyendo un falso negativo que puede desencadenar la transferencia de un embrión afecto[83].

La frecuencia de los fenómenos de ADO en la literatura ha sido ampliamente reportada, con rangos que van hasta casi el 25% de los loci analizados en casos clínicos de DGP[165]. De hecho, el ADO ha sido reportado como la primera causa de errores en análisis DGP-M de fibrosis quística, ya que los embriones enfermos por heterocigosis compuesta son diagnosticados incorrectamente y transferidos debido a que tan solo uno de los alelos es detectado mientras el otro sufre un fenómeno ADO[166,167]. La forma más común de evitar esto consiste en realizar un análisis indirecto a través del estudio de una serie de marcadores polimórficos presentes en el embrión que estarán asociados al alelo sano o al mutado. En este tipo de estudios indirectos se evalúa la relación existente entre varios marcadores

neutros no causales de enfermedad y los haplotipos causantes. La estrategia que comúnmente se sigue es el análisis de ligamiento.

En los análisis de ligamiento se realiza una búsqueda en el genoma de polimorfismos cercanos al gen a estudiar, y se trata de determinar la segregación de cada polimorfismos con el alelo mutado o el sano. Para ello, generalmente se requiere la intervención no sólo de la pareja, sino también de familiares por línea ascendente o descendente, como se verá más adelante.

El principal inconveniente reside en la necesidad de disponer de varias generaciones de familias afectadas de manera que cada miembro es genotipado para determinar los polimorfismos que son heredados más frecuentemente con la enfermedad[168]. La ventaja es que este tipo de metodologías permiten la identificación de nuevos genes involucrados en la patogénesis ya que no se limita al estudio de los polimorfismos causales conocidos. Encontramos múltiples estudios que prueban la efectividad de estos métodos en la identificación de genes causales de patogenicidad[169–172] .

En los estudios de asociación genética por el contrario se buscan polimorfismos individuales implicados en la enfermedad a través de la identificación de genes candidatos y el análisis de la frecuencia de aparición de los polimorfismos en individuos afectos por dicha enfermedad con respecto a una población control sana. El inconveniente de este tipo de técnicas reside en la cantidad de falsos positivos que se generan si un polimorfismo se encuentra sobrerrepresentado.

Capítulo 2: Estado del arte

2.1 DGP-A

La primera técnica desarrollada para DGP-A fue la hibridación de fluorescencia *in situ* (FISH por sus siglas en inglés, *Fluorescent in situ hybridization)*, que emplea sondas fluorescentes de nucleótidos complementarias al ADN para visualizar las regiones de interés. Se aplicó por primera vez en 1982 para detectar la presencia o ausencia de determinadas secuencias de ADN en los cromosomas de *Drosophila melanogaster* [173]. A pesar de que la aplicación de la técnica supuso un hito en la mejora de las tasas de éxito[174], presenta grandes limitaciones que la han desplazado frente a técnicas más recientes. Entre esas limitaciones destaca el hecho de que no se analizan todos los cromosomas[174], por lo que realmente no tenemos toda la información para diagnosticar correctamente al embrión. Además, generalmente requiere de varias rondas de hibridación/deshibridación de sondas lo que, además de ser un proceso muy tedioso, puede provocar falsos positivos si las sondas no se deshibridan correctamente[175].

El primer gran avance surgió de la mano de la aplicación de la técnica de microarrays al DGP-A, con la ventaja de proporcionar resultados para todos los cromosomas[176]. Una de las plataformas más empleadas es el array de hibridación genómica comparativa (aCGH por sus siglas en inglés, *Comparative Genomic Hybridization Aray*)[177]. Aunque esta técnica ha mostrado magníficos resultados durante muchos años, tiene ciertas limitaciones. El principal inconveniente de la técnica consiste en su elevado coste, una sensibilidad media (lo que impide o dificulta la detección de mosaicismos), y un protocolo muy manual. Además, en el caso de progenitores portadores de una traslocación balanceada, no es posible distinguir entre embriones normales y portadores de traslocación equilibrada. Además no detecta mosaicismos con especificidad.

A pesar de que en 2012 Simpson et al. propusieron el CGH-array como la técnica más adecuada para realizar un análisis de DGP-A[82], recientemente varios estudios han validado el éxito de la aplicación de técnicas de secuenciación masiva, en inglés *Next-generation sequencing* (NGS), tanto en biopsias de célula única, como de trofoectodermo[178,179]. La llegada de esta nueva metodología ha supuesto un gran avance en el análisis de aneuploidías con muestras procedentes de embriones debido no solo a la disminución del costo del análisis, sino al potencial de la técnica y a que permite el análisis del ADN mitocondrial[180] y la detección de mosacismos en biopsias de trofoectodermo[181], algo que las técnicas anteriores no eran capaces de detectar. Normalmente los protocolos

de NGS suelen compartir la primera etapa con el aCGH, comenzando con una amplificación del genoma completo, o WGA (Whole Genome Amplification por sus siglas en inglés). Sin embargo, quizá la principal ventaja de estas técnicas NGS actuales es que permiten el análisis en paralelo de varias muestras dentro de la misma carrera, pues cada una es marcada con una secuencia única llamada código de barras o, en inglés, *barcode,* compuesta por una secuencia única de nucleótidos que permiten identificar la procedencia de los productos de PCR[3]. Nos encontramos por tanto ante una técnica con alto nivel de precisión, menos costosa y más rápida que las técnicas previas[3,182–185]. Por estas razones la presente tesis se centra en el desarrollo de un método de análisis eficaz basado en ésta tecnología.

2.1.1 Procesado de la muestras

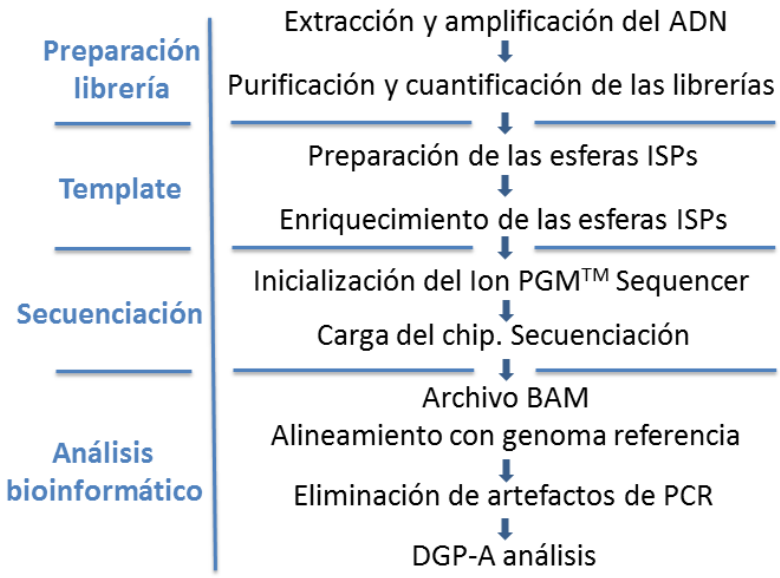

Figura 9: Esquema del proceso de análisis de un DGP-A.

El procesado de las muestras para el análisis DGP-A (Figura 9) se realiza normalmente utilizando el kit Ion ReproSeq™ PGS Hi-Q View (ThermoFisher Scientific). Este kit contiene todo lo necesario desde la amplificación del material genético de la biopsia tomada a partir del embrión hasta las herramientas bioinformáticas para la obtención de resultados. Este kit realiza la amplificación de todo el genoma mediante un método basado en PCR, utilizando oligos aleatorios de 6-mer. La parte de amplificación y preparación de la librería se realiza con el kit SingleSeq. Este kit es una modificación de PicoPlex (Rubicon Genomics) que probablemente constituye el kit más empleado en la amplificación de

muestras para DGP-A. El protocolo para el paso de preparación de la librería está dividido en 3 pasos claramente diferenciados que permiten purificar y cuantificar las librerías:

1. Lisis celular y fragmentación del ADN. En este paso se lisan las células biopsiadas y se fragmenta el ADN mediante la adición de un tampón de lisis y un choque térmico.

2. Preamplificación. Este paso consiste en la amplificación del ADN de las células biopsiadas gracias a la utilización de oligonucleótidos aleatorios. Estos oligos llevan, además, unido un adaptador de secuencia palindrómica que hace que los fragmentos amplificados formen una horquilla, de forma que dejan de estar disponibles como molde para la siguiente ronda de amplificación, lo que asegura que en cada nueva ronda el ADN copiado se forme a partir de moléculas de ADN genómico y no a partir de moléculas ya amplificadas en un proceso de amplificación lineal del genoma.

3. Amplificación y etiquetado. Este paso consiste en la amplificación y linearización de las horquillas formadas en el paso anterior. En esta amplificación, al contrario que la anterior, el ADN sí que se replica de forma exponencial, logrando generar una concentración suficiente como para poder ser secuenciada. Los adaptadores empleados son complementarios a los del paso anterior y llevan, además, unidos los códigos de barras.

Una vez preparadas las librerías se inicia la fase de *Template* (Figura 9), en la cual las librerías se combinan y purifican utilizando un sistema de bolas magnéticas, para posteriormente cuantificarlas con un método fluorimétrico. El conjunto de librerías se diluye a una concentración de 100pM. Tras esto se procede a la preparación del molde de secuenciación a través de la unión de las moléculas de ADN de la librería a un soporte sólido en el cual tiene lugar la amplificación clonal y la posterior secuenciación. En el caso de la plataforma Ion Torrent, este soporte está compuesto por unas microesferas llamadas ISP (por sus siglas en inglés, *Ion Sphere Particle*).

Tradicionalmente, Ion Torrent utiliza la técnica de la PCR en emulsión para realizar la amplificación clonal, que se basa en la dilución y compartimentalización de las moléculas de ADN en gotas de agua en una emulsión con aceite. Idealmente, cada una de las gotas contiene una única molécula de ADN y funciona como un microreactor de PCR. Aunque se trata de un método muy eficaz para la preparación del *template*, también es bastante lento

puesto que puede tardar más de 8 horas. Como se ha mencionado anteriormente, el tiempo es un factor crítico en el DGP-A, por lo que el el kit Ion Reproseq utiliza, en lugar de una PCR en emulsión, un sistema alternativo conocido como *Isothermal Amplification*. Este sistema emplea un reactivo viscoso capaz de generar una malla tridimensional en la que se consigue una compartimentalización similar a la anterior. Además, la reacción tiene lugar a temperatura constante de 37ºC durante 30 minutos, lo que constituye un proceso mucho más rápido que el método tradicional de PCR en emulsión. Sin embargo, su utilización supone un aumento considerable del número de esferas policlonales (donde más de una hebra se amplifica clonalmente), debido a que el movimiento de las hebras no está totalmente restringido, pudiendo por lo tanto fluir de una esfera a otra. La última fase de la preparación del molde de secuenciación consiste en el enriquecimiento de las esferas que contienen ADN, descartando las esferas vacías, en un equipo automatizado que utiliza un sistema de biotina-streptadina.

Finalmente, se procede a la secuenciación en la plataforma IonTorrent™ Personal Genome Machine™ PGM (Life Technologies, Thermo Fisher Scientific Inc., MA USA), para lo cual se sigue un protocolo de secuenciación estándar, pero utilizando 250 flujos en lugar de los 500 habituales debido a que, para esta aplicación, no son necesarias lecturas especialmente largas. Esto se debe a que en el DGP-A no interesa detectar mutaciones, sino saber a qué cromosoma pertenece cada lectura para así calcular el número de lecturas que tiene cada fragmento de cromosoma, por lo que con lecturas de apenas 100pb secuenciadas es más que suficiente. Para la secuenciación se emplean chips 318, con lo que es posible procesar hasta 24 muestras.

2.1.2 Filtrado de duplicados de PCR

Como se ha mencionado anteriormente, el filtrado de duplicados de PCR es un proceso esencial en prácticamente cualquier análisis bioinformático de datos de NGS. Existe una amplia gama de programas destinados a marcar y/o eliminar duplicados y/o artefactos de PCR. Los más populares analizan la cobertura para detectar cambios en el número de copias[144,186]. El método desarrollado por *Bansal* calcula el ratio de duplicados en función de la cobertura teniendo en cuenta si se trata de una región genómica o una no codificante[179]. Por su parte, SAMBLASTER combina la información de la cobertura con la secuencia nucleotídica para ordenar las lecturas y marcar las secuencias duplicadas[140]. Sin embargo, como ya hemos mencionado, la cobertura presenta alta variabilidad que puede generar una

distribución no uniforme que enmascara la verdadera naturaleza de la muestra de origen, de forma que el diagnostico emitido puede no ser fiel a la realidad del material de origen.

Otros algoritmos emplean la posición inicial y final del alineamiento, así como la calidad entre otros parámetros accesorios[187–191]. Un ejemplo es FastUnique, que compara las secuencias pareadas para identificar duplicados en función de las posiciones[181], las herramientas incluidas en SAMtools[182,183] que emplean las posiciones para eliminar los duplicados idénticos. Picard emplea la posición 5' para identificar aquellas lecturas idénticas[184]. Por último, el flujo de trabajo desarrollado por ThermoFisher en su filtro de duplicados tiene un funcionamiento similar a Picard. Sin embargo, ninguno de estos algoritmos ha sido específicamente desarrollado para DGP-A por lo que el proceso de filtrado no es el óptimo para este tipo de muestras. Esto es debido a la naturaleza de la amplicación que se lleva a cabo, que no representa todo el genoma de manera uniforme. Por tanto, resulta necesario desarrollar métodos que estandaricen el ratio de lecturas para lograr un diagnóstico correcto de la ploidía de la muestra[192].

En el momento de la escritura de esta tesis, Ion Reporter presenta dos versiones principales para el análisis DGP-A, la versión 5.0 y la 5.6. Una de las mejoras más importantes implementadas en la nueva versión radica precisamente en lo que a filtrado de duplicados se refiere. Mientras que en la versión 5.0 tan solo se eliminaban secuencias que alineaban en múltiples posiciones genómicas o que no lo hacían en ninguna, a partir de la versión 5.2 y subsiguientes se ha implementado, a través del flujo de trabajo *FilterDuplicates*, un enfoque en dos pasos que agrupa todas las lecturas que tienen la misma coordenada de inicio en el extremo 5' con respecto al genoma referencia y que, además, tienen la misma orientación. Una vez agrupadas marca como duplicados aquellas que poseen la misma coordenada de terminación 3' excepto una de ellas, que considera como lectura original, pero no las elimina. En un segundo paso, las secuencias que hayan sido marcadas son eliminadas del archivo BAM[193], que será empleado en la detección de aneuploidías.

Este método resulta muy efectivo para la eliminación de duplicados generados durante el proceso de amplificación de la librería. Sin embargo, no tiene en cuenta otras fuentes de duplicados debidos a la amplificación del ADN en PGT-A.

Por tanto, podemos definir una serie de debilidades claves a la hora de aplicar al DGP-A las técnicas de filtrado actualmente disponibles en el estado del arte:

- La distorsión generada por la dispersión de las lecturas dificulta la **emisión de resultados** en la determinación de la ploidía de los embriones de los que proceden las células biopsiadas.

- Las técnicas de filtrado **no están diseñadas para filtrar** los archivos BAM procedentes de la secuenciación de librerías DGP-A, ya que las secuencias no se disponen en escalera y la cobertura es del 0,01X aproximadamente, mientras que dichos algoritmos han sido diseñados para filtrado de secuenciaciones en escalera con coberturas entre el 30 y el 1000X.

- La **incapacidad de distinguir duplicados de PCR no idénticos**, es decir, aquellos surgidos por la amplificación de fragmentos previamente amplificados y cuya secuenciación finaliza antes de haber reproducido todo el fragmento.

2.1.3 Detección de la ploidía

Para la detección de aneuploidías, una vez alineadas las lecturas, se emplea el software Ion Reporter (IRS) en varias de sus versiones. Este software utiliza la profundidad de las lecturas a lo largo del genoma para predecir los cambios de número de copia haciendo uso de un modelo oculto de Markov HMM (por sus siglas en inglés, *Hidden Markov Model*). Para el cálculo, en primer lugar, se divide el genoma en fragmentos discretos de aproximadamente 2Mb llamados ventanas, para posteriormente calcular la cobertura dentro de cada ventana a través del conteo del número de lecturas que pertenecen a cada una. Esta información es corregida para las desviaciones debidas al contenido GC mediante la comparación con una línea base predefinida formada por 10 perfiles de embriones euplooides de sexo masculino. Finalmente, el algoritmo basado en HMM determina el estado de ploidía más probable de cada ventana y realiza una normalización y posterior

suavizado de los perfiles, de forma que se indica no solo las aneuploidías presentes en la muestra, sino también las deleciones e inserciones con una sensibilidad de hasta 10Mb.

En el momento de la escritura de esta tesis Ion Reporter™ software (IRS) presenta dos variantes del protocolo para el análisis de la detección de aneuploidías: *ReproSeq Low-pass Whole-genome Aneuploidy PGS 5.2* y *ReproSeq PGS mosaic w1.1*. La diferencia principal entre ambos protocolos en que en el primero el algoritmo HMM asigna la ploidía de mayor probabilidad al considerar la probabilidad de que cada valor de cobertura pertenezca a un estado de ploidía que se ajusta a valores discretos (0, 1, 2, 3, etc). Sin embargo, esto provoca que tan solo ploidías completas sean detectadas, obviando posibles mosaicismos en la muestra. Por su parte, el flujo de trabajo *ReproSeq PGS mosaic w1.1* surge de una modificación del anterior para permitir precisamente la detección de porcentajes de aneuploidía. La modificación radica en la asignación de estados intermedios del número de copias. De esta forma, el algoritmo ajusta los datos de cada región con un "paso entre ploidías" de 0,5. Así, considera la existencia de estados intermedios de ploidía entre los estados absolutos, lo que permitiría detectar teóricamente la presencia de mosaicismo[194].

Si bien es cierto que la llegada del NGS para DGP-A[195,196] convirtió la bioinformática en una herramienta útil en pleno desarrollo exponencial, los algoritmos implementados hasta hoy en el estado del arte presentan ciertas limitaciones clave para su aplicación al DGP-A:

- No permiten determinar con exactitud **el nivel de mosacisimo**[197].
- No permiten detectar bajos porcentajes de mosaicismo.
- Estas dos implican que el riesgo de **emitir un resultado erróneo** con mosaicismos de bajo porcentaje es muy elevado.
- La aplicación de estas técnicas no permite realizar un estudio para **determinar la significancia de los embriones mosaicos** en los ciclos de IVF.

Esto es particularmente cierto cuando se realiza DGP-A en las plataformas de ThermoFisher, al trabajar con un menor número de lecturas.

- Tampoco permite la realización de un estudio acerca de la **conveniencia de la transferencia de embriones con distintos porcentajes** de mosaicismo.

2.1.1 Evaluación de la calidad del resultado

En esta sección se analizan varios parámetros que se utilizan para evaluar la fidelidad de los resultados. Estos parámetros permiten evaluar la probabilidad de que una alteración detectada durante el análisis DGP-A sea un falso positivo o, por el contrario, la probabilidad de que una alteración no sea detectada debido al ruido de la muestra.

2.1.1.1 Confianza y precisión

La confianza es definida como la relación logarítmica entre el estado de ploidía detectado por el algoritmo de análisis y el estado esperado. Valores elevados de confianza indicarán que el algoritmo empleado es capaz de detectar con gran fiabilidad que el estado de ploidía detectado difiere del estado esperado.

La precisión es la relación algorítmica existente entre el probable estado de ploidía asignado y el siguiente estado de ploidía más próximo. Ambas se calculan para cada una de las regiones analizadas.

Una precisión baja (inferior a 10) indica que no hay certeza sobre el valor absoluto de ploidía detectado. Así, es posible que para niveles altos del estado de ploidía, el valor de precisión sea bajo mientras que la confianza sea alta. Esto indicaría que el algoritmo no es capaz de determinar con precisión el estado de ploidía pero que existe una alta certeza de que el estado de ploidía sea diferente al esperado.

2.1.1.2 MAPD

MAPD por sus siglas en inglés, *Median Absolute Pairwise Differences,* se define generalmente como la media de los valores absolutos de las diferencias por pares del logaritmo en base 2 ($\log_2$). Es una medida similar a la desviación estándar que da una estimación de la desviación y ruido producidos durante el proceso de amplificación y define si los datos son apropiados o no para en análisis de aneuploidías. A diferencia de las dos anteriores, este valor es calculado para toda la muestra en general. Esta medida fue diseñada originalmente para ser empleada con datos de microarrays[198,199], pero ha ido tomando importancia en el campo del DGP-A por NGS tras ser adaptada[200] por las principales plataformas de análisis[201]. El valor se obtiene comparando las relaciones en $\log_2$ de regiones adyacentes en el genoma según la fórmula siguiente:

$$MAPD = (|x_{i-1} - x_i|)$$

siendo x_i el $\log_2$ de la posición*i,* con i ordenada según la posición cromosómica.

Típicamente, valores altos de MAPD se asocian con una elevada dispersión o ruido en las ventanas, generalmente debida a una baja calidad de amplificación, pudiendo llegar al punto de enmascarar cualquier posible aneuploidía de la muestra.

En 2014 *Cai et al.* estableció en 0,45 el valor umbral de MAPD admisible para un análisis realista de la ploidía al tomar regiones adyacentes de 500kb[202]. Sin embargo se ha comprobado que este valor decrece al aumentar el tamaño de las regiones[199]. Actualmente, de manera general se considera como aceptable todo valor de MAPD por debajo de 0,3.

2.1.1.3 Número mínimo de lecturas

El número mínimo de lecturas necesario para obtener un diagnóstico fiable de la ploidía es un concepto que está muy relacionado con la dispersión de las lecturas en una muestra, es decir, con el MAPD. Una baja cantidad de lecturas normalmente arroja valores altos de MAPD en la muestra analizada, ya que suele presentar mayor dispersión, lo que podría enmascarar la ploidía debido al ruido.

2.2 DGP-M

La primera aplicación de la técnica ocurrió en 1994, momento en que se documenta el uso de la técnica de amplificación por PCR para seleccionar embriones libres de una enfermedad ligada al cromosoma X a través de la selección de embriones femeninos[203]. Para ello, se seleccionarion aquellos embriones que mostraban un resultado negativo para la amplificación del cromosoma Y eran considerados como femeninos y, por tanto, potencialmente transferibles. Sin embargo resulta evidente apreciar que cualquier test basado en la obtención de un resultado negativo presenta un alto riesgo de falsos negativos debidos a posibles fallos de amplificación, sin los controles adecuados. De hecho, esto desencadenó la transferencia de un embrión masculino enfermo debido a un fallo en el diagnóstico, que finalmente obligó a realizar una terminación voluntaria del embarazo[78].

Inicialmente, se utilizó a técnica de FISH para DGP-M a través del uso de fluorocromos, marcando diferencialmente los cromosomas en la muestra[204]. Sin embargo, debido a sus limitaciones, esta técnica fue rápidamente sustituida por enfoques basados en el análisis de la mutación de interés. Desde entonces, la técnica del DGP ha sido aplicada en múltiples casos como una alternativa al diagnóstico prenatal[205–208], aunque siempre considerando los aspectos éticos que conlleva la selección[99].

Una de las técnicas más utilizadas para DGP-M es la mini-secuenciación. Se emplea para la detección directa de una mutación específica en el estudio de enfermedades monogénicas en embriones[209,210]. Las principales limitaciones se basan en que requiere el diseño de oligos específicos en una región muy concreta y tan solo es posible su empleo en la determinación de mutaciones puntuales. Además, para su realización se requiere una amplificación previa por MDA, lo cual supone la realización de una técnica extra, aumentando el tiempo necesario hasta la obtención de resultados.

Sin embargo, el estudio de marcadores STRs está considerado como la técnica de referencia para la realización de análisis DGP-M[211] gracias a la aplicación de PCR múltiple. Esta técnica consiste en la amplificación de determinados STRs alrededor de la región de interés, junto con la realización de un estudio de ligamiento que nos permita identificar el o los alelos portadores de la mutación. La principal ventaja de esta técnica reside en que puede ser empleada con diferentes parejas con independencia de la mutación que porten. El genotipado y discriminación de los alelos portados puede realizarse mediante diversas técnicas como la amplificación refractaria[212], la combinación de RT-PCR con técnicas de mini-secuenciación a partir de dideoxinucleótidos hibridados con fluorocromos[209] o la

digestión por enzimas de restricción[206]. En esta última se diseñan oligos específicos que amplifican marcadores, generalmente STRs[213,214], que son posteriormente digeridos por enzimas de restricción. Las secuencias diana tendrán sitios de corte específicos que variarán en función del alelo presentado, ya que las variaciones alteran dichos loci, dando como resultado un patrón de migración diferente que puede ser visualizado mediante el corrimiento de los fragmentos resultantes en geles de agarosa. Esta técnica ha sido sustiuda por el análisis de fragmentos mediante electroforesis capilar y la utilización de oligos unidos a fluorocromos específicos para la amplificaión de los STRs. Así, el avance de las técnicas unido al empleo de fluorocromos ha permitido que en los últimos años se pase de la amplificación simultánea de dos o tres STRs a más de quince marcadores por carrera. Sin embargo, en ocasiones, aún supone un número muy bajo de marcadores para hacer frente a los problemas que este tipo de análisis: (1) los STRs aparecen en baja densidad a lo largo del genoma, por lo que a veces los loci disponibles aparecen alejados de la posición de interés; (2) suelen presentar una muy elevada tasa de ADO al emplear material procedente de biopsias de embrión[215]; y (3) se deben seleccionar y validar STRs de manera indivudal no sólo para cada enfermedad, sino prácticamente para cada pareja. Otro inconveniente es que dependen de la disponibilidad de familiares afectos a partir de los que establecer el riesgo de herencia. Todo esto incrementa el coste final de la aplicación de la técnica en DGP-M, a la par que el tiempo necesario para la obtención de los resultados, lo que supone un ingente costo tanto emocional como económico para la pareja en estudio. Por este motivo, el desarrollo de nuevas técnicas aplicadas a DGP-M es un campo de gran interés y en creciente estudio.

En los últimos años han ido surgiendo estudios que demuestran que las técnicas de amplificación del genoma completo WGA (del inglés *"Whole Genome Amplification"*) son perfectamente válidas en el análisis DGP-M[216]. El enfoque pasa por la amplificación del genoma completo para producir suficiente ADN que será posteriormente analizado mediante cualquiera de las técnicas mencionadas anteriormente. La determinación del haplotipo puede ser llevada a cabo mediante PCR a la vez que se combina la realización de análisis directos de la alteración en estudio[216]. Además, las técnicas de WGA permiten combinar el análisis DGP-M con técnicas de aCGH para el análisis DGP-A[205]. Sin embargo, nuevamente el problema de esta metodología se debe al ingente ratio de ADO que experimenta[217], algo que puede ser solventado mediante el análisis de un mayor número de loci.

Karyomapping[218] es una técnica reciente que ha cobrado gran interés en el análisis DGP-M por su gran eficacia y al hecho de que simplifica mucho la aplicación de la técnica al

no necesitar de un diseño específico para cada caso. Emplea arrays diseñados para la evaluación de una alta densidad de SNPs para determinar, el haplotipo ligado a la mutación causante de enfermedad[218]. La principal ventaja de esta metodología reside en su amplia aplicación, ya que no es necesario diseñar un análisis personalizado al caso de los pacientes. Sin embargo, presenta el inconveniente de requerir necesariamente un familiar por vía ascendente o descendente de estatus conocido para establecer las fases haplotípicas de la pareja, por lo que no es una técnica útil cuando no se pueda disponer de dicho familiar (como aquellos casos donde la enfermedad sea causada por una alteración *de novo)*. Por otro lado, es una técnica difícilmente aplicable a casos de alta consanguinidad. Otra ventaja adicional es que detecta aneuploidías a través de las frecuencias alélicas de los marcadores analizados, permitiendo realizar de manera simultánea PGT-A y PGT-M. Sin embargo, no es capaz de evidenciar errores mitóticos y con ello fenómenos de mosaicismo. Además, no es capaz de realizar la detección directa, por lo que precisa del trío para determinar la segregación alélica. Por último, el protocolo es largo y complicado, por lo que no puede ser aplicado para ciclos con transferencia en fresco, donde se requieren resultados en menos de 24 horas.

Por último, las técnicas de secuenciación masiva NGS (del inglés *Next Generation Sequencing*) permiten analizar varios loci al mismo tiempo que se realiza un análisis DGP-A a partir de la misma biopsia. La bibliografía recoge algunos estudios que demuestran la utilidad del NGS para DGP-M[219,220]. El principal inconveniente de la aplicación de estas técnicas al análisis DGP-M se debe a un elevado ADO, lo que podría generar diagnósticos erróneos. La estrategia consiste en el análisis de SNPs por NGS que permita identificar los alelos relacionados con la enfermedad. Además, si la mutación a detectar es del tipo puntual, el estudio directo el SNP responsable de la aparición de la enfermedad es genotipado directamente. Sin embargo, al igual que en el caso anterior, se corre el riesgo de sufrir un fenómeno de ADO, por eso se emplean estas estrategias alternativas.

La ventaja del estudio de SNPs frente a los STRs consiste en su gran densidad y su reparto homogéneo a lo largo del genoma. Esto hace que sea posible diseñar una estrategia de análisis de polimorfismos cercanos a la región de interés. Otra ventaja respecto a las técnicas anteriores es que es capaz de incluir el análisis directo de la alteración a la vez que realiza el análisis indirecto, mientras que las técnicas anteriores requieren la realización de pruebas accesorias. Por estas razones la presente tesis se centra en el desarrollo de un método de análisis eficaz basado en esta tecnología.

2.2.1 Estrategia del DGP-M por NGS

Una limitación técnica que presenta el DGP-M mediante el análisis de SNPs por NGS, es que podría requerir realizar muchas PCRs. Así, si quisiéramos analizar 100 polimorfismos, necesitaríamos 100 PCRs, lo que supone no sólo un coste importante, sino la utilización de una gran cantidad de material genético, de una muestra bastante limitante. Este problema técnico se puede solventar con la utilización de PCRs multiplex, es decir, PCRs que amplifican de manera simultánea un número determinado de polimorfismos. ThermoFisher proporciona una herramienta optimizada para diseñar el mejor conjunto de amplicones que permita secuenciar los marcadores de interés. Para ello, tiene en cuenta la composición nucleotídica de cada amplicón y computa la información para minimizar el número de amplicones necesario para secuenciar el mayor número de marcadores. También asegura que dichos amplicones no se interfieran durante la amplificación, generando conjuntos no solapantes. Por defecto, el tamaño recomendado para cada amplicón es de 200pb. Esta herramienta permite generar un sistema de PCR *multiplex* que amplifique, en una única reacción, varias regiones del genoma, y se llama Ion AmplSeq. A esta plataforma se le sube las coordenadas de los polimorfismos de interés, y la plataforma te devuelve el diseño óptimo. El output de la plataforma de Ion AmpliSeq consiste en dos archivos BED. El primero indica la posición y longitud del amplicon diseñado, mientras que el segundo es de tipo *HotSpot* con la información relativa a la posición concreta del polimorfismos de interés. Este archivo *HotSpot* permite restringir todas las posiciones secuenciadas en el amplicón al análisis de las posiciones de interés, aunque todo el amplicón sea amplificado.

Una de las tareas más complejas de esta aproximación, es la selección de los SNPs a secuenciar. Las bases de datos describen más de 10 millones de SNPs, pero se cree que existen más de 50 millones de SNPs comunes (con una frecuencia mínima para el alelo menor MAF del 1%)[74]. De todos ellos, se deben seleccionar aquellos que nos puedan dar información sobre el alelo mutado. En teoría bastaría con la obtención de dos o tres marcadores ligados a un locus causante de patología, reduciendo el riesgo de no detección de un fenómeno de ADO aproximadamente un 50 y un 75% respectivamente, mientras que con cuatro marcadores el riesgo teórico desaparece completamente[221]. Así, el análisis de los SNPs de una región en lugar de los STRs cercanos parece una estrategia mucho eficiente ya que se dispone de muchos más marcadores y mejor localizados. Sin embargo, se trataría de una técnica muy costosa si tratásemos de analizar todos los polimorfismos de dicha región. Afortunadamente, el número de polimorfismos a estudiar puede ser reducido empleando estrategias de diseño basadas en el desequilibrio de ligamiento (LD) entre alelos[222].

A modo de ejemplo, suponiendo 4 SNPs bialélicos sin recombinación entre ellos, siendo 1 la aparición del alelo de mayor frecuencia y 0 la aparición del alelo de menor frecuencia, podemos imaginar la situación en que solo sean posibles 3 haplotipos distintos del tipo [SNP1 SNP2 SNP3 SNP4] en una población cualquiera [1 1 0 0], [0 0 0 1] y [1 1 1 0]. En este caso vemos que el alelo mayor del SNP 4 tan solo aparece presente cuando el alelo mayor del SNP 2 está ausente y que el SNP 2 es redundante con el SNP 1. Esto implica que un estudio que genotipase tan solo el SNP 1, el SNP 2 o el SNP 4 podría obtener suficiente información para describir la región. Estos SNPs se denominan tagSNPs y pueden ser empleados para inferir el estado alélico del resto de polimorfismos de la región, llamada haplobloque, por ser equivalentes desde el punto de vista del LD. Es decir, un tagSNP representan a los polimorfismos de una región. Así, el diseño de paneles de tagSNPs que permitan inferir el estado alélico de los polimorfismos no genotipados permite minimizar el número de loci que deben ser analizados, maximizando la información contenida por el panel y reduciendo el tiempo y costo de la técnica.

En consecuencia, cómo seleccionar un conjunto mínimo de tagSNPs que contenga la máxima información sobre la región de estudio es un tema de gran interés que puede ser aplicado no solo en el mundo del DGP, sino en el diagnóstico genético en general. La bibliografía recoge un extenso número de estudios algorítmicos destinados a solventar este problema[222–227]. Como ejemplo encontramos el método desarrollado por *Carlson et al*., que calcula los tagSNPs en función del estadístico r^2 de correlación[223]. En el método desarrollado por *Bafna et al*. la selección se realiza con base en una nueva medida que cuantifica la confianza con la que un SNP puede ser considerado como tagSNP[224]. El método de *Chen et al.* realiza la selección por medio de la partición de los polimorfismos en bloques[226]. Finalmente, FasTager emplea el LD en una computación paralelizada para escoger los marcadores más adecuados[227].

Concretamente, en el marco de esta Tesis, el enfoque del DGP pretende, a partir de todos los polimorfismos disponibles, seleccionar tan solo aquellos que permitan reconstruir la región de interés (es decir, que sean tagSNP), de manera que nos permita seleccionar embriones libres de la alteración patogénica portada por los progenitores. Esta es una aplicación novedosa para los tagSNPs. Al principio, los estudios de asociación recibieron una gran atención con el objetivo de poder encontrar polimorfismos correlacionados con las enfermedades. Sin embargo, la relación existente entre los tagSNPs y la probabilidad patogénica quedó descrita tan solo a medias[228–230]. Tras esto, las hipótesis libres de asociación comenzaron a ganar fuerza ya que permiten su aplicación en un rango mayor de pacientes, pues mientras que la identificación de tagSNPs asociados a alteraciones

concretas resulta costoso y complicado de reutilizar, la identificación de tagSNPs asociados a regiones concretas y su uso para reconocer el alelo progenitor causal y su reconstruir su patrón de herencia se convertía en una estrategia mucho más conveniente, al poder ser utilizada en cualquier pareja de pacientes que presentase una alteración para dicha región, con independencia de dicha alteración, de manera que los embriones podrían ser descartados por presentar el haplotipo paterno identificado como cosegregante con la alteración.

Un hecho a destacar es que, mientras los algoritmos diseñados para la selección de tagSNPs se centran tan solo en poder seleccionar aquellos tagSNPs que representan al mayor número de polimorfismos posible, en DGP-M resulta muy interesante obtener también aquellos tagSNPs que no pertenecen a ningún haplobloque, por ser precisamente los que aportarán información diferencial acerca de la línea genética heredada.

2.2.2 Problema de la informatividad

Aunque el uso de marcadores de SNPs conlleva ventajas no solo desde el punto de vista económico, sino que también produce resultados más robustos al reducir la probabilidad de ADO. Sin embargo, al ser más abundantes y menos polimórficos que los STRs, multitud de medidas y parámetros deben ser calculados para determinar si un SNP puede ser o no considerado útil en la predicción otros polimorfismos. En el caso concreto de la selección de SNPs informativos útiles en DGP se deben tener en cuenta más factores que la correlación entre los SNPs de una región.

Empleando tagSNPs en LD con la región de interés, podemos asegurar que los embriones serán portadores o afectos por mostrar aquellos polimorfismos que han sido identificados como cosegregantes con la alteración en los parentales, sin necesidad de realizar el estudio directo de la mutación. El problema de esta estrategia reside en que no todos los haplotipos poblacionales pueden ser reconstruidos a partir de un tagSNP cualquiera por el simple hecho de que represente su haplobloque. Suponiendo dos SNPs bialélicos presentes en un individuo, siendo los alelos del SNP 1 pertenecientes al conjunto {a, A} y los del SNP 2 al conjunto {b, B}; como muestra la Tabla 1, nos encontramos que existen 9 posibles genotipos pero 10 ordenaciones distintas para los pares de alelos que corresponden a ese genotipo, es decir, 10 pares haplotípicos. Por tanto, existe una incertidumbre acerca de la combinación haplotípica presentada para el doble heterocigoto rodeado por un cuadrado rojo en la Tabla 1, a pesar de ser el caso más interesante para la

determinación del haplotipo heredado por cada progenitor. Esta dualidad es lo que hemos denominado *problema de la informatividad* y dificulta el empleo de los paneles de tagSNPs escogidos por los métodos tradicionales en casos de DGP.

Genotipo	Par haplotípico		Genotipo
A B	AB	Ab	A b
	AB	Ab	
a B	aB	ab	a b
	aB	ab	
A Bb	AB	AB	Aa B
	Ab	aB	
Aa b	Ab	aB	a Bb
	Ab	ab	
Aa Bb	AB	Ab	Aa Bb
	Ab	aB	

Tabla 1: Pares haplotípicos y genotipos posibles para un individuo con dos SNPs bialélicos.

Por tanto, a pesar de los métodos disponibles en el estado del arte, se hace necesario el desarrollo de nuevos métodos de selección de un conjunto mínimo y óptimo de tagSNPs útiles en DGP-M que permita solventar el problema de la informatividad.

El uso de SNPs informativos permite establecer los haplotipos heredados por cada embrión y descartar para la transferencia aquellos que muestren los polimorfismos cosegregantes con el cromosoma paterno asociado a la alteración.

Suponiendo un SNP bialélico cuya frecuencia alélica para el alelo referencia o alelo mayor sea p y para el alelo alternativo o alelo menor sea q, encontraremos dos estados:

- Que el SNP sea informativo es decir, que un parental sea homocigoto pp o qq y el otro sea heterocigoto pq.
- Que no sea informativo; resto de combinaciones: $pp - pp$, $pp - qq$, $qq - qq$.

Determinar *a priori* la probabilidad de que un SNP sea informativo no es fácil. De lo anterior podemos apreciar que el MAF (frecuencia alélica del alelo menor) juega un papel decisivo directamente relacionado con la probabilidad de informatividad de un

polimorfismo. Cuanto mayor sea esta frecuencia alélica, mayor número de individuos podrán portar el alelo. Sin embargo, esto no es necesariamente sinónimo de que la informatividad aumente, como muestra la Figura 10.

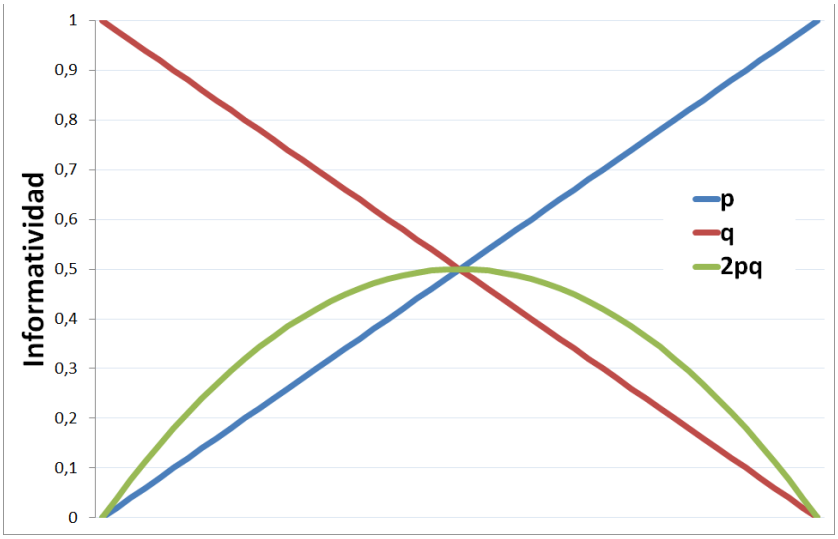

Figura 10: Convergencia de las frecuencias alélicas de un polimorfismo en términos de informatividad. El eje X representaría los distintos experimentos llevados a cabo con las distintas frecuencias de los alelos p y q

Con base en los principios de Hardy-Weinberg y las ecuaciones de la genética de poblaciones en el equilibrio[231] el suceso de informatividad podría expresarse por la ecuación $2pq * (1 - 2pq)$ a partir de la cual se obtiene el desarrollo $p - 3p^2 + 4p^3 - 2p^4$, a la que denominaremos MaxP. Cuando MaxP es derivada e igualada a cero obtenemos una ecuación de tercer grado cuyas soluciones son 0,5. Esto implica que 0,5 es, por tanto, el máximo teórico alcanzable para el valor de informatividad de un polimorfismo, tal como muestra la Figura 10 en el punto en que la línea correspondiente a los heterocigotos converge con la de los homocigotos. Así, cualquier polimorfismo con un valor de frecuencia alélica entre 0,4 y 0,6 es un buen candidato teórico para maximizar la informatividad. La Figura 11 muestra la relación existente entre la frecuencia alélica y la informatividad para una región cualquiera. Como se puede observar, la selección de tagSNPs en base únicamente a su valor de frecuencia alélica podría suponer un criterio suficiente para permitir maximizar la informatividad del panel final escogido. Por su parte, debido a que las poblaciones están en un equilibrio donde las frecuencias de los genotipos varían de una generación a otra, la aplicación de un criterio correspondiente a la selección de heterocigotos al 0,5 es un buen criterio de selección cuando se trate de datos poblacionales, pero puede fallar cuando se pretende aplicar a los individuos concretos.

Así, empleando polimorfismos con mayor probabilidad de ser informativos el problema de la informatividad quedaría solventado. Sin embargo, todavía queda por determnar la procedencia del alelo heredado; por ello se hace necesario el desarrollo de métodos de determinación del alelo portador mediante el fasado (predicción de los alelos que cosegregan con cada cromosoma progenitor) de dichos tagSNPs, permitiendo la selección de los embriones libres de la alteración en estudio.

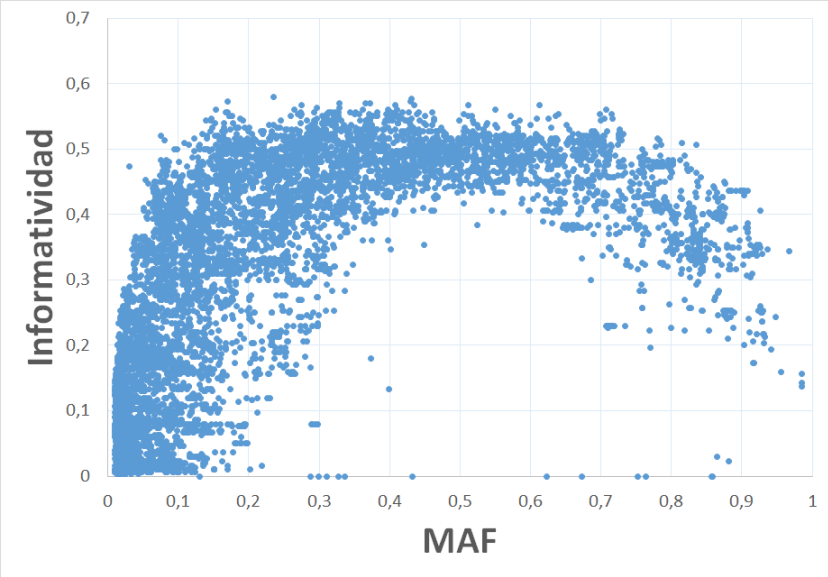

Figura 11: Distribución de los valores de informatividad con respecto al valor de MAF.

2.2.3 Estrategias de selección de tagSNPs

A pesar de la ingente cantidad de estrategias existentes a la hora de seleccionar tagSNPs, todas ellas están basadas en uno de estos tres modelos: modelos basados en el cálculo del desequilibrio de ligamiento, modelos basados en bloques y modelos libres de bloques.

Los modelos basados en el desequilibrio de ligamiento parten de datos genotípicos de poblaciones referencia para tratar de identificar un conjunto mínimo de regiones denominadas ventana cuyos polimorfismos estén en alto desequilibrio de ligamiento, de tal manera que los SNPs de dos ventanas nunca estén en LD entre sí[223].

Por otro lado se ha estimado que más del 50% de los fenómenos de recombinación tienen lugar en menos del 10% del genoma[232]. Este hecho permite a los modelos basados en bloques, definir regiones denominadas bloques haplotípicos donde el desequilibrio de ligamiento existente entre los polimorfismos es compatible con la ausencia de recombinación. Estos bloques además estarán separados por regiones con niveles de recombinación mucho más elevados. Un hecho diferencial con la estrategia anterior reside en que los bloques haplotípicos son previamente definidos por el usuario, de forma que la selección de tagSNPs se centra en encontrar, de entre todos los polimorfismos pertenecientes al bloque haplotípico, un conjunto mínimo de tagSNPs que permita distinguir todos los haplotipos existentes en el bloque[228,233,234]. La desventaja principal de las estrategias basadas en estos métodos reside en la multitud de formas disponibles para dividir el genoma en bloques haplotípicos, lo que resulta en que, incluso la aplicación del mismo algoritmo puede resultar en paneles de tagSNPs muy diferentes para la misma región de análisis. Dos SNPs serán considerados como suficientemente correlacionados cuando ocurran dentro del mismo bloque haplotípico, es decir, dentro de la misma región con baja evidencia de recombinación. Un ejemplo de este tipo de estrategias está implementado en la herramienta SNPinfo[235].

Finalmente, los métodos libres de bloques consideran que, dado que la frecuencia alélica de los polimorfismos y las tasas de recombinación varían a lo largo del genoma, predefinir regiones de búsqueda carece de sentido, por lo que basan sus algoritmos en la cuantificación de estadísticos de vecindad, asumiendo el número de tagSNPs total del panel como parámetro de entrada fijado por el usuario[225,236–238].

Los métodos basados en este tercer modelo permiten encontrar tagSNPs capaces de predecir haplotipos en muestras desconocidas, mientras que los métodos basados en los dos primeros logran minimizar el número de tagSNPs necesario para realizar estudios de asociación[239]. Por su parte, los métodos basados en LD identifican tagSNPs capaces de representar polimorfismos que están distantes en el genoma, mientras que los métodos basados en bloques representan regiones contiguas[236]. Por esto los primeros pueden fallar a la hora de distinguir entre los distintos haplotipos existentes en una región ventana, pero los segundos permiten diferenciar todos los haplotipos existentes para una región haplotípica.

2.2.4 Estrategias de fasado de los polimorfismos

La determinación de la procedencia del alelo heredado se hace a través del fasado de dichos tagSNPs. Sin embargo, la ingente cantidad de marcadores empleados en los estudios de informatividad a través de SNPs por técnicas de NGS dificulta la determinación de la procedencia del alelo heredado. Por tanto, hace patente la necesidad del desarrollo de métodos automatizados para el fasado de dichos tagSNPs.

Handyside et al [240] afirmaron que, a pesar de que los SNPs son esencialmente bialélicos y, por tanto, potencialmente menos informativos que los STRs, pueden ser empleados a través del establecimiento del genotipo paterno y el de un hijo enfermo preferiblemente, u otro familiar cercano a la pareja cuyo estado de enfermedad sea conocido, para establecer cuatro conjuntos distintos de SNPs segregantes en los cromosomas parentales de acuerdo a las teorías de herencia mendeliana [218]. Así, mediante la comparativa de los alelos parentales, los alelos del hijo afecto y los alelos presentes en la biopsia embrionaria, puede quedar establecida la presencia o ausencia del alelo afecto en los embriones mediante un estudio indirecto.

Con base en estas afirmaciones, durante el transcurso de esta tesis se desarrolló en la misma empresa (Bioarray S.L) el trabajo Fin de Master de José Leonardo Díaz Chacón, titulado *"Estudio de la cantidad de ADN mitocondrial como marcador de la calidad embrionaria y Desarrollo de nuevas técnicas bioinformáticas para el diagnóstico preimplantacional",* presentado en julio de 2017 para la obtención del título de Master Universitario en Bioinformática por la *Universitat de València*. Dicho trabajo desarrolló un método de fasado automático de polimorfismos en muestras trío, es decir, a partir de las muestras obtenidas de una pareja y su hijo enfermo. Para ello, el algoritmo parte de los archivos HOTSPOT generados a partir de la secuenciación de los oligos diseñados por la plataforma AmliSeq y establece locus a locus, por medio de una matriz temporal, la procedencia de los alelos presentes en el hijo afecto mediante la generación de una matriz temporal. Además, el método permite el establecimiento de keySNPs, definidos por Handyside como aquellos polimorfismos presentes en los embriones que, sin lugar a duda, no se encuentran bajo un efecto ADO[240].

Sin embargo, a pesar de la gran utilidad del método propuesto, encontramos una serie de debilidades a mejorar: en primer lugar el algoritmo no permite el establecimiento de las fases haplotípicas en familias compuestas por miembros diferentes a los que componen las muestras trío ni su aplicación en aquellos casos en los que no se dispone del

material procedente de un hijo afecto. Esto es debido a que el algoritmo fue desarrollado de forma específica atendiendo a los familiares involucrados en cada situación y su extensión a familias con más componentes necesitaría el modelado individualizado de cada caso para considerar el grado de parentesco entre los individuos disponibles. Por otro lado, el empleo de matrices guardadas en archivos temporales complica la aplicación del mismo a casos con más individuos, a la vez que incrementa el tiempo de computación al precisar de la realización reiterada de procesos de lectura y comprobación en cada individuo y posterior escritura en cada uno de los locus secuenciados en el análisis de informatividad, lo cual crece con el número de individuos para los que hay que realizar estos procesos.

2.3 Combinación de DGP-A y DGP-M

Por último, debemos destacar que, a pesar de todas las técnicas disponibles, aún no se ha podido establecer un método rápido, eficaz y económico que combine DGP-A y DGP-M en una única biopsia a través de NGS. Éste hecho resulta muy ventajoso debido a la posibilidad de reducir tanto el tiempo necesario para la obtención de resultados, y con ello el tiempo de espera de los pacientes, reduciendo los niveles de estrés sufridos por la madre, como el costo de todo el proceso, al no ser necesario recurrir a técnicas accesorias y permitir el manejo simultáneo de las librerías destinadas a cada análisis.

El principal motivo es que, hasta ahora, las técnicas utilizadas para DGP-A y DGP-M eran radicalmente distintas. Mientras que para la primera, como ya se ha comentado, se solía utilizar FISH o microarray hasta la incoporación reciente de las tecnologías NGS, para la segunda se utilizaba secuenciación capilar o análisis de fragmentos. Sin embargo, el trabajo desarrollado en esta Tesis permite unificar ambas técnicas bajo el paraguas de una única tecnología: la secuenciación masiva.

II. Hipótesis y Objetivos

- **Propósito general**

Se han identificado las siguientes carencias a la hora de aplicar las soluciones existentes en el estado del arte para disponer de un método rápido, eficaz y económico que combine DGP-A y DGP-M en una única biopsia a través de NGS.

- Las técnicas de filtrado disponibles no están diseñadas para filtrar los archivos BAM procedentes de la secuenciación de librerías DGP-A, al no manejar de manera adecuada duplicados generados con oligos semialeatorios y que, por tanto, no idénticos.

- Los algoritmos de análisis implementados en el momento de escritura de la tesis no permiten determinar con exactitud el nivel de mosaicismo.

- Las distintas técnicas de selección de tagSNPs que podemos encontrar en el estado del arte calculan conjuntos mínimos de tagSNPs que no pueden ser empleados de forma óptima en los análisis DGP-M, ya que no permiten determinar con exactitud los alelos segregantes en cada cromosoma.

- Al no estar contemplado el empleo de los tagSNPs propiamente dicho en los análisis DGP-M, el fasado de los mismos tampoco ha sido automatizado.

El principal objetivo de los algoritmos desarrollados en el marco de esta tesis persigue el diseño de un método de análisis rápido y eficaz que permita aunar los procesos de análisis de DGP-A y DGP-M mediante secuenciación por NGS a partir de una biopsia única. Este hecho permitiría además la realización de biopsias con transferencia en fresco, ya que el proceso completo de laboratorio dura menos de 12 horas y los resultados bioinformáticos son obtenidos en tiempo real, a la par que mejorar la confianza en la información obtenida con respecto a los métodos de diagnosis actuales.

- **Hipótesis de investigación**

La hipótesis de partida de esta tesis doctoral es que se puede conseguir un proceso unificado de análisis de DGP-A y DGP-M mediante secuenciación por NGS a partir de una

biopsia única, si bien requiere solucionar las carencias anteriores. Esto lleva a las siguientes hipótesis de investigación:

- La disminución de la dispersión de las lecturas a través del proceso de filtrado de duplicados y artefactos de PCR permite aumentar la confianza de los resultados obtenidos.

- Mejorar el filtrado de duplicados y artefactos de PCR facilitará la determinación de la ploidía.

- Mejorar la detección del porcentaje de aneuploidía facilitará la determinación de la ploidía y el nivel de mosaicismo.

- Mejorar los niveles de sensibilidad y especificidad en la determinación de la ploidía y el nivel de mosaicismo permitirá no tener que descartar muestras con bajo número de lecturas o alta dispersión de las mismas.

- Mejorar la selección de tagSNPs permitirá ahorrar en tiempo y costo de secuenciación al seleccionar un conjunto menor de polimorfismos.

- Mejorar la selección de tagSNPs a través de la consideración de la informatividad permitirá su aplicación en DGP-M para resolver el problema de la informatividad.

- El diseño de una metodología automatizada de fasado de los polimorfismos analizados facilitará la determinación de los alelos portados, permitiendo así la selección de embriones libres de alteración y reducirá el tiempo necesario para la obtención de resultados.

- **Objetivos específicos**

Las hipótesis de investigación permiten definir los siguientes objetivos específicos de esta tesis doctoral.

Filtrado de duplicados y artefactos de PCR:

Desarrollo de un algoritmo de filtrado de artefactos y duplicados de PCR denominado MiNFilterDups específico para muestras de DGP por NGS, permitiendo que los software de análisis de DGP-A detecten el modelo real de ploidía presentada por la muestra con independencia del número de lecturas y la dispersión de las mismas. Este algoritmo debe ser capaz de:

- Disminuir el valor de MAPD y aumentar la confianza de los resultados de las muestras filtradas con respecto al estado del arte.
- Permitir que el diagnóstico de muestras con valores de MAPD superiores a 0,3 aún sea fiable.

Determinación de la ploidía y el nivel de mosaicismo:

Desarrollo de un algoritmo de análisis DGP-A denominado MiNmos específicamente diseñado para ser capaz de detectar bajos niveles de mosaicismo, controlando los niveles de sensibilidad y especificidad. Dicho algoritmo deberá:

- Gradar y determinar el nivel de mosaicismo de las muestras, indicando el porcentaje de células aneuploides con respecto a las normales.
- Detectar bajos porcentajes de aneuploidía y determinar el nivel de mosacismo.
- Emplear el *valor Z-Score del log10 de los niveles de cobertura corregidos respecto a los valores de las dos líneas base*, como indicador del estado de ploidía y el nivel de mosaicismo.
- Detectar el estado de ploidía y el nivel de mosaicismo incluso con valores de MAPD mayores que 0,3.
- Mostrar una mayor sensibilidad y especificidad en la distinción entre embriones euploides y mosaicos de bajo nivel respecto a las técnicas actuales del estado del arte.
- Precisar de un menor número de lecturas mínimas para obtener resultados fiables respecto a los algoritmos del estado del arte.

Estudio de la dispersión de las lecturas:

A partir del estudio de la aplicación de nuevas técnicas al DGP-A se deduce que el MAPD es un valor inversamente relacionado con el número de lecturas. Esto sugiere que no es posible realizar la comparación de la dispersión de las lecturas a través de los valores de MAPD entre muestras con distinto número de lecturas. Por ello, en este estudio se pretende proponer una alternativa a la medida existente actual de la dispersión de lecturas en las muestras y debatir acerca de su idoneidad en el DGP-A.

También se discutirá acerca de la relación existente entre el MAPD de la muestra y el número de lecturas, y la implicación de dicha relación en la comparación entre muestras con distinto número de lecturas.

Se estudiará si la ploidía de un embrión pueda ser correctamente determinada con independencia de su valor de MAPD si ha sido correctamente filtrada.

Por último se presentará un análisis de medidas alternativas al MAPD para estudiar y comparar la dispersión de las lecturas entre las muestras.

Selección de tagSNPs útiles en análisis DGP-M:

Desarrollar un algoritmo denominado MiNtagSNP capaz de determinar un set mínimo y óptimo de tagSNPs útiles en DGP-M a través de la maximización de la informatividad de los mismos. Dicho algoritmo debe ser capaz de seleccionar:

- Un conjunto mínimo de tagSNPs útiles para los análisis de DGP-M por técnicas de NGS.
- Un menor número de tagSNPs que las técnicas del estado del arte y, además, estos son más informativos.
- tagSNPs que son, en promedio, más independientes que los polimorfismos seleccionados por el estado del arte.

Fasado de SNPs y determinación de embriones aptos para transferencia:

Realizar una validación del algoritmo de fasado de polimorfismos, determinando si es capaz de discernir entre los distintos tipos de herencia y los individuos considerados en el análisis, reconociendo los cromosomas heredados por los embriones a través de polimorfismos cosegregantes para su transferencia. Dicho algoritmo de fasado debe tener la capacidad de:

- Fasar polimorfismos en un análisis DGP-M, distinguiendo entre los distintos tipos de patrón de herencia y grados de relación entre los familiares disponibles.
- Distinguir los embriones portadores no por presentar la alteración, sino por mostrar los polimorfismos cosegregantes con el cromosoma paterno identificado como portador.

A partir de los tagSNPs seleccionados mediante MiNtagSNP:

- Permitir realizar un análisis indirecto de la segregación de la alteración evitando el riesgo de realizar una determinación errónea debido a fenómenos de ADO.
- Permitir disponer de un número de marcadores mucho más elevado que los disponibles con la aplicación de otras técnicas del estado del arte, lo que permite identificar los cromosomas portados por cada embrión a pesar del riesgo de ADO.
- Permitir la posibilidad de aplicar la técnica en casos donde no se disponga de un familiar. Esto se consigue combinando el análisis directo e indirecto, al incluir la mutación en el panel.

100| HIPÓTESIS Y OBJETIVOS

III. Material y métodos

Capítulo 1: Procedimientos comunes

1.1 Selección de muestras

Todas las muestras seleccionadas para la presente tesis proceden de parejas que acudieron a distintos centros de reproducción a someterse a un tratamiento de fecundación *in vitro*. Dichos centros acordaron, junto a la pareja, la realización de un procedimiento DGP-A y/o DGP-M, para lo cual firmaron un consentimiento informado de acuerdo a los postulados establecidos durante el acuerdo firmado en la Declaración de Helsinki[241], y remitieron la muestra (una biopsia embrionaria de trofoectodermo o de blastómera) a Bioarray SL para su análisis siguiendo los protocolos y procedimientos estandarizados tanto en la clínica de reproducción como en Bioarray SL.

Dado que la mayor parte del trabajo aquí presentado es de carácter bioinformático, una vez finalizado el proceso de DGP e informadas de los resultados tanto las clínicas de reproducción como los pacientes implicados, los datos fueron empleados para esta investigación.

Para el estudio de validación de los algoritmos desarrollados para la detección de aneuploidías se seleccionaron datos procedentes de distintas parejas sometidas a DGP-A. Los **criterios de inclusión** de muestras fueron:

- La indicación para la realización de DGP-A debía ser edad materna avanzada o que alguno de los padres fuese portador de traslocación balanceada.
- Calidad embrionaria buena[242].
- Buena calidad de biopsia y entubado.
- Realización de todo el procedimiento anterior a la recepción de la biopsia en Bioarray SL en un laboratorio con experiencia en el campo.
- Proceso de fecundación realizado mediante ICSI[243].

Los **criterios de exclusión** fueron:

- Datos procedentes de parejas portadoras de aneuploidías en cromosomas sexuales.
- Muestras procedentes de embriones, biopsias o procesos de entubado de mala calidad.

Para el estudio de la validación de los algoritmos desarrollados para detección de enfermedades monogénicas se seleccionaron los datos procedentes de parejas que habían

sido previamente diagnosticadas como portadoras de algún trastorno monogénico. Las muestras incluidas procedieron de pacientes que presentaron alteraciones consistentes en mutaciones puntuales, excluyendo aquellas parejas cuyas alteraciones patogénicas hubieran sido diagnosticadas como inserciones o deleciones de más de una base, así como aneuploidías o traslocaciones que afectasen al cromosoma para el cual se realizó el análisis DGP-M, incluso en estado de mosaicismo.

1.2 Ciclo de Fertilización in Vitro (IVF)

Cada centro de reproducción siguió su propio protocolo para el ciclo IVF, así como para la recogida de la biopsia embrionaria. A modo ilustrativo, este protocolo suele consistir en la obtención de los espermatozoides y oocitos de las parejas sometidas al tratamiento. Para recoger los oocitos, las pacientes se someten a un ciclo de estimulación ovárica mediante la administración de Clomifeno, un inductor de gonadotropinas endógenas[244], o de gonadotropina menopáusica humana (HMG por sus siglas en inglés, *Human Menopausic Gonadotropin*)[245] y de hormona estimulante de folículos (FSH por sus siglas en inglés *Follicle-Stimulating Hormone*)[246], según se considere más oportuno en relación a la clínica de la paciente. La administración de hormonas se realiza siguiendo un ciclo largo (el fármaco se inició en fase folicular precoz o lútea del ciclo ovárico anterior) o corto (durante el ciclo que se pretende estimular). Durante todo el ciclo de estimulación las pacientes se monitorizan mediante ecografías seriadas y determinaciones séricas de los niveles de estradiol, a fin de evitar la aparición del Síndrome de Hiperestimulación Ovárica (SHO)[247].

La maduración folicular se induce mediante la administración de gonadotropina coriónica humana (hCG por sus siglas en inglés *Human Chorionic Gonadotropin*)[247,248]. Posteriormente se procede a la punción folicular transvaginal mediante guía ecográfica y aspiración de los oocitos pertenecientes a los folículos que habían madurado[249].

Finalmente, los oocitos se desproveen de la capa de células granulosas de la zona pelúcida (denudación) y fecundan *in vitro* por microinyección gracias a una aguja ICSI con espermatozoides previamente capacitados[250].

1.3 Biopsia embrionaria

Al igual que en el apartado anterior, cada centro realizó la biopsia embrionaria siguiendo sus propios protocolos. En general, esta biopsia se realiza tras cultivar los

embriones en día 5 en placas de Petri Falcon 1006 y biopsiados empleando un láser de diodo de 1,48-µm (Zylos-TK laser, Hamilton Thorne, Beverly, MA, USA). Las biopsias en día 5 se realizaron sobre medio de cultivo atemperado tamponado con Hepes o MOPS. suplementado con albúmina.

Tras el lavado de las células biopsiadas con una solución PBS + PVA 1mg/ml se procedió al entubado de las mismas empleando una pipeta Stripper®. Las muestras se almacenaron para su transporte junto a un control negativo (tubo con el medio de lavado sin muestra). El transporte de las muestras se realizó en frío.

1.4 Preparación de librería para PGT-A

Una vez recibidas las muestras en el laboratorio se procedió a la amplificación del ADN y preparación de las librerías siguiendo el protocolo correspondiente al kit Ion Reproseq (Thermo Fisher Scientific). Este kit está diseñado para la amplificación del material genético a partir de una o pocas células, como son las biopsias embrionarias. Está basando en PCR, y consta de tres pasos:

1. Lisis celular: Al tubo conteniendo la biopsia embrionaria, se le añaden 2,5 ul de un tampón de lisis, y se incuba a XX grados durante 10 minutos.
2. Amplificación: En este paso, se añaden los oligos para producir la amplificación del genoma del embrión. Una característica especial de estos oligos es que, por un extremo, son semialeatorios para producir la amplificación inespecífica del genoma. Por el otro extremo, tiene una secuencia autocomplementaria. Esta secuencia hace que, tras la amplificación, la hebra generada forme una horquilla, de manera que no estará disponible para su amplificación en los siguientes ciclos. De esta manera, se trata de evitar la sobreamplificación de las hebras generadas durante los primeros ciclos, de forma que la amplificación es más homogénea.
3. Amplificación exponencial y adición de códigos de barras: en este paso, se utilizan unos nuevos oligos que son, por un lado, complementarios a los anteriores. Por el otro extremo, tienen un código de barras molecular que utiliza el secuenciador para detectar de qué paciente se trata.

Tras este paso de amplificación del genoma del embrión, las muestras ya están listas para secuenciar, o bien se purifican mediante bolas magnéticas y se utilizan para el siguiente paso.

En la validación de los algoritmos desarrollados para el análisis de aneuploidías se siguió el protocolo descrito más adelante en el bloque 2.1.1 Procesado de la muestras del bloque *III* Material y métodos*;* para las muestras de la validación de los algoritmos desarrollados para el análisis de alteraciones monogénicas se siguió el protocolo descrito más adelante en el apartado *3.2.3.* Procesado de muestras del bloque *III* Material y métodos. Todas las librerías fueron secuenciadas empleando el secuenciador Ion PGM™ Sequencing Hi-Q (ThermoFisher Scientific), usando chips 318 (que permiten secuenciar hasta 24 muestras a la vez).

Finalmente, las lecturas obtenidas para cada biopsia embrionaria fueron alineadas con el genoma humano de referencia[251] empleando el Software Ion Torrent Suite v5.0.4 para generar los archivos BAM que se emplearon en las validaciones.

2.1 Algoritmo para el filtrado de duplicados

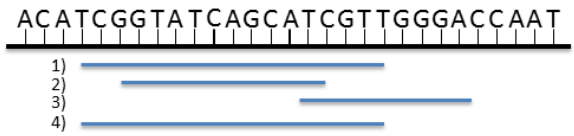

Figura 12: Ejemplo de distintos duplicados de PCR.

La Figura 12 muestra distintos tipos de artefactos de PCR que se forman como consecuencia del protocolo de amplificación y/o preparación del molde utilizado. Si tomamos la lectura 1) como el producto de PCR incial, la lectura 2) puede haberse formado como resultado de la unión de los oligos aleatorios a la lectura 1) y su posterior amplificación. La lectura 3) también no es un duplicado de PCR porque si bien su origen solapa con la lectura 1), la parte final está más allá de ésta. Sin embargo, este solapamiento parcial podría afectar al algoritmo de cálculo de ploidía, dado que para ese en concreto existe una mayor cantidad de lecturas. Finalmente, la lectura 4 puede tener dos posibles orígenes: en primer lugar, haberse formado por una unión de oligos al primer fragmento, o bien haberse formado a consecuencia de la unión del fragmento de ADN original a dos esferas ISP. Como ya se ha comentado previamente en el Bloque I Introducción, las herramientas disponibles en el estado del arte están diseñadas para detectar duplicados de PCR originados a partir del oligos idéntidos, no de oligos semialeatorios. Por tanto, detectan que la lectura 4) es un duplicado, pero no son capaces de identificar las lecturas 2) y 3). Todos estos artefactos contribuyen por igual a la distorsión del perfil de las lecturas de las muestras, dificultando la identificación del estado de ploidía de las mismas.

2.1.1 MiNFilterDups

Con el objetivo de poder desarrollar un método rápido y eficaz que permita eliminar los artefactos de PCR de los archivos BAM en un tiempo de computación razonable, se decidió codificar el algoritmo combinando los lenguajes de programación Perl[252] y Bash[253] con las opciones *sort* y *view* del software *SAMtools*[189].

La Figura 13 muestra el desarrollo del algoritmo implementado. En primer lugar, MiNFilterDups emplea la función *sort* de *SAMtools* para ordenar las lecturas del archivo

BAM según la posición de origen en el genoma de referencia. Cada lectura es leída por el algoritmo, que analiza y guarda la posición de origen y el tamaño de dicha secuencia; si la siguiente lectura comienza en una posición localizada dentro de la región que cubre la primera lectura, (es decir, si ambas solapan en algún punto), entonces MiNFilterDups compara el tamaño de ambas lecturas y elimina la de menor tamaño, conservando como referencia la posición inicial de la primera. Si la segunda secuencia comienza en un nucleótido posterior al último nucleótido de la primera secuencia, el proceso comienza de nuevo tomando como referencia esta segunda lectura y los datos se actualizan con los nuevos valores. Este proceso se repite hasta que todas las lecturas del archivo BAM han sido analizadas.

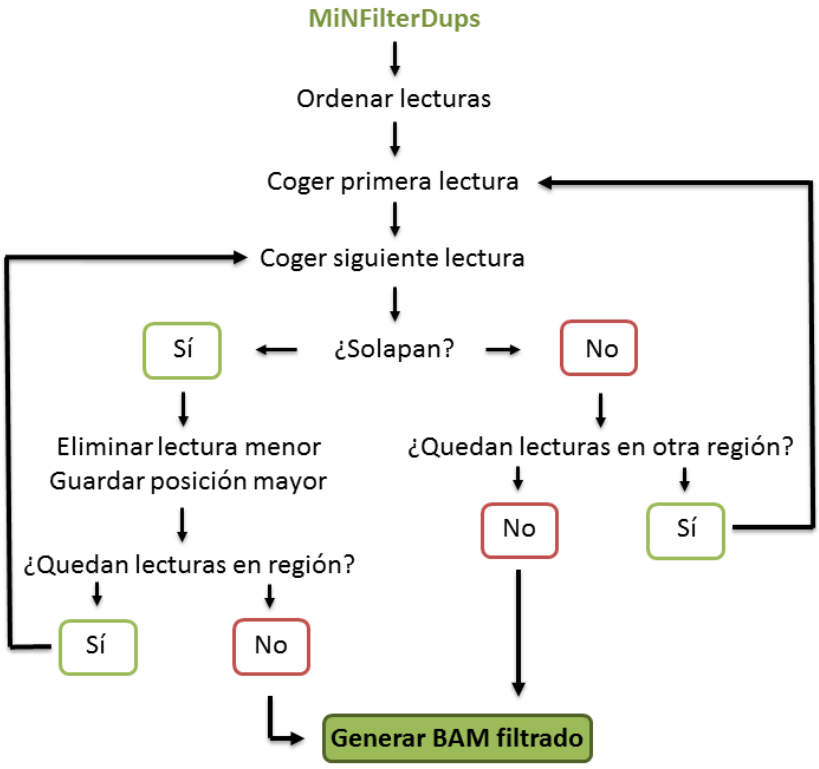

Figura 13: Desarrollo de MiNFilterDups .

2.1.2 Diseño de la validación

Para demostrar la eficiencia de nuestro algoritmo de filtrado se decidió seguir el protocolo descrito en la Figura 9 del bloque I Introducción pero empleando MiNFilterDups en el paso "Eliminación de artefactos de PCR". Los resultados obtenidos fueron comparados con los resultados arrojados al emplear el algoritmo incluido en Ion Reporter Server en su

versión 5.0 (sin filtro de duplicados) y 5.2 (con el flujo de trabajo *FilterDuplicates*). Los aspectos evaluados fueron el valor de MAPD y el número de lecturas eliminadas. También se midió el tiempo de ejecución empleado, un parámetro que resulta esencial a la hora de obtener resultados en análisis DGP-A, ya que un algoritmo computacionalmente costoso, a pesar de ser eficiente, provocaría ralentizar demasiado el proceso, si hubiera que analizar un número alto de embriones.

Todos los embriones analizados fueron obtenidos a partir de parejas que se habían sometido previamente a ciclos de DGP-A en nuestro laboratorio. La selección de las muestras que conformaron la validación se realizó con base en los cromosomas afectados, de forma hubiese representación de la mayoría de ellos o, al menos, de un cromosoma perteneciente a cada cromosoma.

Así, se utilizaron los archivos BAM correspondientes a 15 embriones biopsiados en día 5 (4 euploides masculinos, 4 aneuploides masculinos, 3 euploides femeninos y 4 aneuploides femeninos), que contenían monosomías para los cromosomas 1, 3, 5 ,8 ,12, 14, 19 y 22 y trisomías para los cromosomas 13, 14, 18 y 21. Además, uno de los embriones presentó una trisomía parcial del brazo largo del cromosoma 1 y monosomía para su brazo corto.

Para determinar el impacto que tiene el filtrado de duplicados en la calidad global de los resultados del DGP-A, se decidió no solo evaluar el valor de MAPD, sino también la relación respecto al número de lecturas. Como se ha comentado anteriormente, en general, a menor número de lecturas, mayor dispersión y, por tanto, mayor MADP. Lo que tratamos de determinar en este punto es si, gracias a la mejora del filtro de duplicados se puede mejorar el valor de MAPD. Cabe recordar que reducir el número de lecturas necesarias para tener resultados fiables es un punto importante en este proceso, puesto que permite un mayor nivel de multiplexación, lo que se traduce en una reducción de costes.

Para ello, utilizando la función *subsampling* del software *SAMtools 1.3.1*[189], se realizó una selección aleatoria de las lecturas de cada archivo BAM de los embriones originales, generando nuevos archivos BAM con rangos entre 5.000 y 100.000 lecturas. Estos archivos se generaron por triplicado de manera independiente, generando un set final compuesto por 450 archivos BAM (15 embriones $_{x}$ 10 categorías de lecturas $_{x}$ 3 repeticiones). Se comprobó que la proporción de lecturas por cromosoma se mantenía tras la eliminación de las lecturas durante la generación de los archivos.

Finalmente, se analizó un embrión denominado *T1*. Este embrión fue señalado por IRS como 46,XX,del9(pter), pero la inspección visual del perfil sugiere que el embrión presenta un cariotipo aberrante diferente al señalado.

2.1.3 Evaluación del tiempo de ejecución

La teoría de la complejidad es una rama de la teoría de computación centrada en la clasificación de los problemas computacionales por medio de la cuantificación de los recursos necesarios para su resolución. El estudio del tiempo de ejecución de un proceso permite evaluar si un método es realmente efectivo y puede ser empleado para aquello que ha sido diseñado.

En este caso particular, el dominio del problema se centra en el diseño de un algoritmo efectivo a la hora de eliminar artefactos de PCR para DGP-A, pero también debe ser suficientemente rápido como para permitir que todo el proceso de análisis de las muestras embrionarias biopsiadas se realice en el menor tiempo posible, de manera que sea viable realizar transferencias en el menor intervalo de tiempo. Esto es especialmente crítico cuando se pretende realizar la transferencia en D6 tras una biopsia en D5.

En nuestro estudio, este parámetro fue evaluado usando un ordenador con procesador *Xeon E5-2407 V2 2.4GH, 64GB RAM.* Para ello se midió el tiempo empleado en el filtrado de los archivos generados por cada uno de los algoritmos evaluados.

2.2 Algoritmo para la detección del porcentaje de mosaicismo

2.2.1 MiNmos

El algoritmo para la determinación del porcentaje de ploidía está basado en la aplicación de un modelo Z-Score[254]. Un modelo Z-Score es una medida de la relación entre los elementos de un grupo e indica cuántas veces la desviación estándar un elemento está alejado respecto de la media. Así, mediante el uso de este modelo se puede comparar eficientemente datos en distinta escala. La ecuación que define este modelo se recoge en la siguiente línea:

$$Z = \frac{x_i - \bar{x}}{\sigma}$$

siendo x_i el valor de cada elemento, $\bar{x}$ la media poblacional y σ la desviación estándar. Se trata por tanto de una medida adimensional obtenida a partir de la sustracción de la media a cada valor individual y su división por la desviación. El valor del Z-Score puede ser positivo o negativo, de manera que los valores positivos indican que el valor se encuentra tantas desviaciones por encima de la media, mientras que los valores negativos sitúan los elementos por debajo.

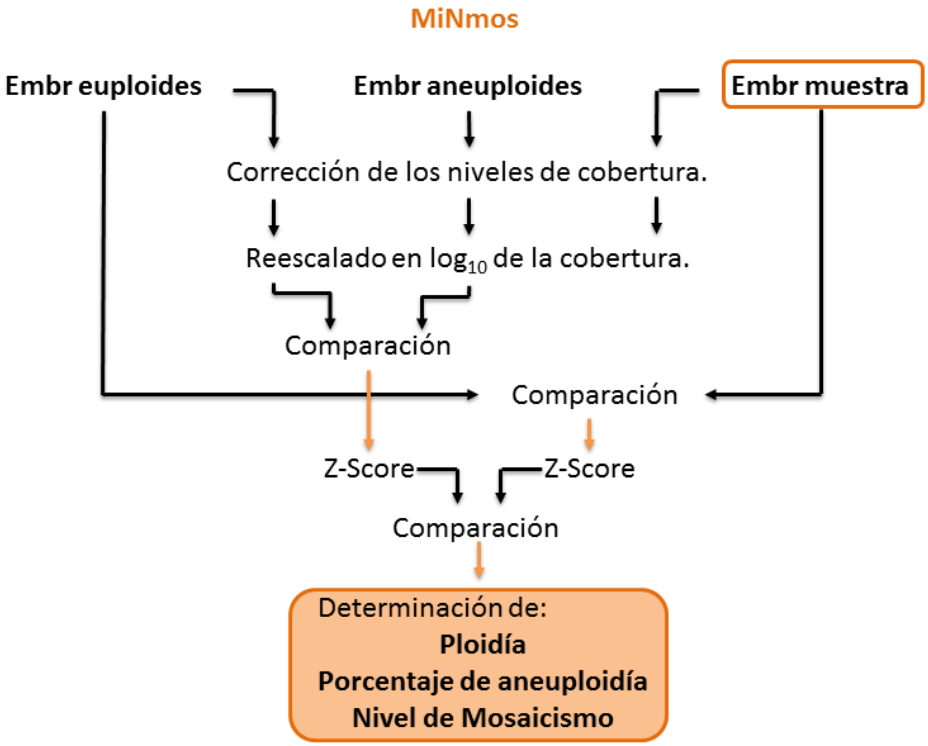

Figura 14: Desarrollo de MiNmos.

Así, MiNmos emplea la información de la cobertura en un modelo estadístico que determina a qué estado de ploidía pertenece una región genómica concreta, permitiendo una determinación más precisa del porcentaje de mosaicismo. La Figura 14 muestra el desarrollo del algoritmo diseñado. Para ello, MiNmos realiza un reescalado en logaritmo en base 10 de los valores de cobertura de las muestras analizadas y compara la información con una línea base formada por los archivos BAM de 10 embriones masculinos euploides, de manera que se puede obtener el valor de Z-Score correspondiente. El resultado obtenido es entonces comparado con una segunda línea base reconfigurada formada a partir de embriones cuyo porcentaje de aneuploidía es conocido, de manera que se puede obtener el valor de Z-Score correspondiente a cada porcentaje de aneuploidía. El uso de 10 archivos

BAM para cada línea base simplifica la varianza que se podría generar con el uso de un solo archivo BAM aleatoriamente escogido[255].

El valor de Z-Score obtenido puede variar de cromosoma a cromosoma debida al contenido en secuencias repetitivas, contenido en GC, tamaño del cromosoma etc. Por este motivo se decidió establecer una segunda línea base, en este caso diseñada para cada aneuploidía en cada cromosoma. Esta línea base se construyó utilizando embriones con aneuploidías conocidas. Así, el valor de Z-Score obtenido a partir de la muestra en comparación con la línea base euploide es comparado con la escala de valores obtenida por la línea base aneuploide, permitiendo el establecimiento del porcentaje de aneuploidía, que se encuentra previamente gradado en categorías.

2.2.2 Validación del algoritmo

La validación se realizó utilizando los archivos BAM correspondientes a 15 embriones biopsiados tanto en día 3 como en día 5 que contenían aneuploidías para distintos cromosomas.

A partir de los archivos de estos embriones y usando la función *merge* del software SAMtools 1.3.1[189], se construyó un set de archivos BAM mezclando aleatoriamente lecturas procedentes de un embrión aneuploide con uno euploide. Los porcentajes de lecturas tomadas de cada embrión variaron desde la combinación aneuploide-euploide 90%-10% hasta el 10-90%, en categorías decrecientes de 10 en 10%. Además, se controló que las lecturas tomadas aleatoriamente de cada embrión respetasen la proporción de lecturas de cada cromosoma al igual que se ha descrito en la validación de MiNFilterDups.

También se generó un segundo set *in silico* para medir el número mínimo de lecturas necesario para identificar correctamente el mosaicismo, siguiendo la misma metodología empleada en la validación del filtro de duplicados. Cada archivo BAM se procesó tres veces de manera independiente a fin de evitar posibles desviaciones de los datos debidas al efecto del azar.

El análisis de los resultados se realizó utilizando el algoritmo MiNFilterDups para el filtrado de duplicados seguido de la normalización del contenido GC con IRS[256].

La efectividad del algoritmo MiNmos desarrollado en este trabajo en la detección del porcentaje de aneuploidía y determinación del nivel de mosaicismo se evaluó

considerando los valores de MAPD, número mínimo de lecturas necesarias para detectar correctamente la ploidía y nivel de mosaicismo, el porcentaje de aneuploidía detectado, la sensibilidad y la especificidad.

Tanto el MAPD como el número mínimo de lecturas fueron descritos anteriormente. Por su parte, la sensibilidad se define como el número de verdaderos positivos detectados por una técnica; en este caso, se consideraron verdaderos positivos los embriones con aneuploidías en mosaicismo que fueron correctamente diagnosticados respecto al total de embriones mosaico. Por su parte, la especificidad se define comúnmente como el número de verdaderos negativos es decir, embriones euploides correctamente identificados con respecto al total de embriones euploides.

$$Sensibilidad = \frac{VP}{VP + FN} \quad ; \; Especificidad = \frac{VN}{VN + FP}$$

MiNmos realiza una catalogación del nivel de mosaicismo de la muestra, es decir, establece una serie de categorías para distintos niveles de mosaicismo. Estas categorías partían de un 10% de aneuploidía en incrementos del 10% hasta llegar al 90%. Así, nos encontramos ante una clasificación multiclase que provoca que el dominio del problema no pueda ser analizado empleando una curva ROC[257], ya que estas curvas están diseñadas para variables categóricas de tan solo dos categorías.

Para poder analizar si el algoritmo es realmente eficiente, se decidió que cada nivel de mosaicismo representaría una clase distinta, de manera que la sensibilidad y especificidad se calcularon para cada clase del nivel de mosaicismo. Así, la sensibilidad finalmente mide el número de muestras pertenecientes a determinada categoría de mosaicismo que fueron correctamente diagnosticadas (verdaderos positivos), mientras que la especificidad midió el número de muestras pertenecientes a la categoría superior de mosaicismo que no fueron catalogados dentro de la categoría de estudio (verdaderos negativos).

2.2.3 Gradación del mosaicismo

La gradación de las diferentes clases del nivel de mosaicismo se realizó comparando el valor Z-Score del logaritmo en base 10 de los niveles de cobertura de la línea base aneuploide con la línea base euploide, con base en el valor del porcentaje de aneuploidía

detectado en la muestra analizada. Estos valores fueron calculados para cada ventana de 2 Mb.

El nivel de mosaicismo se estableció de manera discreta mediante la creación de varias categorías de mosaicismo. Como se ha mencionado anteriormente, estas categorías partían de un 10% de aneuploidía, en incrementos de 10 en 10 hasta llegar al 90%. Para establecer el valor umbral óptimo para cada clase del nivel de mosaicismo se analizaron y compararon los valores de sensibilidad y especificidad de cada categoría utilizando tres de los estadísticos más extendidos en cuanto a categorización se refiere:

a) En primer lugar se consideró establecer como límite de clase aquel valor que contuvo el 95% de los datos centrales de dispersión de los valores Z-Score de la cobertura de la línea base aneuploide. Así, el valor umbral A de una clase i respecto a la siguiente j se estableció en función a la ecuación $PtoCorte_{Aij} = Z_x - 2 * \sigma$, siendo Z_x el valor del Z-Score en cada muestra y σ la desviación típica.

b) Como segunda opción y teniendo en cuenta el modelo de distribución probabilística, se decidió emplear el valor del tercer cuartil de cada clase, calculado a partir de la media X_i y desviación estándar de los valores de Z-Score en cada clase. El tercer cuartil por definición establece un punto de corte que permite estimar correctamente el 75% de los valores pertenecientes a la categoría de estudio. Clases de mosaicismo con ganancia de material $+i$ tienen valores umbrales positivos, por lo que la ecuación empleada fue $PtoCorte_{B+ij} = Z_x + 0{,}674 * \sigma$; mientras que las clases con pérdida de material $-j$ presentan valores negativos y la ecuación empleada fue $Z - 0{,}674 * \sigma$ (el uso de la primera ecuación resultaría en el cálculo del primer cuartil). Nuevamente Z_x es el valor de Z-Score y σ la desviación entandar.

c) La última opción fue considerar como punto de corte el punto medio resultante del Z-Score de las muestras pertenecientes a un nivel de mosaicismo i con respecto a las muestras de la clase siguiente j. La ecuación sería $PtoCorte_{Cij} = (Z_x + Z_j)/2$.

Además de la sensibilidad y especificidad, para comparar qué conjunto de valores umbral establece una mejor clasificación de las categorías de mosaicismo se calcularon dos índices más:

El Índice exacto o de exactitud I_e muestra la probabilidad de que el test realice un diagnóstico correcto[258] y su ecuación es:

$$I_e = \frac{(V_p + V_n)}{T_n}$$

siendo V_p los verdaderos positivos o embriones que pertenecen realmente a la clase de estudio, V_n los verdaderos negativos o embriones pertenecientes a la siguiente clase que no son clasificados en la clase de estudio y T_n el total de embriones analizados en ambas clases.

El segundo índice calculado fue el Índice de Youden Y_i :

$$Y_i = \max\ (Sensibilidad + Especificidad - 1)$$

Debido a que sensibilidad y especificidad son generalmente complementarias, Y_i permite maximizar la diferencia entre verdaderos y falsos positivos[259]. Así, un punto de corte óptimo sería aquel que maximizase la diferencia del Y_i, generando un consenso entre especificidad y sensibilidad que garantice la correcta clasificación y evitar errores de sub o súper estimación del nivel de mosaicismo.

Tanto el valor de I_e como el de Y_i se mueven entre 0 y 1, siendo 1 cuando todas las muestras son correctamente clasificadas para el I_e y el compromiso óptimo entre sensibilidad y especificidad en el Y_i.

2.3 Estudio de la medida de dispersión de las lecturas

Como ya se ha comentado anteriormente, la dispersión de las lecturas es un parámetro esencial para determinar la fiabilidad del resultado arrojado por el análisis DGP-A. Generalmente, el MAPD es el parámetro escogido para reflejar esta fiabilidad. Sin embargo, el estudio de los datos a través de los algoritmos presentados en esta tesis arrojaron cierta incertidumbre acerca de la idoneidad de emplear esta medida en comparaciones de muestras con distinto número de lecturas debido a la fuerte correlación existente entre el número de lecturas y el valor de MAPD mostrado. Este hecho provoca que muestras con menor número de lecturas muestren un valor de MAPD mayor, lo que se traduce en una mayor dispersión teórica y, por tanto, una menor fiabilidad de los resultados obtenidos. Sin embargo esto no tiene que ser necesariamente cierto, ya que muestras con menor número de lecturas pueden presentar menor dispersión que muestras con gran número si se reparten homogéneamente entre las ventanas, aunque esto es poco probable

que suceda, y el MADP no permitiría comparar la fiabilidad de los resultados obtenidos de manera absoluta y con independencia del número de lecturas de la muestra.

Para estudiar posibles medidas alternativas de la dispersión de las lecturas se decidió emplear los valores de cobertura corregidos para las desviaciones de GC y evaluar diferentes descriptivos. Así, los estadísticos escogidos para el conjunto de valores de cobertura $X = \{x_1, \dots, x_k\}$ con media $\bar{x}$ y desviación típica σ fueron:

- Rango: Diferencia entre el límite superior (valor mayor) y el inferior (valor menor).

$$R = x_k - x_1$$

- Rango medio: Media del límite superior y el inferior.

$$Rm = \frac{\min(X) + \max(X}{2}$$

- Rango Intercuartílico: Diferencia entre el tercer y el primer cuartil.

$$Rq = Q3(X) - Q1(X)$$

- Varianza: Media de los cuadrados de las distancias de los datos a la media o, lo que es lo mismo, suma de cada distancia al cuadrado dividida entre el total de datos.

$$Var = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_k - \bar{x})^2}{N}$$

- Desviación típica o estándar: Raíz cuadrada de la varianza

$$\sigma = \sqrt{Var}$$

- Coeficiente de la variación: Expresa la desviación estándar como porcentaje de la media aritmética.

$$Cv = \frac{\sigma}{|\bar{x}|}$$

- Z-Score o unidad tipificada: Muestra el número de desviaciones típicas que un valor dado se sitúa por encima o debajo de la media de su muestra o población.

$$Z = \frac{x_i - \bar{x}}{\sigma}$$

- Z-ScoreAbs: Como el valor de Z-Score puede ser negativo o positivo según el valor se sitúe por encima o por debajo de la media, se decidió representar el valor absoluto de éste valor.

$$Z_{abs} = |Z|$$

- Z-ScoreAbs_NegLogaritmo: Valor negativo del logaritmo del valor absoluto del Z-Score.

$$Z_{abs\_Neglog} = -Log(Z_{abs})$$

Capítulo 3: Enfermedades monogénicas: DGP-M

3.1 Algoritmo de selección de tagSNPs para maximizar la informatividad en DGP-M

3.1.1 MiNtagSNP

"Esta tesis doctoral está sometida a procesos de protección o transferencia de tecnología o de conocimiento, por lo que esta sección está inhibida en la publicación en los repositorios institucionales. Esta sección contiene, además, los siguientes elementos inhibidos:

Figura 15: Desarrollo MiNtagSNP. En ella se representan las dos partes que componen el algoritmo: SPA y SSA

Autorizado por la Comisión General de Doctorado de la Universidad de Murcia con fecha 24 de febrero de 2021"

3.1.1.1 SPA: Algoritmo de predicción de SNPs

"Esta tesis doctoral está sometida a procesos de protección o transferencia de tecnología o de conocimiento, por lo que esta sección está inhibida en la publicación en los repositorios institucionales.

Autorizado por la Comisión General de Doctorado de la Universidad de Murcia con fecha 24 de febrero de 2021"

3.1.1.2 SSA: Algoritmo de selección de SNPs

"Esta tesis doctoral está sometida a procesos de protección o transferencia de tecnología o de conocimiento, por lo que esta sección está inhibida en la publicación en los repositorios institucionales.

Autorizado por la Comisión General de Doctorado de la Universidad de Murcia con fecha 24 de febrero de 2021"

3.1.1.3 Especificaciones del usuario

"Esta tesis doctoral está sometida a procesos de protección o transferencia de tecnología o de conocimiento, por lo que esta sección está inhibida en la publicación en los repositorios institucionales.

Autorizado por la Comisión General de Doctorado de la Universidad de Murcia con fecha 24 de febrero de 2021"

3.1.1.4 Tiempo de ejecución

"Esta tesis doctoral está sometida a procesos de protección o transferencia de tecnología o de conocimiento, por lo que esta sección está inhibida en la publicación en los repositorios institucionales.

Autorizado por la Comisión General de Doctorado de la Universidad de Murcia con fecha 24 de febrero de 2021"

3.1.2 Valores óptimos de MaxP, HETrate, r^2 y D'

El valor mínimo recomendado de r^2 y *D'* para asegurar una correlación significativa es 0,75[264]. Con el fin de recabar el máximo de información a partir del set de tagSNPs elegido, este valor umbral se estableció en 0,9 y 0,85 respectivamente.

Como el usuario también puede fijar el valor de MaxP y HETrate, se decidió realizar un pequeño experimento previo analizando los SNPs contenidos en la región chr2:9.000.000-11.000.000, que permitió establecer los valores óptimos de corte para estos parámetros en población europea.

Para ello se obtuvieron los valores de fase de los SNPs de los 404 individuos pertenecientes a la superpoblación Europea registrada en la base de datos *1000GenomesDB*. *Finnish inFinland* (FIN), *British in England and Scotland* (GBR), *Iberian Population in Spain* (IBS) y *Toscani in Italy* (TSI)[265]. Algunos de estos individuos fueron seleccionados aleatoriamente para generar 300 cruces, simulando la situación real actual de la población Europea, donde los cruces suceden al azar sin tener en cuenta el valor de *fitness* de cada individuo implicado en el cruzamiento[266] y varios individuos pueden formar varios cruces mientras que otros no forman parte de ninguno.

La informatividad media para cada SNP quedó definida como el promedio de cruces en los que el SNP resultó informativo. La informatividad por cruce se definió como el ratio de tagSNPs que resultaron informativos en el cruce. La informatividad total del panel se definió como el promedio de la Informatividad media para cada SNP. Los valores de MaxP y HETrate fueron representados contra los valores de informatividad media obtenida para cada SNP para establecer el valor umbral óptimo de cada parámetro.

3.1.3 Diseño de la validación

La validación del poder estadístico del algoritmo MiNtagSNP se evaluó mediante dos parámetros diferentes:

Por un lado se estudió la capacidad del algoritmo de seleccionar tagSNPs informativos útiles en los análisis de DGP-M. Para ello se realizó una simulación *in silico* a partir de los datos registrados en la base de datos *1000GenomesDB*[265].

Por otro lado se analizó la precisión de imputación de dichos tagSNPs, es decir, la capacidad de ser empleados para inferir el valor del resto de polimorfismos a los que cubren. Nuevamente, esto se realizó *in silico* con base en los datos recogidos por la población europea de *1000GenomesDB*. Por otro lado, en laboratorio resulta muy interesante poder inferir los valores de los otros tagSNPs cuando se produce un fallo de amplificación de alguno de ellos. Esta idea también fue testeada en una validación *in vitro*.

3.1.3.1 Validación in silico

La informatividad de los tagSNPs seleccionados con MiNtagSNP se evaluó en relación al resultado obtenido por otros 3 métodos de cálculo de tagSNPs a través de una simulación *in silico* de cruces aleatorios entre los individuos registrados en la población Europea de *1000GenomesDB*[265] en el diagnóstico DGP-M de 4 genes de interés:

- VHL chr13:10,181,319 – 10,197,354
- CFTR chr7:117,118,017 - 117,310,718
- ATXN2 chr12:111,888,018 - 112,039,480
- PKD1 chr16:2,136,711 - 2,187,899

Estos genes fueron escogidos por encontrarse entre los más prevalentes. Para ello se estableció un perímetro de interés de 2Mb aguas arriba y 2Mb aguas abajo y se hallaron los tagSNPs de dicha región. Las mutaciones en el gen *VHL* son responsables del Síndrome de von Hippel-Lindau o angiomatosis familiar cerebeloretinial, causante de la aparición de tumores, principalmente de riñón, cerebelo, bulbo y médula espinal, así como afectar a la retina. Por su parte, las mutaciones del gen *CFTR* provocan la enfermedad conocida como fibrosis quística o mucoviscidosis, lo cual ocasiona que aumenten las concentraciones de cloro y sodio en las secreciones corporales. También genera la ausencia bilateral congénita de conductos deferentes, azoospermia y esterilidad en varones. Las alteraciones en el gen

ATXN2 provocan ataxia espinocerebelosa tipo II o tipo Holguin, un proceso caracterizado por problemas progresivos del movimiento que afectan a la coordinación motora y al equilibrio (ataxia). También pueden verse afectadas la deglución, el habla, la espasticidad y la musculatura motora ocular. Por último, las mutaciones del gen *PKD1* provocan poliquistosis renal dominante tipo I, que genera la formación de quistes en los riñones y su hipertrofia.

Por tanto, se obtuvieron 4 sets de tagSNPs en función del método de cálculo empleado:

a) TagSNPs seleccionados por el algoritmo *SNPinfo*, una herramienta bastante intuitiva de emplear, disponible en versión web, que selecciona los tagSNPs con base en análisis de LD[235] y la asignación de pesos estadísticos basados en el p-GWAS obtenido del análisis GWAS entre los polimorfismos.

b) TagSNPs seleccionados por el algoritmo MiNtagSNP.

c) TagSNPs incluidos en el array de SNPs GWAS Omni2.5 de Illumina[267] correspondientes a la misma región cubierta por a) y b). La razón de emplear esta comparación reside en que dicho array es la base sobre la que se diseñó el Karyomapping[218].

d) Mismo número de tagSNPs obtenidos en b) pero elegidos de forma aleatoria.

Los valores de los SNPs para cada individuo fueron obtenidos a partir de la base de datos *1000GenomesDB*. Esta base de datos fue escogida debido a que la base de datos *HapMap*[238,268–271], referencia en la mayor parte de los estudios desarrollados sobre SNPs durante mucho tiempo, fue retirada en 2016. En su dominio web puede leerse actualmente una referencia citando 1000GenomesDB como la base de datos más apta para el desarrollo de trabajos centrados en el cálculo con polmorfismos[272].

Los individuos de la población fueron aleatoriamente separados en dos grupos: el primero estuvo compuesto por 101 individuos que actuaron como *población muestral*; el otro grupo lo formaron los 303 individuos restantes y fueron considerados como la *población referencia*. Para los métodos b y d, se calcularon tagSNP seleccionados por el panel MiNtagSNP y por el panel aleatorio. Los tagSNP de a) y c) fueron obtenidos respectivamente

a partir del sitio web y de las posiciones descritas para el array de SNPs, públicamente accesibles en la red.

Finalmente, los ratios de informatividad fueron calculados siguiendo una metodología similar a la descrita en la sección anterior. En este caso, una vez descartados los individuos relacionados por parentesco familiar (información disponible dentro de la propia base de datos), se procedió a la simulación de todos los cruces posibles dentro de la población seleccionada teniendo en cuenta el sexo del individuo, de tal manera que todos los individuos varones fueron cruzados con las mujeres en la base de datos.

Por otro lado, se evaluó el poder de imputación, es decir, la capacidad de los paneles de tagSNPs de predecir el valor del resto de polimorfismos de una muestra. Para ello se empleó el software BEAGLE en su versión 4.1[273,274], el cual permite realizar la imputación por medio de la comparación de los tagSNPs secuenciados en la *población muestral* frente al valor de todos los polimorfismos contenidos la población que actúa como *referencia*. La imputación se calculó para los mismos cuatro sets en los cuatro genes descritos en esta sección. La eficiencia de imputación del panel quedó definida como el promedio del número de SNPs correctamente imputados.

3.2 Fasado de SNPs en estudios de DGP-M

Una parte fundamental para poder completar cualquier estudio de PGT-M es realizar el fasado de polimorfismos, es decir, determinar inequívocamente a qué alelo pertenece (al sano o al portador de la mutación) cada uno de los polimorfismos identificados, ya sean STR o SNP.

El fasado consiste en, una vez obtenido y secuenciado el ADN de la pareja sometida a IVF y de al menos un familiar cuyo estatus de enfermedad sea conocido, emplear la información relativa a los tagSNPs informativos de la pareja para fasar al familiar. Posteriormente, el resultado es empleado para fasar los embriones y conocer qué alelo ha sido heredado en cada caso, permitiendo así descartar embriones que posean la mutación (debido a la detección directa de la misma) o los SNPs ligados a ella (lo que indicaría un efecto ADO de la posición de la alteración). Como se ha mencionado anteriormente, previamente se había desarrollado un algoritmo para el fasado de SNPs para los casos más simples (trios padre-madre-hijo) en el trabajo Fin de Master de Don José Leonardo Díaz Chacón, y que ha sido utilizado en esta parte.

En primer lugar el algoritmo analiza para un locus todos los tagSNP e identifica aquellos que son informativos. Tras esto, se genera una matriz de datos donde cada columna representa uno de los cromosomas del individuo o embrión. Posteriormente, el genotipo del individuo afecto es considerado como la referencia y empleado para establecer las fases genotípicas de cada locus informativo en la pareja teniendo en cuenta el grado de parentesco que los relaciona y el tipo de herencia que cursa con la alteración estudiada.

Una vez realizado este proceso, se analiza cada SNPs en los embriones para localizar polimorfismos clave o key SNPs, que son aquellos que, además de ser informativos, son heterocigotos en el embrión. Estos polimorfismos son muy importanes, puesto que evidencian inequívocamente que no se ha producido un efecto ADO en las muestras embriónicas. Debemos recordar que no existe garantía de que los SNPs no *key* detectados en homocigosis sean efectivamente homocigóticos, pues pueden estar bajo un efecto ADO que impida detectar ambos alelos, o incluso ser consecuencia de una monosomía. El genotipo fasado de los embriones se compara con la referencia en aquellos loci que resultan llave, evitando así efectos de ADO. El algoritmo representa cada polimorfismo de manera ordenada según la posición cromosómica, coloreando la casilla de acuerdo al haplotipo parental heredado. Finalmente, compara los posibles sucesos de sobrecruzamiento y corrige cualquier posible desviación de los datos.

Este fasado de SNPs permite detemrinar qué polimorfismos segregan con el alelo sano, y cuáles con el patogénico.

La implementación de este algoritmo fue realizada íntegramente mediante el lenguaje de programación Python y es operable a través de la línea de comandos. También está disponible una versión web para clientes accesible únicamente a través del servidor local de la empresa Bioarray S.L pues, como puede consultarse más adelante, toda la invención se encuentra protegida bajo una patente que se encuentra solicitada en el momento de escritura de esta tesis doctoral.

Un ejemplo del proceso puede observarse en la Figura 16 la cual refleja el caso más simple, una muestra trio formada por una pareja donde el padre y su hijo son portadores afectos de una alteración autosómica dominante.

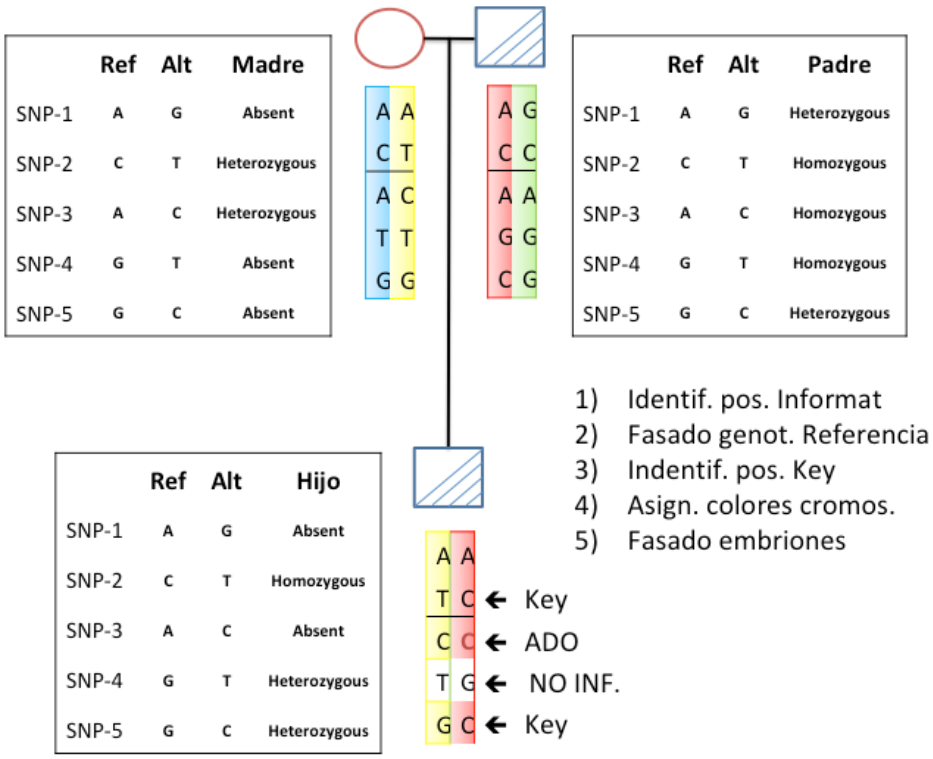

Figura 16: Ejemplo de fasado. En los cuadros se esquematiza la llamada de cada SNP para cada individuo en los archivos HotSPot.

Supongamos cinco loci bialélicos situados en orden según su posición cromosómica y próximos a la alteración de interés, la cual se encuentra situada entre el segundo y el tercero: el SNP-1 con alelos referencia y alternativo respectivamente en el conjunto {A, G}, el SNP-2 {C, T}, el SNP-3 {A, C}, el SNP4 {G, T} y el SNP-5 {G, C}. El algoritmo interpreta que las tres primeras posiciones y la quinta son informativas y descarta la cuarta, pues a pesar de que se sabe que el hijo será heterocigoto y, por tanto, tendrá un alelo referencia y otro alternativo, no es posible distinguir de cuál de los dos cromosomas paternos heredó el alelo G o de los dos maternos heredó el alelo T. Como observamos en la figura, el SNP-1 es informativo para el alelo paterno y el hijo es homocigoto para el alelo referencia, por tanto, si no ha ocurrido un efecto ADO, habrá heredado el alelo referencia de su padre. El SNP-5 también es informativo hacia el padre, pero en este caso estamos seguros de que el hijo no está bajo un efecto ADO (es un keySNP), ya que es heterocigoto. Esta posición permite al algoritmo establecer la fase haplotípica, considerando que tanto padre como hijo están afectos por la alteración, por lo que, como ambos son heterocigotos, el hijo debe haber heredado el alelo alternativo, y el cromosoma que porte dicho alelo, salvo que se produzca una evidencia de sobrecruzamiento, es el que porta la alteración causante de enfermedad. El SNP-2 es informativo hacia la madre y permite identificar que el hijo ha heredado el

cromosoma que porta el alelo alternativo T. Como la madre no es afecta y estamos ante una alteración dominante, podemos asegurar que dicho cromosoma será no portador. Finalmente, el SNP-3 está afectado por efecto ADO en el hijo, pues aparece como homocigoto para un alelo que no ha podido heredar de su padre. Tras el fasado, el software añade las posiciones no informativas y colorea cada cromosoma en función de su procedencia para facilitar la interpretación en un fichero tipo Excel (extensión .xls). También devuelve un fichero donde se registran las posiciones donde no ha habido coherencia de asignación. Muchas de estas posiciones serán debidas a fenómenos de ADO, aunque también pueden existir posiciones imposibles de interpretar debidas a fallos de secuenciación. También se genera un fichero que contiene las posiciones informativas en ambos parentales, un fichero con las posiciones informativas hacia el padre y otro con las posiciones hacia la madre, para facilitar la revisión del caso y selección de embriones. Estos ficheros son del tipo texto plano con formato tabular.

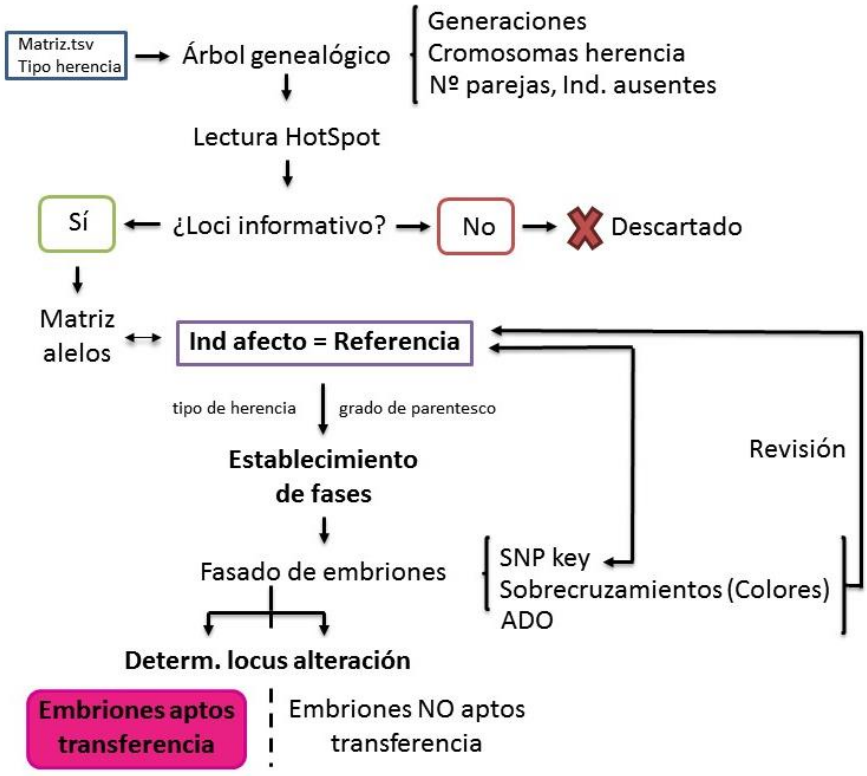

Figura 17: Desarrollo de la estrategia de fasado de polimorfismos.

En la Figura 17 se muestra el desarrollo de la estrategia de fasado paso a paso. A partir del archivo *HotSpot* se realiza el análisis y fasado de la pareja para determinar los polimorfismos informativos y, con ello, los embriones aptos para ser transferidos, siguiendo la misma técnica explicada previamente.

3.2.1 Validación mediante STR

Para validar la fiabilidad de las herramientas desarrolladas para DGP-M, se decidió realizar en paralelo un estudio clásico con STRs. Para ello se escogieron las muestras de ADN de 3 parejas que se habían sometido previamente a DGP-M, debido a que presentaron alguna alteración en el gen *PKD1* (chr16: 2136711-2187899) con herencia autosómica dominante, junto al ADN de las biopsias de los embriones obtenidos en el ciclo IVF y se aplicó el método aquí desarrollado en contraposición con el estado del arte basado en STRs. Se contabilizó el número de SNPs y STRs informativos y se empleó toda la información para fasar los cromosomas de los embriones, de manera que se pudieran seleccionar aquellos embriones libres de la alteración parental.

STR	Posición Chr16	Forward	Reverse
D16S521	94296-94420	GAGCGAGACTCCGTCTAAA	CAGCAGCCTCAGGGTT
D16S3399	145245-145427	ACCTAGATCCCTCCAGGTT	GGGCCATTATTCAGCCAATC
D16S3024	1654203-1654429	ACATGCTGTGCCACCT	AGCTGCCAGTATATGGAGGA
KG8	2138793-2138911	CACAGAAGTGGTACACAGAAGCAG	CAGGGTGGAGGAAGGTGAC
CW2	2457076-2457225	GTCCCTAGAAATAAGACCAAGTATGTG	CATTGCAGVAAGACTCCATCT
SM7	2567638-2567782	ATATGAAGAGGAATGGATGGGGTAG	CAAACAACAAGAGTGAATCTCTGAC

Tabla 2: Resumen de los STRs empleados en la validación

Los 6 STRs escogidos para realizar la validación se citan en la Tabla 2; el esquema de su localización respecto al gen PKD1 se muestra en la Figura 18. Se trata de STRs comunes muy utilizados en estudios de informatividad de PKD1 debido a su localización y a que la técnica de amplificación ha sido ampliamente depurada permitiendo obtener los mejores resultados[275,276].

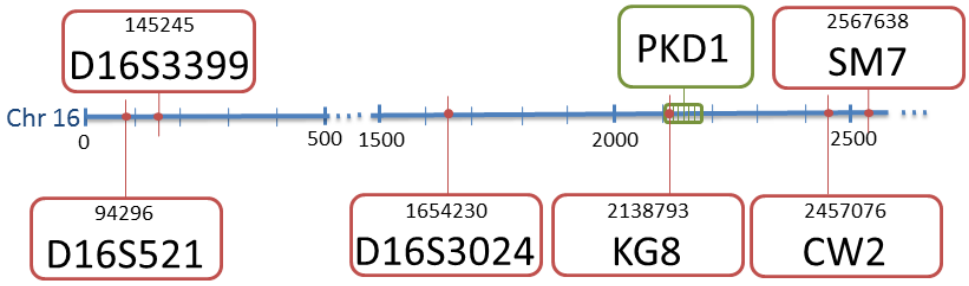

Figura 18: Esquema de localización de los STRs escogidos con respecto al gen PKD1. La distancia se mide en pares de bases.

3.2.2 Selección de muestras

La mujer de la pareja 1 presentaba la alteración *PKD1*:c1261C>T p.(Arg421Cys) en heterocigosis. A partir del ciclo de IVF se obtuvieron 12 embriones que pudieron ser biopsiados en día 3. Como no se disponía de muestra procedente de ningún familiar, se analizó directamente la mutación en todos los embriones. Se detectó la mutación en uno de ellos, que fue considerado referencia, fasando el resto de los embriones en función del resultado de éste.

El varón de la pareja 2 era portador de la alteración en heterocigosis *PKD1*:c6921_6922dup (pAla230Gly fs*7). También se dispuso de material perteneciente a la madre de dicho varón, portadora de la misma alteración. Ocho embriones fueron biopsiados en día 5.

El varón de la tercera pareja presentó una alteración en heterocigosis en *PKD1*:c2315T>C. También se dispuso de muestras de ADN de dos tías paternas, del hermano y del padre de dicho varón, todos ellos portadores en heterocigosis de la misma alteración. Del ciclo de IVF se biopsiaron siete embriones en día 3.

La siguiente tabla muestra un resumen de los tres casos escogidos indicando datos de los familiares recogidos en cada caso.

Pareja	Padre	Madre	Familiares	Embriones
1	No portador	c1261C>T	No	12; D3
2	c6921_6922dup	No portadora	Abuela Paterna Portadora Abuelo Paterno Sano	8; D5
3	c2315T>C	No portadora	TiaPat1 Portadora TiaPat2 Portadora Hermano Paterno Portador Abuelo Paterno Portador	7; D3

Tabla 3: Resumen de casos para la validación.

3.2.3 Procesado de muestras

El output de MiNtagSNP es reportado como un archivo tabular en formato BED y un archivo VCF. Los tagSNP seleccionados son señalados por filas indicando la posición cromosómica y la información relativa a los alelos referencia y alternativo, frecuencia alélica del alelo mayor y valor de MaxP. Posteriormente el archivo BED es utilizado por la plataforma de ThermoFisher para generar un panel personalizado AmpliSeq mediante el

uso del software Ion AmpliSeq comentado en 2.2.1 Estrategia del DGP-M por NGS del bloque 0 Introducción. La salida de la plataforma de Ion AmpliSeq consiste en un archivo BED tipo *HotSpot* con la información relativa a la posición cromosómica de los amplicones diseñados y a su identificador que permite restringir el análisis de todas las posiciones secuenciadas en el amplicón al análisis de las posiciones de interés (Las posiciones de los tagSNP y cstSNPs de interés), aunque todo el amplicón sea amplificado.

Una vez diseñado el panel se obtuvieron muestras de sangre en tubos EDTA de los pacientes sometidos al ciclo de IVF. El ADN fue extraído empleando el kit de purificación *Maxwell® 16 Blood DNA Purification*, siguiendo las instrucciones del fabricante. Por su parte, el ADN de las células biopsiadas de los embriones fue amplificado siguiendo el protocolo previamente descrito.

Para la preparación de las librerías, el ADN se amplificó con los amplicones diseñados. Tras la amplificación, estos amplicones fueron parcialmente digeridos mediante la adicción de FuPa y las muestras volvieron a ser incubadas. Los correspondientes códigos de barras moleculares fueron añadidos mediante ligación. Por último, la librería fue purificada empleando los kits *Agencourt® y AMPure®XP.*

La librería obtenida fue amplificada y se realizaron dos pasos de purificación del ADN, posteriormente se cuantificó con *Tape Station Agilent Technologies 2200*. Las librerías fueron clonadas empleando el *kit Ion PGM HiQ View OT2* (Life Technologies) en un sistema *Ion OneTouch 2* que emplea un sistema de amplificación sobre *Ion Sphere Particles (ISPs)* para realizar el enriquecimiento. Finalmente se realizó la secuenciación empleando la plataforma *Ion Torrent PGM* (LifeTechonologies).

El resultado del protocolo de laboratorio consistió en un set de archivos VCF con la llamada obtenida para cada SNP registrado en el archivo *HotSpot*, es decir, el valor de cada tagSNP y cada cstSNP en cada individuo. Genéricamente, de acuerdo a las instrucciones del fabricante, estas llamadas pueden tener el valor "Homocygous" cuando el SNP es homocigoto para el alelo alternativo; "Absent", cuando lo sea para el alelo referencia; "Heterocygous", cuando sea heterocigoto y "NoCall" si se ha experimentado un fallo de amplificación en dicha posición o la cobertura no es suficiente para realizar la determinación alélica con confianza.

Por otro lado, todas las muestras fueron procesadas siguiendo el protocolo facilitado por el fabricante para obtener los mejores resultados en técnicas de informatividad por STRs.

3.2.3.1 Implementación

La implementación en el laboratorio se realizó estudiando los datos recogidos en 51 parejas que se sometieron a análisis de DPG-M en nuestro laboratorio debido a distintas patologías. Los diferentes paneles de tagSNPs empleados fueron diseñados con MiNtagSNP atendiendo a la alteración presentada. Finalmente, se anotó el número de tagSNPs diseñados en el panel, el número de tagSNPs que pudieron ser secuenciados en ambos individuos de la pareja (y el porcentaje que estos representaron con respecto al total diseñado), el número de informativos hacia cada individuo entendido como el número de polimorfismos donde el individuo en estudio fue heterocigoto mientras que su pareja fue homocigoto y el número total de SNPs informativos del panel, así como el porcentaje que representó.

Por otro lado se decidió realizar una validación *in vitro* del poder de imputación sobre los polimorfismos diseñados en el panel de tagSNPs. Para ello, se escogieron como representativos 10 casos a partir de las 5 parejas donde se obtuvo el mayor porcentaje de SNPs secuenciados y las 5 parejas con el menor porcentaje. Los tagSNPs diseñados por MiNtagSNP fueron secuenciados y los resultados anotados para cada polimorfismo. Dado que BEAGLE exige que todos los individuos de la población a imputar (en este caso cada una de las parejas en cada uno de los 10 casos) presenten información para todas las posiciones indicadas como tagSNP, se decidió imputar las posiciones donde uno de los dos individuos no había sido correctamente secuenciado. Así, las posiciones de los tagSNPs donde ambos integrantes de la pareja fueron correctamente secuenciados fueron empleadas para imputar aquellas posiciones donde en uno de los dos integrantes se había producido un fallo de amplificación, empleando los individuos de la base de datos *1000GenomesDB* como población referencia. El valor de imputación obtenido para cada tagSNP en el individuo que sí había sido secuenciado se comparó con el valor imputado por BEAGLE a partir de los tagSNPs, permitiendo estimar la precisión de la imputación.

IV. Resultados y discusión

Capítulo 1: MiNFilterDups: Algoritmo específicamente diseñado para eliminar duplicados y artefactos de PCR en muestras de DGP-A

1.1 Set In sílico

La Tabla 4 recoge los valores de número de lecturas, MAPD y ploidía de cada embrión seleccionado para formar el set *in silico* de la validación (De B2 a CP7). Como se puede observar en la tabla, todas las muestras presentaron al menos 100.000 lecturas y un MAPD inferior a 0,3.

Embrión	MAPD	Lecturas	Cariotipo
B2	0,177	123168	45 XX, -3
T7	0,159	276470	46 XX, +18, -21
B10	0,183	161522	44 XY, -5, -19
B11	0,162	243812	48 XY, +13, +18
B12	0,168	153632	46 XX, +13, -8
B13	0,171	134481	46 XY, +21, -14
B14	0,122	165944	44 XX, +14, -1, -12, -22
B15	0,149	106961	46 XY, +1p, -1q
CP1	0,168	150776	46 XY
CP2	0,171	119791	46 XX
CP3	0,170	162725	46 XX
CP4	0,189	165178	46 XY
CP5	0,178	100962	46 XY
CP6	0,199	104398	46 XY
CP7	0,193	218502	46 XX
T1	0,210		46 XX, del(9p?)

Tabla 4: Resumen de las muestras seleccionadas en la validación de MiNFilterDups.

Resulta lógico asumir la idea de que el número de artefactos y duplicados de PCR experimente un crecimiento lineal con el número de lecturas de la muestra. Sin embargo, debido a la generación de nuevos "targets" en cada ciclo de amplificación este crecimiento se observa de manera exponencial, como se verá más adelante al estudiar el porcentaje de lecturas eliminado para cada muestra. Debemos recordar que este hecho fue tenido en cuenta a la hora de generar los archivos del set *in silico*, de forma que la proporción de lecturas de cada cromosoma en el BAM original se mantuviese en los archivos simulados. Finalmente, este set final fue procesado 3 veces de manera independiente:

a) Por un lado, los 495 archivos fueron filtrados empleando el plugin de filtrado de artefactos de PCR de ThermoFisher implementado en la versión 5.6 del IonReporter Reproseq™.

b) Por otro lado, los archivos fueron procesados de la misma forma pero empleando la versión 5.0, que no incluye flujo de trabajo para filtrado.

c) Por último, se empleó el algoritmo MiNFilterDups para generar el tercer set de archivos resultado.

Los archivos BAM resultantes de la simulación a partir de los archivos originales (n=495), el número de lecturas de cada archivo, así como los datos para el número de lecturas y MAPD obtenidos tras el filtrado con cada uno de los algoritmos se pueden consultar en la tabla adjunta en el repositorio GitHub a través de la dirección https://github.com/nacasfer/thesis_tables ; debemos destacar que, observando esta tabla, se puede apreciar una relación proporcional entre el número de lecturas y el valor de MAPD, de forma que a medida que el primero disminuye, aumenta el valor del segundo. En la tabla se indica el nombre del archivo BAM simulado, las lecturas de dicho BAM sin filtrar y las lecturas y el MAPD de cada archivo tras el filtrado con los distintos protocolos evaluados. El nombre de cada embrión simulado presenta el formato *LLLL_repX_EE*, siendo LLLL la categoría de número de lecturas de la muestra, repX la réplica (1ª, 2ª o 3ª) y EE el nombre del embrión original a partir del cual se obtuvo el archivo BAM. (Ej. 5000_re2_B2 corresponde al segundo archivo BAM simulado a partir del archivo del embrión B2 con aproximadamente 5000 lecturas).

1.2 MAPD y número de lecturas

Como el propio concepto indica, cuanto mayor sea el valor del MAPD para una muestra, mayor será la dispersión que presenten sus lecturas; este hecho podría enmascarar la verdadera ploidía del embrión al analizar el archivo BAM, pues el ruido producido por las lecturas no permitiría distinguir una aneuploidía de una dispersión natural. Por ello, con nuestro algoritmo MiNFilterDups hemos querido demostrar que, mediante un correcto filtrado de las lecturas del archivo BAM, es posible disminuir el valor de MAPD, reduciendo la dispersión de la muestra y aportando mayor credibilidad al diagnóstico del análisis DGP-A.

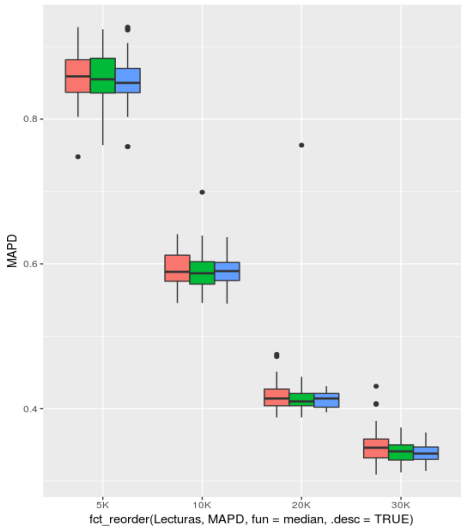

Figura 19: Diagrama de caja y bigote de la distribución del MAPD en muestras pertenecientes a categorías con bajo número de lecturas.

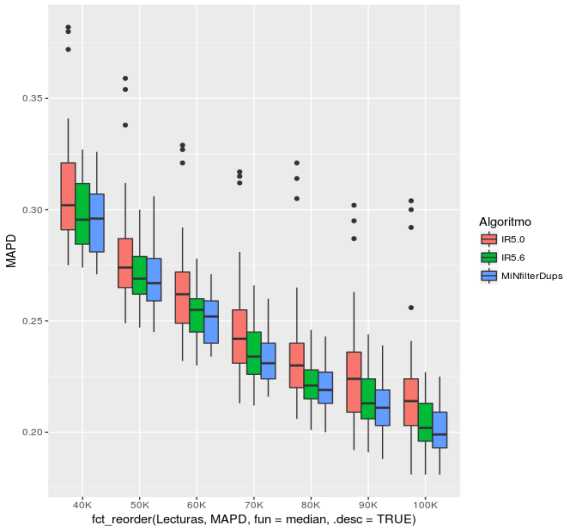

Figura 20: Diagrama de caja y bigote de la distribución del MAPD en muestras pertenecientes a categorías con alto número de lecturas.

IonReporter™ establece en 150.000 el número mínimo de lecturas necesario para realizar una correcta determinación de la ploidía, aunque valores experimentales previos no publicados establecen 100.000 lecturas como un límite más que suficiente; a partir de dicho valor, cuantas más lecturas, mejor será el nivel de sensibilidad. Como podemos observar en la tabla adjunta en el repositorio GitHub a través de la dirección https://github.com/nacasfer/thesis_tables, por debajo de 30.000 lecturas la sensibilidad disminuye drásticamente provocando un gran número de falsos negativos, por lo que será recomendable repetir la secuenciación[277].

En la Figura 19 y la Figura 20 se puede observar la relación existente entre el valor del MAPD y el número de lecturas del embrión antes de ser filtrado. La razón de separar ambas gráficas reside en facilitar el análisis por separado de las categorías que presentaron un valor de MAPD por debajo de 0,3 que, recordemos, es el valor marcado por IRS como umbral máximo para dar por significativo un diagnóstico de DGP-A ya que valores mayores generarían tal dispersión de las lecturas que el estado de ploidía podría quedar enmascarado..

En la Figura 19 se observa que apenas existió diferencia en los resultados obtenidos al filtrar los archivos con el plugin 5.6 de IRS o con MiNFilterDups. Por el contrario, en todos los casos mostrados en la Figura 20 los archivos filtrados con MiNFilterDups mostraron

valores más bajos de MAPD y número de lecturas que los valores arrojados al filtrar los mismos archivos con los otros algoritmos.

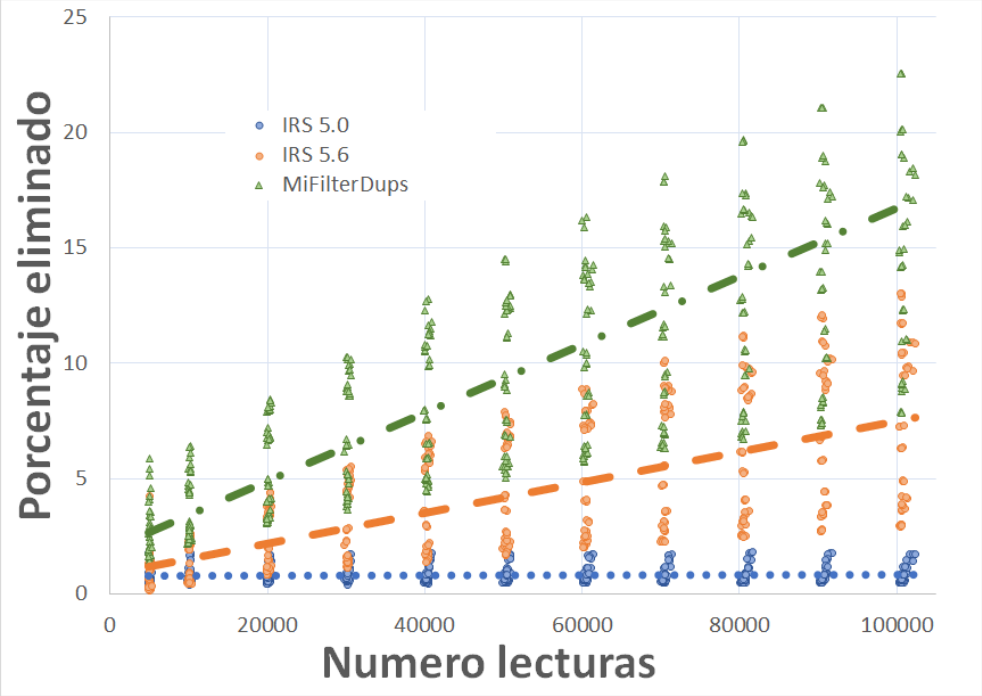

Figura 21: Promedio de lecturas filtradas por cada flujo de trabajo con respecto a las lecturas originales.

En segundo lugar quisimos demostrar que nuestro filtro es más restrictivo a la hora de detectar y eliminar artefactos de PCR. Se observó que el promedio de lecturas filtradas por los plugins IRS 5.0 y 5.6 fue del 0,81% y 4,23% respectivamente mientras que el porcentaje filtrado por MiNFilterDups ascendió al 9,5%. Esta diferencia se puede observar en la Figura 21. A mayor número de lecturas presentes en el embrión original, mayor es el porcentaje de lecturas eliminadas. Además, se observó que, a bajo número de lecturas originales, los plugins de IRS apenas presentaron diferencias en el número de lecturas eliminadas, mientras que MiNFilterDups siguió presentando porcentajes mayores. También resulta interesante señalar que MiNFilterDups mostró una varianza, en cuanto al porcentaje de lecturas eliminadas, de entre el 3,6 y el 22,5%, mientras que los valores del plugin de IRS 5.6 fueron entre el 2.9 y el 12%. Por su parte, el plugin IR5.0 siempre eliminó el 0,85%.

Por último, la Figura 22 se obtuvo al realizar una captura de pantalla al software IGV[278] y muestra una comparativa de las lecturas remanentes en dos regiones del genoma antes y después del procesamiento del mismo archivo BAM por los algoritmos de IRS y por MiNFilterDups. Estas dos regiones se han escogido para poder mostrar las diferencias de

filtrado de los algoritmos. En la región mostrada a la izquierda no se han producido duplicados; la región de la derecha presenta varios duplicados de PCR.

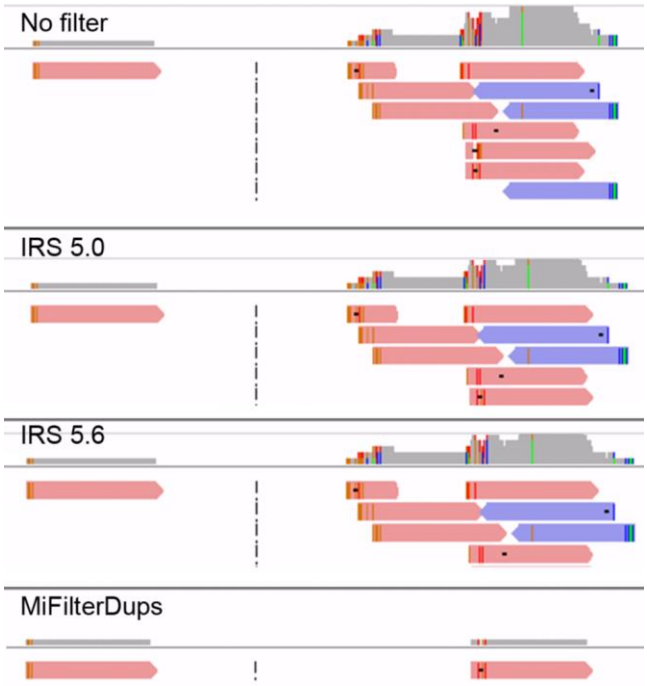

Figura 22: Comparativa de las lecturas filtradas para cada uno de los distintos flujos de trabajo evaluados frente a la situación sin filtrado.

La primera zona solo presentó una lectura, es decir, no presentó duplicados ni artefactos en esa región. Todos los algoritmos mantuvieron dicha lectura intacta. Por su parte, en la segunda región encontramos varios duplicados y artefactos procedentes de los distintos procesos de amplificación. Se observa claramente como, al contrario de lo que sucede con los duplicados de PCR más comunes, en este caso éstos presentan tamaños y puntos de origen y finalización diversos.

Se observó que, respecto al caso sin filtrar, mientras IRS 5.0 eliminaba las secuencias repetidas que presentaban exactamente la misma posición inicial y final , IRS 5.6 era capaz de eliminar más secuencias. Por su parte, MiNFilterDups filtró todas las secuencias, manteniendo la más larga.

1.3 Tiempo de ejecución

Se observó una relación lineal $\vartheta(n)$ entre el número de lecturas presentes en el archivo BAM a filtrar y el tiempo que emplea el algoritmo para procesarlo. La Figura 23 muestra que, a mayor número de lecturas, mayor fue el tiempo necesario para procesar el archivo.

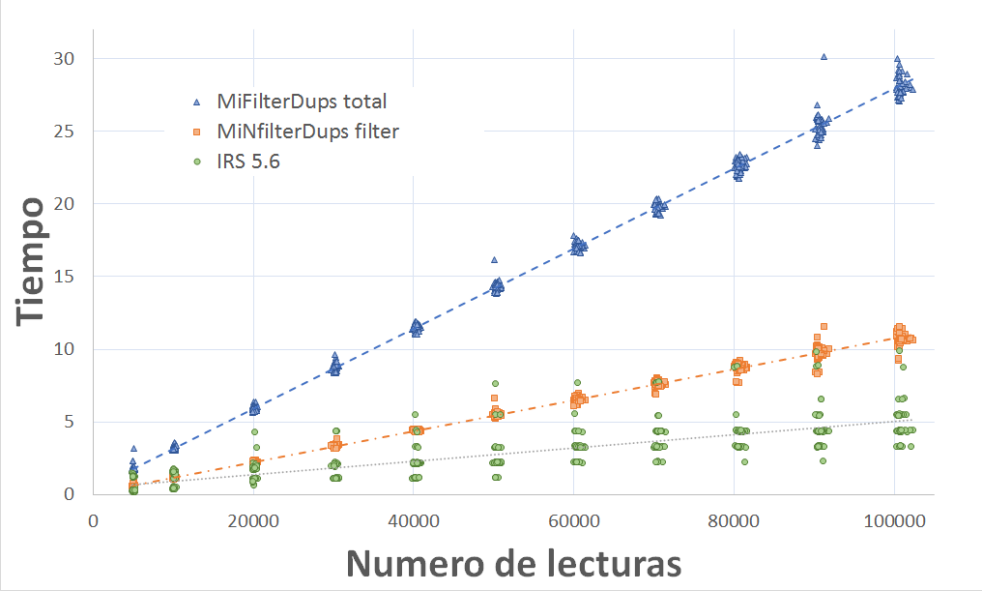

Figura 23: Representación del tiempo necesario para filtrar cada muestra con los distintos algoritmos. MiNfilterDups se representa dos veces, la primera en azul representa el tiempo total frente a una muestra cualquiera. La segunda, en verde, tiene en cuenta tan solo el proceso de filtrado de lecturas. La unidad temporal es el segundo.

Debemos destacar que el proceso completo del archivo usando MiNFilterDups supuso siempre un tiempo mayor respecto al tiempo empleado por IRS 5.6. Se observó también que la mayor parte del tiempo empleado por MiNFilterDups es empleado en ordenar las lecturas del archivo BAM, pues al descontar el tiempo empleado en este proceso, el tiempo total necesario para filtrar cada archivo descendió notablemente a pesar de continuar siendo ligeramente superior a los valores arrojados por IR5.6. Así, la muestra con mayor número de lecturas (perteneciente a la categoría 100.000 lecturas) fue procesada en un máximo de 29,3 segundos por MiNFilterDups, de los cuales solo 10,7 segundos fueron verdaderamente empleados en la fase de eliminación de las secuencias. Por su parte, IR5.6 empleó 9,9 segundos en filtrar el mismo archivo, lo que supone una diferencia insignificante en cuanto a tiempo de espera para obtención de resultados se refiere.

1.4 Embrión T1

El embrión T1 fue obtenido a partir de los embriones clasificados como "no aptos para transferencia" en una pareja que se sometió a un tratamiento de DGP-A. Diagnosticado en primera instancia como *46XX,del(9)(pter,p1.1)*, el embrión presentaba un perfil que sugería la posible tenencia de un cariotipo algo más complejo que el establecido por el workflow de ReproSeq. La Figura 24 muestra un pantallazo del software IGV implementado en el IonReporter (IRGV)[279] del archivo BAM filtrado con el IRS 5.6 (imagen superior) y el MiNFilterDups (imagen inferior).

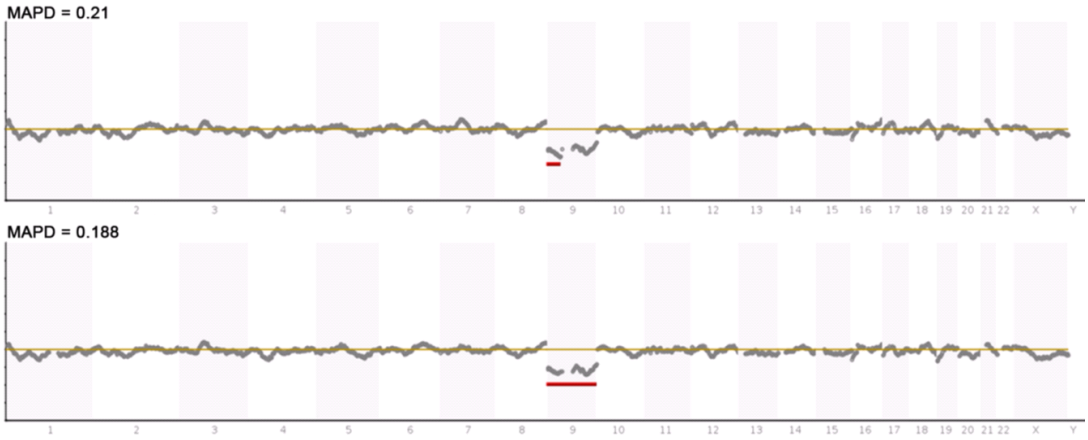

Figura 24: Vista en IGV del perfil del embrión T1. arriba filtrado con IRS 5.6, abajo con MiNFilterDups.

En la imagen superior de la Figura 24 observamos que el flujo de trabajo de IRS tan solo detectó una monosomía del brazo corto del cromosoma 9. La aneuploidía detectada es señalada con la línea roja. Sin embargo, el perfil de la dispersión de lecturas mostró claramente una alteración que afectaría a todo el cromosoma, pues las lecturas se agrupan por debajo de la marca de disomía (línea amarilla) en la que se agrupan las lecturas del resto de cromosomas. Este es un claro ejemplo de cómo la dispersión de las lecturas podría enmascarar la ploidía de la muestra para el algoritmo de DGP-A.

Si nos centramos en la imagen inferior, donde se muestra el análisis DGP-A realizado por el mismo algoritmo de IRS pero empleando MiNFilterDups como algoritmo de filtrado, observamos que todo el cromosoma 9 pudo ser detectado y diagnosticado como aneuploide.

Por otro lado, el valor de MAPD arrojado por el archivo BAM filtrado con el flujo de trabajo de IonReporter fue de 0,21, mientras que el filtrado con MiNFilterDups logró

disminuir el valor hasta 0,188, haciendo que el diagnóstico del análisis DGP-A más creíble debido a la disminución de la dispersión de la lecturas y a la mayor eliminación de ruido.

1.5 Discusión

Si nos centramos en la complejidad de las librerías génicas, lo ideal sería obtener un *pool* que refleje fielmente la complejidad del material original para poder analizarlo y obtener resultados fiables. Sin embargo, el proceso de amplificación necesario para poder analizar dicho material introduce desviaciones que se producen de manera heterogénea a lo largo de todo el genoma, dificultando el alcance de un diagnóstico final fiable con la realidad.

La estrategia de MiNFilterDups se justifica con base en dos hechos. En primer lugar, como ya se ha mencionado, el sistema IonReporter obtiene una media teórica μ de 150.000 lecturas, con una longitud media l de 110pb. Dado que el tamaño de ADN secuenciable en pares de bases es de $n = 289731462$, podrían formarse un total de $t = n - l + 1 = 2897310353$ lecturas de 110pb diferentes, que diferirían entre sí en un solo nucleótido. Teniendo esto en cuenta, la probabilidad de que dos lecturas sean exactamente iguales es $P_i = (1/t)^2 = 1{,}19 * 10^{-19}$.

Por otro lado, cada lectura tiene tantas lecturas solapantes no idénticas posibles como bases sea su tamaño menos una, de forma que el grado de solapamiento para cada lectura es $G_s = ((l-1)/t)^2 = 1{,}41/* 10^{-15}$.

Así, la probabilidad de que dos lecturas aparezcan representando la misma región es del orden de $P = P_i * (P_i + G_s) = 1{,}31 * 10^{-17}$. Esto quiere decir que, si dos lecturas aparecen solapantes dentro de un mismo archivo BAM, es muy probable que se hayan originado a partir del mismo producto de PCR, pues la probabilidad de que se hayan originado de forma independiente es despreciable.

Por su parte, la mayoría de los algoritmos de filtrado de artefactos suponen que la librería de ADN ha sido formada a partir de la fragmentación y posterior ligación de adaptadores que permiten la amplificación de todo el fragmento. En una situación ideal tan solo se secuenciaría una copia completa del fragmento, pero a menudo más de un fragmento de ADN se une a cada esfera magnética ISP para secuenciarse. Estos algoritmos aceptan por tanto que los duplicados de PCR son copias exactas del mismo fragmento de ADN

secuenciado de manera independiente y que se originan en la etapa de amplificación de la librería tras añadir los adaptadores. Sin embargo, la preparación de la librería para DGP-A es ligeramente diferente pues los fragmentos a secuenciar no se originan por fragmentación del material original sino por amplificación empleando cebadores semialeatorios que se unen al ADN y generan fragmentos a los cuales se unirán más tarde los adaptadores para construir la librería. A veces, estos cebadores pueden hibridar con un fragmento de ADN ya amplificado en lugar de con el material original de manera que se generan dos posibles fuentes de artefactos: el primero debido a la hibridación de los cebadores aleatorios en el ADN ya amplificado y el segundo debido a la amplificación de la biblioteca tras la ligación del adaptador.

En un principio observamos que existe una relación entre el número de lecturas y el valor de MAPD, de forma que al aumentar el primero disminuye el segundo. MiNFilterDups es capaz de disminuir el valor de MAPD al disminuir la dispersión de las lecturas gracias a la eliminación de los artefactos, lo cual aumenta la credibilidad del valor de confianza del diagnóstico de ploidía.

Como observamos en la Figura 20, cuando los archivos BAM a filtrar tienen menos de 30.000 lecturas, la diferencia en el valor de MAPD de los embriones filtrados con los diferentes algoritmos es inapreciable, el valor asciende por encima del valor umbral fijado por IRS y el diagnóstico deja de ser, supuestamente, fiable según el proveedor. Sin embargo, hemos visto que el diagnóstico sigue siendo correcto cuando los embriones son filtrados por MiNFilterDups. Esto nos indica que MiNFilterDups está realizando una mejor eliminación de los artefactos de PCR. En categorías con muchas lecturas la diferencia es más acusada, siendo siempre menor el MAPD de embriones filtrados con MiNFilterDups.

IRS emplea las coordenadas cromosómicas de alineamiento para determinar qué lecturas son duplicados (originadas a partir de la misma molécula de ADN). A diferencia de cómo lo hacen otros algoritmos también basados en las coordenadas, tiene una mayor flexibilidad en cuanto a los puntos de origen. Como ReproSeq™ es un sistema basado en amplificación por oligos semialeatorios, todas las lecturas pertenecientes a un mismo par de amplicones pueden no tener la misma posición de inicio y de final y por tanto además presentar diferentes tamaños. Este hecho hace que el enfoque de IRS no sea capaz de eliminar todos los artefactos producidos durante la fase de amplificación en la PCR. La presencia de estos artefactos provoca un aumento de la dispersión de lecturas, generando un ruido que se traduce en un aumento del valor de MAPD disminuyendo la credibilidad del diagnóstico de ploidía emitido, debido a que el ruido enmascara el modelo real.

En este estudio demostramos que los embriones filtrados con MiNFilterDups presentan siempre valores menores de MAPD y número de lecturas que los mismos archivos filtrados por los algoritmos de IRS; esta disminución de la relación procede de la disminución de la dispersión de las lecturas, lo cual provoca un aumento del valor de confianza del algoritmo HMM sobre el análisis DGP-A, haciendo más creíble el diagnóstico emitido.

Debemos destacar que los embriones pertenecientes a algunas categorías con bajo número de lecturas presentaron menores valores de MAPD al ser filtrados con IRS que con MiNFilterDups, esto puede deberse a que IRS eliminó menor cantidad de lecturas (menor cantidad de artefactos). Cuando una muestra presenta un elevado número de lecturas la dispersión se densifica, ya que las lecturas ocupan todo el espacio como si de una línea continua se tratase, por lo que es identificado como una distribución uniforme y el valor de MAPD disminuye. Por el contrario, un archivo BAM con bajo número de lecturas será observado como un conjunto de lecturas discretas y dispersas y el MAPD de la muestra ascenderá. Sin embargo, esto es un efecto secundario y trampa, pues no refleja realmente la dispersión real de la muestra y aporta un falsa sensación de seguridad que podría conducir a un mal diagnóstico en la ploidía del embrión, por haber quedado oculta tras el exceso/falta de lecturas y el ruido que éstas producen.

Por otro lado, como era de esperar, a mayor número de lecturas mayor es el porcentaje de artefactos eliminados, independientemente del filtro que se emplee. Observamos que en todos los casos MiNFilterDups superó el porcentaje de eliminación de IRS. El mayor porcentaje de lecturas eliminadas, unido al descenso del MAPD, hace que el diagnóstico emitido sobre los embriones filtrados con MiNFilterDups sea más fiable que el emitido sobe aquellos filtrados con los flujos de trabajo de IRS, incluso cuando estos embriones presentan, al final, menor número de lecturas.

Sorprendentemente, IRS 5.0 eliminó siempre un porcentaje constante de lecturas alrededor del 0,8%, independientemente del número de lecturas del archivo BAM filtrado. Este hecho puede deberse a que, en realidad, el algoritmo está descartando lecturas con baja calidad y lecturas multialineadas, llamadas *multimapping* (lecturas que alinean en más de una posición genómica), un efecto que se espera constante en todos los casos. Además, cabe destacar que en categorías bajas de número de lecturas, el protocolo 5.6 de IRS apenas presentó diferencia en el porcentaje de lecturas eliminadas con respecto a IRS 5.0, lo que implica que el algoritmo no es capaz de distinguir los duplicados presentes. Por su parte, MiNFilterDups continuó eliminando un buen porcentaje de artefactos.

Observando la dispersión del porcentaje de lecturas filtradas, observamos que embriones pertenecientes a la misma categoría de número de lecturas presentaron porcentajes de filtrado muy diferentes. Esto se debe a que la formación de artefactos es un proceso aleatorio que surge espontáneamente y, por tanto, es diferente para cada muestra. Por esta razón la proporción de lecturas de cada cromosoma ha sido respetada en los embriones simulados a partir de cada muestra original y, por ello, el porcentaje de filtrado final es diferente. MiNFilterDups presentó la mayor dispersión de este valor de filtrado, lo cual apoya la idea de que nuestro algoritmo es capaz de filtrar mucho mejor los artefactos de la muestra.

En cuanto al tiempo de ejecución, dado que $\vartheta(n)$ es una de las complejidades computacionales más eficientes, podemos considerar MiNFilterDups como una herramienta efectiva para el análisis DGP-A. Los embriones con 100.000 lecturas fueron filtrados en un máximo de 27 segundos en total, lo cual supone un tiempo razonable dentro del tiempo de manejo de cada muestra al considerar todo el proceso, pero excesivo en comparación con el tiempo empleado por IRS. Sin embargo, si tan solo contabilizamos el tiempo empleado en el filtrado de artefactos, los embriones fueron filtrados en un máximo de 10,6 segundos, lo cual resulta perfectamente asumible y está cerca de los 9,9 segundos del IRS.

Finalmente, el filtrado del embrión T1 permitió que el algoritmo basado en HMM detectase la monosomía completa del cromosoma 9, lo cual confirma que nuestro algoritmo supera la capacidad de los plugins de IRS para eliminar duplicados y artefactos. Además, MiNFilterDups disminuyó el valor de MAPD arrojado por el embrión con respecto al valor arrojado por el análisis del mismo archivo BAM con los plugins de IRS, lo que hace que el análisis DGP-A del archivo filtrado con nuestro algoritmo sea más fiable que el análisis realizado sobre el archivo filtrado con IRS.

A modo de resumen, dado que el filtrado de duplicados realizado por las técnicas del estado del arte no era suficientemente eficiente, MiNfilterDups es un algoritmo específicamente diseñado para muestras de DGP-A por técnicas de NGS.

El filtrado con MiNFilterDups permite disminuir el valor de MAPD y aumentar la confianza de los resultados de las muestras filtradas con respecto al estado del arte.

El filtrado de duplicados realizado por MiNFilterDups permite que muestras con menor número de lecturas muestren valores de MAPD inferiores a 0,3, mientras que el filtrado de dichas muestras con los algoritmos disponibles en el estado del arte muestra valores por encima de dicho umbral.

El filtrado con MiNFilterDups permite que el diagnóstico de muestras con valores de MAPD superiores a 0,3 aún sea fiable.

El filtrado con MiNFilterDups siempre elimina mayor número de lecturas que otros algoritmos del estado del arte ya que reconoce duplicados idénticos y secuencias formadas a partir de fragmentos previamente amplificados.

Capítulo 2: MiNmos: Algoritmo para la determinación de la ploidía y el nivel de mosaicismo en muestras de DGP a través de técnicas de NGS

2.1 Diseño

La Tabla 5 recoge los valores de número de lecturas, MAPD y ploidía de cada embrión seleccionado para formar el set *in silico* de la validación (De B2 a CP7). Como se puede observar en la tabla, todas las muestras presentaron al menos 100.000 lecturas y un MAPD inferior a 0,3.

Embrión	Biopsia	MAPD	Lecturas	Cariotipo
B2	5	0,177	123168	45 XX, -3
T7	5	0,159	276470	46 XX, +18, -21
B10	5	0,183	161522	44 XY, -5, -19
B11	5	0,162	243812	48 XY, +13, +18
B3	3	0,172	154685	46 XX, +13, -8
B13	5	0,171	134481	46 XY, +21, -14
B4	3	0,125	156234	44 XX, +14, -1, -12, -22
B5	3	0,151	125483	46 XY, +1p, -1q
CP1	5	0,168	150773	46 XY
CP2	5	0,171	119791	46 XX
CP3	5	0,170	162725	46 XX
CP4	5	0,189	194396	46 XY
CP5	5	0,178	165178	46 XY
CP6	5	0,199	99962	46 XY
CP7	5	0,193	104398	46 XX

Tabla 5: Muestras empleadas en la validación de MiNmos.

Tras la simulación de las muestras se recogieron los valores de MAPD, número de lecturas y ploidía emitido por el análisis de cada muestra. Estos datos quedan reflejados en la Tabla 2 y la Tabla3 adjuntas en el repositorio GitHub a través de la dirección https://github.com/nacasfer/thesis_tables.

2.2 Gradación del mosaicismo

Para estudiar si el valor Z-Score del logaritmo en base 10 de los niveles de cobertura corregidos, en cada embrión, con respecto a los valores de la línea base es un buen indicador del estado de ploidía y el nivel de mosaicismo a partir del porcentaje de aneuploidía, se representaron los valores de todos los cromosomas aneuploides presentes en los embriones analizados en función del nivel de mosaicismo simulado para esa muestra. Como

podemos observar en la Figura 25, se generó un gradiente de valores de Z- Score que indica la posibilidad de usar este valor continuo para gradar la categoría de mosaicismo.

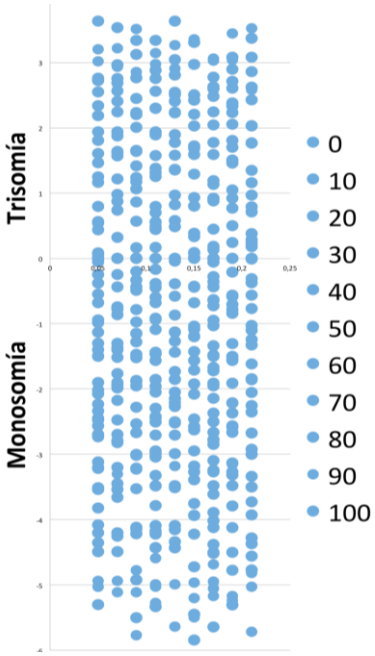

Figura 25: Distribución del Valor de Z-Score del log_{10} de los niveles de cobertura corregidos en cada embrión del set in silico.

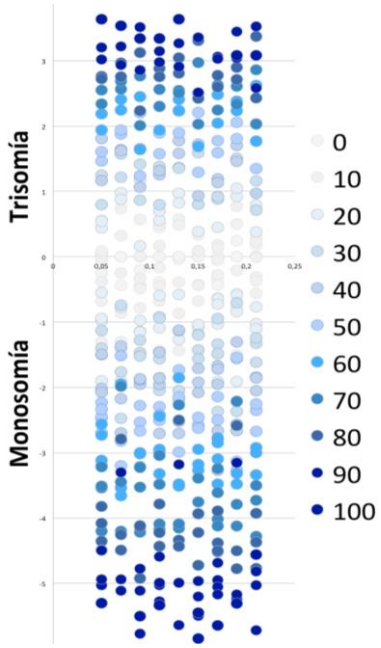

Figura 26: Gradación del Valor de Z-Score del log_{10} de los niveles de cobertura corregido en función de su categoría de mosaicismo.

La Tabla 25, Tabla 26 (adjuntas en la sección ANEXOS) y Tabla 7 muestran los valores determinados como punto de corte de cada categoría según la opción con la que se calcularon. Para determinar el punto de corte más adecuado se calculó el Ie y el YI, cuyo resultado se muestra en la Tabla 6. Observamos que el conjunto de puntos de corte con mayor Ie fue el compuesto por los puntos de corte determinados por el cálculo del tercer cuartil (el 77% de las muestras fueron correctamente clasificadas). Sin embargo, el mejor YI fue mostrado por el conjunto de puntos de corte promedio de los valores de Z-Score de una categoría y la siguiente (0,56 frente a 0,52 del conjunto de puntos marcado por el tercer cuartil). Por tanto, este último conjunto de puntos de corte fue considerado como el conjunto de puntos de corte óptimo para determinar las distintas categorías de mosaicismo. Además, observamos que el valor de Ie resultó perfectamente admisible, pues el 75% de muestras (frente al 77%) continúan siendo correctamente clasificadas dentro de su categoría.

	Índice exactitud	Índice Youden	Sensibilidad	Especificidad
Media -2dv	0,71	0,38	0,405±0,19	0,97±0,03
3er cuartil	0,77	0,52	0,79±0,17	0,74±0,2
Media Z-Score	0,75	0,56	0,84±0,14	0,72±0,17
Media Z-Score PGDIS	0,83	0,62	0,87±0,12	0,76±0,16

Tabla 6: Valor del índice de exactitud, índice de Youden, sensibilidad y especificidad para cada uno de los conjuntos de puntos de corte evaluados.

A partir de estos puntos de corte se determinaron 10 categorías del nivel de mosaicismo en embriones con trisomías y otras 10 en embriones con monosomías: Trisómico 100% , trisómico90%-diploide10%, trisómico80%-diploide20% … euploide 100%, diploide90%-monosómico10%, … diploide10%-monosómico 90% y monosómico 100%.

Como podemos ver en la Figura 26, en los valores de Z-Score que antes representábamos en la Figura 25 se genera ahora un gradiente que confirma que podemos usar este conjunto de puntos determinados por el promedio del valor de Z-Score de las categorías como valor de segregación de las muestras en función del nivel de mosaicismo.

Z-Score promedio		Valor unbral	Índice exactitud	Sensibilidad	Error_Se	IC_Se+	IC_Se-	Especificidad	Error_Esp	IC_Esp+	IC_Esp-
Trisom.	**100-90%**	2,968	0,095	1	0	1	1	0,133	0,172	0,305	-0,038
	90 - 80%	2,864	0,8	1	0	1	1	0,6	0,247	0,847	0,352
	80 - 70%	2,598	0,866	0,933	0,126	1,059	0,80709734	0,8	0,202	1,002	0,597
	70 - 60%	2,337	0,666	1	0	1	1	0,333	0,238	0,571	0,094
	60 - 50%	1,900	0,833	1	0	1	1	0,666	0,238	0,905	0,428
	50 - 40%	1,521	0,566	0,666	0,238	0,905	0,428	0,466	0,252	0,719	0,214
	40 - 30%	1,192	0,633	0,8	0,202	1,002	0,597	0,466	0,252	0,719	0,214
	30 - 20%	0,945	0,733	0,733	0,223	0,957	0,509	0,733	0,223	0,953	0,509
	20 - 10%	0,592	0,793	0,733	0,223	0,957	0,509	0,857	0,183	1,040	0,673
	10 - 0%	0,357	0,743	0,733	0,223	0,957	0,509	0,857	0,183	1,040	0,673
	Promedio		**0,673**	**0,86**	**0,123**	**0,983**	**0,736**	**0,591**	**0,219**	**0,810**	**0,673**
Monos.	**0- 10%**	-0,529	0,6388	0,428	0,211	0,640	0,216	0,933	0,126	1,059	0,807
	10 - 20%	-0,743	0,829	0,809	0,167	0,977	0,641	0,85	0,156	1,006	0,693
	20 - 30%	-1,1908	0,880	0,857	0,149	1,006	0,707	0,904	0,125	1,030	0,779
	30 - 40%	-1,6608	0,904	0,904	0,125	1,030	0,779	0,904	0,125	1,030	0,779
	40 - 50%	-2,1642	0,880	0,904	0,125	1,030	0,779	0,857	0,149	1,006	0,707
	50 - 60%	-2,7316	0,857	0,857	0,149	1,006	0,707	0,857	0,149	1,006	0,707
	60 - 70%	-3,3639	0,857	0,809	0,167	0,977	0,641	0,904	0,125	1,030	0,779
	70 - 80%	-4,0318	0,785	0,761	0,182	0,944	0,579	0,809	0,167	0,977	0,641
	80 - 90%	-4,7512	0,857	0,857	0,149	1,006	0,707	0,857	0,149	1,006	0,707
	90-100%	-5,1351	0,733	0,888	0,205	1,094	0,683	0,666	0,201	0,868	0,465
	Promedio		**0,817**	**0,861**	**0,141**	**1,003**	**0,720**	**0,767**	**0,182**	**0,949**	**0,584**

Tabla 7: Resumen de datos obtenidos al clasificar las muestras empleando los puntos de corte relativos al set c) Valor promedio de Z-Score. Valor Umbral: Valor del punto de corte entre categorías según el método estudiado; Error_Se: Tasa de error de la sensibilidad; IC_Se+: Valor superior del intervalo de confianza al 95% para la sensibilidad; Error_Esp: tasa de error de la especificidad; IC_Se-: Valor inferior del intervalo de confianza al 95% para la sensibilidad; Error_Esp: Tasa de error de la especificidad; IC_Esp+: Valor superior del intervalo de confianza al 95% para la especificidad; Error_esp: tasa de error de la especificidad; IC_esp-: Valor inferior del intervalo de confianza al 95% para la especificidad

2.3 Sensibilidad y especificidad

IonReporter™ establece que se obtiene un 100% de sensibilidad y un 64,29% de especificidad al analizar embriones 100% aneuploides[280] realizando el análisis con los valores por defecto.

La Tabla 25, Tabla 26 (adjuntas en la sección ANEXOS) y la Tabla 7 muestran los cálculos de sensibilidad y especificidad para cada categoría del nivel de mosaicismo en cada uno de los sets candidatos a puntos de corte al usar MiNmos. Los resultados quedaron resumidos en la Tabla 6. Se observó que, para el conjunto de puntos de corte procedentes de la media de los Z-Score entre categorías (Tabla 7), el promedio de sensibilidad y especificidad de la clasificación de muestras dentro de su categoría fue de 0,84 y 0,72 respectivamente. Por su parte, el I_e fue de 0,75 (El 75% de muestras quedaron correctamente clasificadas dentro de la categoría de mosaicismo para la que fueron simuladas). Finalmente se observó que, como muestra la Tabla 8, aparentemente, resulta más sencillo de detectar la monosomía que la trisomía, con valores para el I_e para el conjunto de puntos de corte escogido de 0,82 y 0,67 respectivamente.

Tipo mosaicismo	Puntos de corte	Índice exactitud	Índice Youden	Sensibilidad	Especificidad
Trisomía	Media -2dv	0,68	0,36	0,37±0,19	0,99±0,01
	3er cuartil	0,74	0,45	0,72±0,20	0,73±0,22
	Media Z-Score	0,67	0,45	0,86±0,12	0,59±0,21
	Media Z-Score PGDIS	0,76	0,49	0,86±0,11	0,63±0,21
Monosomía	Media -2dv	0,73	0,4	0,44±0,20	0,96±0,05
	3er cuartil	0,81	0,62	0,86±0,14	0,76±0,18
	Media Z-Score	0,82	0,66	0,81±0,16	0,85±0,14
	Media Z-Score PGDIS	0,88	0,75	0,87±0,13	0,88±0,12

Tabla 8: Valor del índice de exactitud, índice de Youden, sensibilidad y especificidad para cada uno de los sets de puntos de corte evaluados según el tipo de mosaicismo presente.

2.4 Reformulación según el PGDIS

Para el caso específico de la distinción entre embriones euploides y embriones con la categoría inferior de mosaicismo (hasta 10%), la sensibilidad fue de 0,73 para mosaicos con trisomía y 0,42 en monosomías. La especificidad fue de 0,85 y 0,93 respectivamente y el I_e de 0,74 y 0,63 (Tabla 8).

Además, observamos en la Tabla 7, que el 26% de muestras analizadas quedaban clasificadas fuera de su categoría al intentar distinguir si se trataba de un embrión euploide o un mosaico del 10% con trisomía (el 84% fueron correctamente clasificadas). Por su parte, el 37% lo hacían al tratar de distinguir si se trataba de un mosaico con monosomía al 10% (63% de acierto).

Con base en estos resultados y siguiendo las recomendaciones del PGDIS[281] se reformularon las categorías del nivel de mosaicismo (Tabla 9), agrupando las categorías con un 10 y un 20% de mosaicismo con la categoría euploide; de la misma forma se agruparon las categorías de aneuploidía con las categorías de un 90 y 80% de mosaicismo. De esta forma todo embrión con menos de un 20% de mosaicismo sería considerado euploide y con más del 80% sería tomado como aneuploide.

Tras la agrupación se observó que tanto la sensibilidad (0,87) como la especificidad (0,76), el valor del índice de exactitud (0,83) y el Índice de Youden (0,62) aumentaron con respecto al modelo general diseñado con todas las categorías (Tabla 6). De manera más específica (Tabla 9) se observó que los valores de sensibilidad y especificidad para los niveles de mosaicismo con trisomía aumentaron a 0,86 y 0,63 respectivamente, mientras que en niveles de monosomía aumentaron a 0,87 y 0,88. También se observó que, respectivamente, el 76% y el 88% de las muestras analizadas quedaron correctamente clasificadas dentro de su categoría.

En la distinción entre categoría euploide y mosaicismo del 30% con triploidía (Tabla 9), el 87% de las muestras fueron correctamente clasificadas, mientras que el 94% lo fueron al distinguir entre euploides y mosaicos al 30% con monosomía. Los valores de sensibilidad aumentaron a 0,74 y 0,88 respectivamente, mientras que los valores de especificidad fueron 0,91 y 0.81.

También se realizó una última prueba de reclasificación; en este caso las categorías de mosaicismo se superponían, de manera que la primera categoría de mosaicismo comprendería embriones entre 0 y 20% (euploides), la segunda entre 17 y 33%, la tercera

entre 27 y 43% y así sucesivamente. Observando la Tabla 7 y realizando los cálculos (no publicados) observamos que los valores de sensibilidad y especificidad ascendieron a 0,99.

La Figura 27, la Figura 28, y la Figura 29 muestran gráficamente los valores comentados en este apartado, para facilitar su comprensión y comparación.

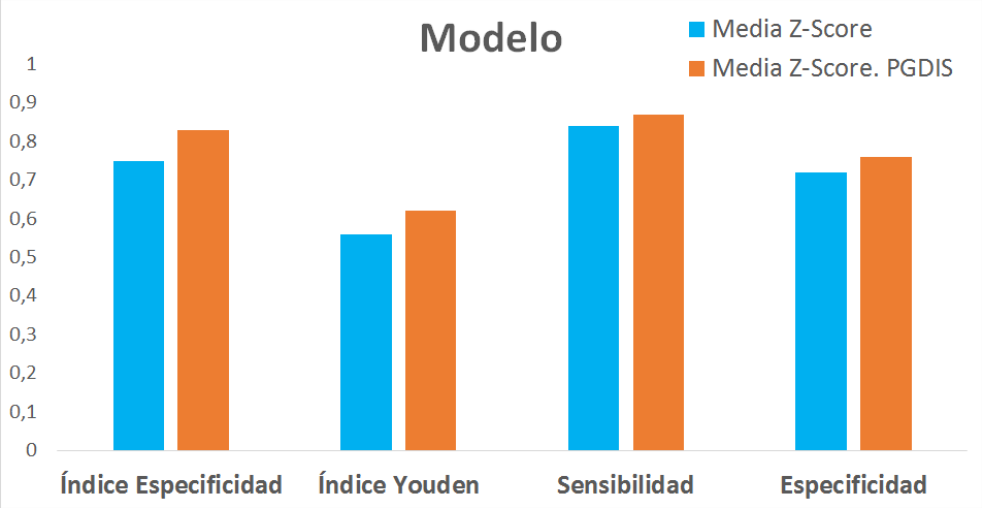

Figura 27: Resumen de los datos obtenidos para el Modelo Media de Z-Score y Media d Z-Score según las recomendaciones del PGDIS.

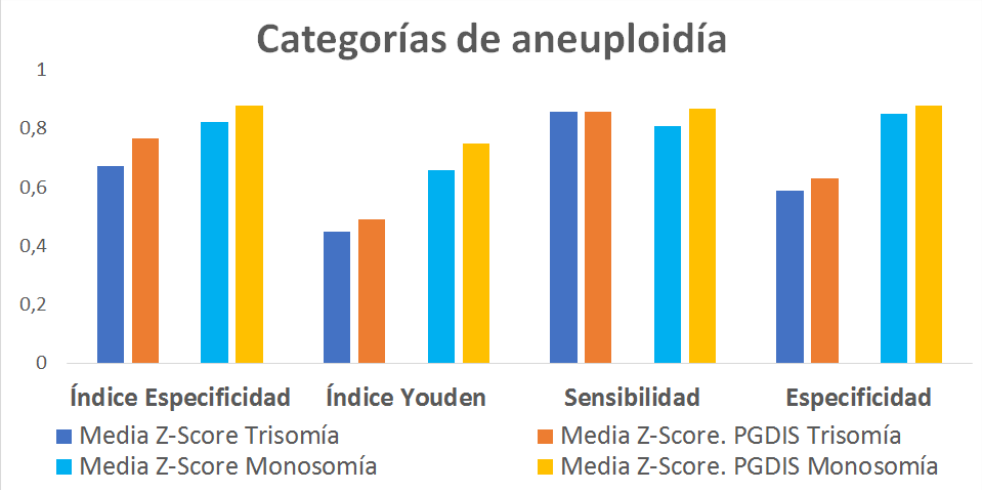

Figura 28: Resumen de los datos obtenidos para el Modelo Media de Z-Score y Media d Z-Score según las recomendaciones del PGDIS para cada tipo de aneuploidía.

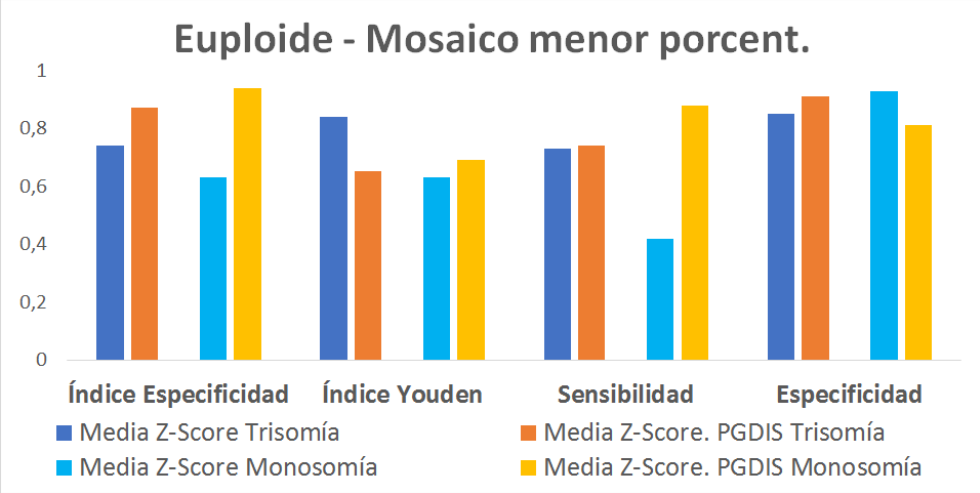

Figura 29: Resumen de los datos obtenidos para el Modelo Media de Z-Score y Media d Z-Score según las recomendaciones del PGDIS para a distinción entre embriones euploides y mosaicos del menor nivel.

Z-Score promedio PGDIS		Valor unbral	Índice exactitud	Sensibilidad	Error_Se	IC_Se+	IC_Se-	Especificidad	Error_Esp	IC_Esp+	IC_Esp-
Trisom.	100-70%	2,598	0,862	0,972	0,053	1,025	0,9187	0,6	0,247	0,847	0,352
	70 - 60%	2,337	0,766	1	0	1	1	0,533	0,252	0,785	0,280
	60 - 50%	1,900	0,9	1	0	1	1	0,8	0,202	1,002	0,597
	50 - 40%	1,521	0,566	0,666	0,238	0,905	0,428	0,466	0,252	0,719	0,214
	40 - 30%	1,192	0,633	0,8	0,202	1,002	0,597	0,466	0,252	0,719	0,214
	30 - 0%	0,945	0,866	0,733	0,223	0,957	0,509	0,911	0,083	0,994	0,827
	Promedio		**0,766**	**0,862**	**0,119**	**0,981**	**0,742**	**0,629**	**0,215**	**0,844**	**0,414**
Monos.	0 - 30%	-1,190	0,935	0,857	0,149	1,00	0,707	0,964	0,047	1,012	0,917
	30 - 40%	-1,660	0,904	0,904	0,125	1,030	0,779	0,904	0,125	1,030	0,779
	40 - 50%	-2,164	0,880	0,904	0,125	1,030	0,779	0,857	0,149	1,006	0,707
	50 - 60%	-2,731	0,857	0,867	0,149	1,006	0,707	0,857	0,149	1,006	0,707
	60 - 70%	-3,363	0,957	0,809	0,167	0,977	0,641	0,904	0,125	1,030	0,779
	70-100%	-4,031	0,861	0,882	0,088	0,970	0,793	0,809	0,167	0,977	0,641
	Promedio		**0,887**	**0,871**	**0,141**	**1,003**	**0,720**	**0,882**	**0,182**	**0,949**	**0,584**

Tabla 9: Resumen de datos obtenidos al clasificar las muestras empleando los puntos de corte relativos al set c) Valor promedio de Z-Score, pero considerando la clasificación establecida por el PGDIS. Valor Umbral: Valor del punto de corte entre categorías según el método estudiado; Error_Se: Tasa de error de la sensibilidad; IC_Se+: Valor superior del intervalo de confianza al 95% para la sensibilidad; Error_Esp: tasa de error de la especificidad; IC_Se-: Valor inferior del intervalo de confianza al 95% para la sensibilidad; Error_Esp: Tasa de error de la especificidad; IC_Esp+: Valor superior del intervalo de confianza al 95% para la especificidad; Error_esp: tasa de error de la especificidad; IC_esp-: Valor inferior del intervalo de confianza al 95% para la especificidad

2.5 Nivel de mosaicismo

El análisis del set in silico formado por embriones mosaico nos permitió establecer los valores mínimos de mosaicismo detectables por cada uno de los algoritmos sometidos a validación en este estudio (Tabla3 adjunta en el repositorio GitHub a través de la dirección https://github.com/nacasfer/thesis_tables). Como podemos observar en la Tabla 10, *ReproSeq low-pass Whole-genome aneuploidy* es capaz, en promedio, de detectar aneuploidía cuando hay presente un 70% de células aneuploides (desviación del 11%) siendo 60% de mosaicismo el caso de detección en el que se detectó el menor porcentaje de aneuploidía; por debajo de dicho porcentaje la aneuploidía queda oculta e indetectable y el embrión es catalogado como completamente euploide.

Por su parte, el protocolo *ReproSeq Mosaic PGS w1.1* detecta mosaicismo en todas las categorías, pero se observa un alto número de falsos positivos en muchos de los cromosomas de las muestras.

Por el contrario, observamos que el promedio de detección mínima de mosaicismo de MiNmos fue de 25,4% de aneuploidía. Recordemos que, según los criterios de PGDIS, embriones con porcentajes de aneuploidía inferiores al 20% son considerados euploides, por tanto nuestro algoritmo es capaz de detectar prácticamente todo el amplio espectro del nivel de mosaicismo. Además, a diferencia de IRS nuestro algoritmo en capaz de asignar dicho porcentaje de aneuploidía a una categoría del nivel de mosaicismo, permitiendo establecer una clasificación de las muestras en función del nivel de mosaicismo y no de la ploidía que presente mayor porcentaje en el embrión.

Embrión	ReproSeq Low-pass Whole-genome aneuploidy				MiNmos			
	Detección mosaicismo (%)		Núm. Mínimo lecturas		Detección mosaicismo (%)		Núm. Mínimo lecturas	
	Min. detectado	Categoría promedio	Min. detectado	Categoría promedio	Min. detectado	Categoría promedio	Min. detectado	Categoría promedio
B2_rep1	70		19796		10		10076	
B2_rep2	70	70	19836	20000	10	10	19926	10000
B2_rep3	70		20165		10		19990	
T7_rep1	70		20008		30		18630	
T7_rep2	90	70	20064	20000	30	30	18770	20000
T7_rep3	80		20030		30		18711	
B10_rep1	80		29954		30		20163	
B10_rep2	90	80	30000	30000	30	30	20031	20000
B10_rep3	80		30216		30		20277	
B11_rep1	80		19913		40		19016	
B11_rep2	90	80	19978	20000	40	40	19096	20000
B11_rep3	80		20095		40		19217	
B3_rep1	70		50023		40		20356	
B3_rep2	60	60	49744	50000	30	30	19933	20000
B3_rep3	60		50147		30		20187	
B13_rep1	50		20084		20		9786	
B13_rep2	60	50	20270	20000	20	20	9753	10000
B13_rep3	60		19365		20		9757	
B4_rep1	60		20242		20		18389	
B4_rep2	60	60	19723	20000	20	20	18655	20000
B4_rep3	60		29909		20		27429	
B5_rep1	60		10092		20		10000	
B5_rep2	60	60	19898	20000	20	20	10000	10000
B5_rep3	70		19878		20		20000	
Average	**70**		**24976,2**		**25,4**		**17422,8**	
Desv. est	11,42		10634,6		9,31		4757,8	

Continúa a partir de la tabla de la página anterior								
CP1_rep1			19907				9965	
CP1_rep2			19897	20000			19665	10000
CP1_rep3			20170				19682	
CP2_rep1			19940				19624	
CP2_rep2			20000	20000			19731	20000
CP2_rep3			20072				19806	
CP3_rep1			19888				9921	
CP3_rep2			20032	20000			9872	10000
CP3_rep3			30153				19806	
CP4_rep1			19863				18623	
CP4_rep2			20028	20000			18584	20000
CP4_rep3			30301				18763	
CP5_rep1			20243				19000	
CP5_rep2			20108	20000			18882	20000
CP5_rep3			19815				18653	
CP6_rep1			19986				19505	
CP6_rep2			20094	20000			19646	20000
CP6_rep3			19895				19435	
CP7_rep1			9993				9724	
CP7_rep2			20259	20000			9770	10000
CP7_rep3			19992				19594	
Average			**20506,4**				**17059,5**	
Desv. est			3900,1				4149,6	

Tabla 10: Categorías mínimas del set in silico correctamente detectadas por los distintos algoritmos.

2.6 Mínimo número de lecturas

Para establecer el número mínimo de lecturas necesario para establecer una correcta determinación de la ploidía del embrión, se analizaron las tres réplicas del segundo set *in silico*, así como el número mínimo de lecturas necesario para detectar las aneuploidías y su nivel de mosaicismo (Tabla 1 adjunta en el repositorio GitHub a través de la dirección https://github.com/nacasfer/thesis_tables).

La Tabla 10 muestra la categoría de mosaicismo con el menor porcentaje de aneuploidía correctamente detectado para cada embrión, así como el número mínimo de lecturas determinado para una correcta detección de la ploidía tanto por nuestro algoritmo como por los algoritmos de IRS.

Realizando un promedio de todas las categorías podemos concluir que el flujo de trabajo *ReproSeq Low-pass Whole-genome aneuploidy* realiza una correcta determinación de la ploidía con un promedio de, al menos, 24976 lecturas cuando analiza embriones aneuploides y de 20506 ante embriones euploides. El embrión que menor número de lecturas necesitó para ser correctamente clasificado perteneció a la categoría de 20.000 lecturas.

El workflow *Reproseq Mosaic w1.1* nuevamente detectó mosaicismo en todas las categorías, como muestra la Tabla1 adjunta en el repositorio GitHub a través de la dirección https://github.com/nacasfer/thesis_tables, pero volvió a detectar un número altísimo de falsos positivos, incluso en muestras biopsiadas en día 3 (Recordemos que en día 3 solo se biopsia una célula, por lo que la determinación de un diagnóstico de mosaicismo carece de sentido).

Por su parte, MiNmos fue capaz de detectar correctamente la categoría de mosaicismo con un promedio de 17422 lecturas en el caso de embriones con aneuploidías y de 17059 en embriones euploides. El embrión con menos lecturas correctamente detectado perteneció a la categoría de 10.000 lecturas

2.7 MAPD

La tabla 2 y la tabla 3 adjuntas en el repositorio GitHub a través de la dirección https://github.com/nacasfer/thesis_tables muestran las lecturas y valor de MAPD de cada

archivo BAM analizado. Podemos observar que, aun en los casos en que el valor de MAPD fue superior a 0,3, MiNmos realizó una correcta determinación no solo de la ploidía, sino del nivel de mosaicismo. El mayor valor de MAPD de un embrión correctamente analizado fue 0,637, lo que supone una gran mejora en cuanto a la fiabilidad del método, ya que IRS dejaba de serlo a partir de 0,3 (más de la mitad de la dispersión) debido a la cantidad de falsos positivos que presenta.

2.8 Discusión

Aunque el efecto del mosaicismo sobre las tasas de implantación es aún desconocido, es razonable asumir que la presencia de este fenómeno en los embriones transferidos disminuye las tasas de éxito de los ciclos de IVF[90].

Hasta ahora, la identificación de embriones mosaico era un proceso difícil y costoso que, en muchos casos, suponía la disgregación celular y, en aquellos casos en que se lograba, suponía la no transferencia debido al desconocimiento de las consecuencias sobre la descendencia[152]. Sin embargo, diversos autores han afirmado que la transferencia de embriones mosaico puede resultar en la implantación y nacimiento de un bebé sano[181]. Esto sugiere la existencia de algún mecanismo por el cual la aneuploidía es "corregida" en estos embriones mosaico y las células aneuploides son "relegadas" por las euploides; hecho que viene apoyado por las bajas tasas de mosaicos reportados en los test prenatales[153,155]. De esta manera, estos embriones podrían suponer una esperanza en parejas con tasas de fertilidad bajas de cuyos ciclos de fertilización no se hubiesen obtenido embriones euploides[152].

En la otra cara de la moneda nos encontramos con la ingente ambigüedad de consecuencias que podría suponer la transferencia de un embrión mosaico, pasando por fallos de implantación hasta el posible nacimiento de descendencia afecta[4] pues, aunque el mosaicismo está reconocido como un fenómeno prevalente en los ciclos de IVF, hasta ahora no es posible de determinar con fiabilidad el nivel al cual dichos embriones pasarían de ser irrelevantes a ser un problema[90,282].

Por estos motivos resulta esencial el establecimiento de una gradación de aneuploidía de los embriones mosaico que permita fijar y diagnosticar un valor umbral de porcentaje de células aneuploides presentes a partir del cual sea o no aconsejable la transferencia.

2.8.1 Diseño

Como ya hemos comentado, las células de trofoectodermo dan lugar a la placenta, por lo que, teóricamente, incluso una biopsia diagnosticada como anormal en este tejido podría resultar en un feto cromosómicamente normal; además, se cree que las células aneuploides presentan menores tasas de replicación y mayor tasa de apoptosis que las células euploides, debido a la mayor carga genética que portan, que provoca errores durante la división. Es por ello que la obtención posterior de muestra a partir del mismo embrión para su validación con el nivel de mosaicismo, podría arrojar resultados muy diferentes a los diagnosticados sin implicar, necesariamente, que el método de análisis estuviese errando durante su aplicación con el material original. El uso de simulaciones *in silico* a partir de archivos BAM secuenciados permite no solo generar todas las categorías de mosaicismo posibles a partir de lecturas de una biopsia procedente de un embrión humano real, algo que sería éticamente inviable en laboratorio, sino controlar todos los parámetros asegurando que el nivel de mosaicismo del archivo BAM final corresponda realmente con el porcentaje de aneuploidía esperado. Pero dichos archivos proceden de la biopsia de un embrión real.

La idea de que las condiciones de manejo en laboratorio pueden causar aneuploidías post-fertilización o aparición de embriones mosaico es un concepto muy provocativo pero ampliamente demostrado en un estudio realizado con 623 embriones pertenecientes a 7 clínicas distintas, el cual concluyó que un mismo embrión podía ser diagnosticado con un rango de mosaicismo entre el 32 y el 60% en función del laboratorio en el que se hubiese realizado la biopsia[283,284]. Para evitar este efecto, los archivos BAM seleccionados para realizar la validación de nuestro algoritmo proceden de embriones recogidos en distintos laboratorios.

Además, el uso del filtro de duplicados MiNFilterDups permite eliminar artefactos y duplicados de PCR, manteniendo tan solo la lectura original amplificada durante la reacción de PCR, disminuyendo la dispersión de las lecturas y permitiendo que la cuantificación de los cromosomas no esté afectada por las posibles desviaciones introducidas durante la elaboración de la técnica en los distintos laboratorios.

2.8.2 Gradación del mosaicismo

Teniendo todo esto en cuenta hemos desarrollado MiNmos, un algoritmo que permite no solo la detección de aneuploidías en embriones mosaico, sino el establecimiento del nivel de mosaicismo a través del porcentaje de células aneuploides presentes. Así, MiNmos realiza la determinación del porcentaje de aneuploidía presente a través de la gradación del valor *Z-Score de la transformación logarítmica en base 10 de las intensidades de cobertura corregidas* con base en los puntos de corte arrojados por el conjunto *Z-Score promedio de las categorías de mosaicismo* pues, aunque el Índice de exactitud de este conjunto de puntos de corte fue ligeramente inferior al I_e de otro de los conjuntos propuesto (0,75 frente a 0,77), el Índice de Youden mostró que este conjunto de valores umbral maximiza la diferencia entre verdaderos y falsos positivos, generando un consenso entre especificidad y sensibilidad que garantiza la mejor clasificación de las muestras dentro de la categoría de mosaicismo a la que pertenece, evitando así errores de sub y súper estima del nivel de mosaicismo y el porcentaje de aneuploidía presente.

Debemos además destacar el caso de las muestras B10 y B5, en inicio consideradas aneuploides para la monosomía de los cromosomas 5 y 19 en B10 y con una deleción/duplicación en los brazos del cromosoma 1 en el caso de B15.

A analizar B10 con MiNmos se observó que el cromosoma 19 exhibió un nivel de aneuploidía correspondiente a un nivel de mosaicismo del 60%. En el caso del brazo largo del cromosoma 1 de B15, observamos que este grado de mosaicismo fue del 90%. De ser cierto, ambos niveles entrarían dentro del nivel de detección de las técnicas actuales, por lo que habrían sido erróneamente tomados como aneuploides completos y, por tanto, incluidos en la validación de los métodos por medio de los cálculos de sensibilidad y especificidad. Es por ello que nos atrevemos a afirmar que dichas muestras son en realidad mosaicos no detectados por el estado del arte de las técnicas actuales, y su inclusión en la validación pudo provocar la subestimación de los valores de sensibilidad y especificidad.

2.8.3 Falsos positivos

A pesar de la existencia de estudios que afirman fervientemente el éxito de la transferencia de embriones mosaico[153,155], podemos encontrar en la bibliografía autores que contradicen estas afirmaciones alegando que dichos ratios de éxito son realmente debidos a que los embriones mosaico transferidos son en realidad falsos positivos de la técnica de

detección empleada y, por lo tanto, embriones euploides con un perfil lleno de ruido que los hace ser erróneamente clasificados como mosaicos[158]. Es precisamente debido a estas contradicciones que se hace imperiosa la necesidad de desarrollar un método preciso de determinación del porcentaje de aneuploidía presente que controle los niveles de sensibilidad y especificidad de la técnica y permita establecer un valor umbral (si existe y es realmente posible de determinar) a partir del cual un embrión mosaico es viable para la transferencia.

Como ya comentamos, IonReporter™ establece que, realizando el análisis con los valores por defecto, se obtiene un 100% de sensibilidad y un 64,29% de especificidad al analizar embriones 100% aneuploides[280], pero no se indica nada de embriones mosaicos. En el caso de MiNmos, las tasas ascendieron al 100% al realizar la misma comparativa entre embriones euploides y aneuploides. Además, se estableció una tasa global de especificidad del 72%, mientras la sensibilidad ascendió al 84% al distinguir entre embriones euploides o embriones con un máximo de aneuploidía del 10%. Las tasas ascendieron al 100% tanto para sensibilidad como para especificidad al distinguir entre embriones euploides. Centrándonos en los tipos de aneuploidía, observamos que los valores de sensibilidad y especificidad fueron de 86 y 59% al distinguir embriones mosaico con trisomías al 10% frente a embriones euploides, y del 81 y 85% cuando los mosaicos fueron de monosomías.

Al realizar la reclasificación con base en los criterios del PGDIS, que considera que cualquier embrión con un máximo del 20% de aneuploidía puede ser considerado euploide[281], los valores de sensibilidad y especificidad global ascendieron tanto globalmente como en la distinción entre las categorías inferior de mosaicismo y la categoría de euploides, algo que resulta realmente interesante, pues siempre debe priorizar la transferencia de embriones euploides sobre los mosaicos[281].

Cabe destacar que el uso de esta reclasificación no solo se fundamenta en los consejos publicados por el PGDIS. Si consideramos que una biopsia de blastocisto contiene entre 4 y 10 células (se desaconseja una biopsia con mayor número de células debido a que podría perturbar el normodesarrollo del embrión[285]) y que, por lo general, se denomina mosaico a aquellos embriones que presentan al menos 2 células aneuploides[286], nos encontramos que el caso con menor porcentaje de aneuploidía sería un mosaico con un 20% de eneuploidía (2 células aneuploides en 10) y nunca uno del 10% (1 célula aneuploide en 10). Según esto, establecer este valor umbral de determinación resulta lógico y admisible, permitiendo el incremento de los valores de sensibilidad y especificidad de la técnica.

La última prueba de reclasificación realizada nos sirve para darnos cuenta del poder de diagnóstico de MiNmos. A diferencia de otros algoritmos, que no son capaces de determinar el porcentaje de mosaicismo, MiNmos no solo es capaz de detectar la aneuploidía en embriones mosaico, sino que es capaz de determinar el porcentaje de aneuploidía presente y con ello el nivel de mosaicismo. Así, las tasas de sensibilidad y especificidad tomadas en la Tabla 7 y la Tabla 9 de este estudio refieren a la catalogación exacta del embrión dentro de la categoría para la cual fue simulado. Dada la existencia de embriones mosaico y su creciente importancia en el campo del DGP, la simple catalogación en embriones euploides y aneuploides (transferibles o no) se queda corta en términos de análisis de resultados y explicaría la alta tasa de sensibilidad mostrada por IRS. Por tanto, para el análisis de validación de MiNmos se consideró fallo cuando el embrión fue catalogado en la categoría superior o inferior, lo cual no significa en ningún caso que el algoritmo MiNmos señale el embrión como euploide en caso de ser aneuploide (o viceversa), sino que puede sobre estimar o subestimar, con un error del 3-5% el porcentaje de mosaicismo presente, hecho que confirman las tasas de sensibilidad y especificidad obtenidas al considerar categorías solapantes.

2.8.4 MAPD, nivel de mosaicismo y número mínimo de lecturas,

Se observó que el algoritmo *ReproSeq Low-pass Whole-genome aneuploidy* es capaz de detectar en promedio embriones mosaico con un 70% de aneuploidía (el mínimo detectado fue de 50%), pero no es capaz de determinar dicho porcentaje ni gradarlos en categorías de mosaicismo. Por debajo de estos niveles de mosaicismo, la aneuploidía queda oculta e indetectable y el embrión es clasificado como euploide, lo que lo convierte en un embrión potencialmente transferible que puede generar diversos efectos en función del grado de aneuploidía y el tipo celular afectado.

Por su parte, *ReproSeq Mosaic w1.1* es capaz de detectar aneuploidía en todos los niveles de mosaicismo pero también presenta tal número de falsos positivos que todos los embriones parecen mosaico. Este hecho quedó reflejado en el análisis de los embriones del set *in silico* de lecturas mínimas biopsiados en día 3. Debemos recordar que las biopsias en día 3 capturan tan solo una célula y que este set se realizó extrayendo lecturas aleatoriamente a partir de los archivos BAM, por lo que la detección de niveles de mosaicismo o porcentajes de aneuploidía en el análisis de estos archivos carece de sentido.

Además *ReproSeq Mosaic w1.1* tampoco es capaz de gradar el nivel de mosaicismo a través del porcentaje de aneuploidía.

MiNmos fue capaz de detectar aneuploidía en embriones mosaico al 25,4% (el mínimo detectado fue del 10%), lo cual resulta un éxito si tenemos en cuenta las recomendaciones del PGDIS. Además, MiNmos fue capaz de detectar de forma fiable la aneuploidía en embriones con al menos 17506 lecturas, mientras que, aunque se observó acierto hasta con un mínimo de 24976 lecturas al emplear el workflow *ReproSeq Low-pass Whole-genome aneuploidy*, el proveedor asegura que son necesarias al menos 100.000 lecturas para obtener un diagnóstico fiable, pues un menor número de lecturas provoca un aumento del valor de MAPD, con el consecuente aumento de la dispersión de las lecturas y el ocultamiento del verdadero modelo de ploidía del embrión.

Finalmente, respecto al valor de MAPD, se observó que aún cuando este valor fue muy superior al umbral de 0,3 impuesto por IRS, nuestro algoritmo continuó realizando una correcta determinación de la ploidía y el nivel de mosaicismo del embrión, lo cual supone una gran mejoría en cuanto a la fiabilidad del método (IRS deja de ser fiable a partir de 0,3 mientras que MiNmos parece no ser susceptible a dicho valor) y su utilidad pues es difícil conseguir que los embriones sean secuenciados con más de 100.000 lecturas. Estos avances suponen un ahorro en tiempo y dinero pues a menor número de lecturas necesario, mayor cantidad de muestras podrían secuenciarse de una sola vez.

Así, a diferencia del protocolo *ReproSeq Low-pass Whole-genome aneuploidy* de IRS, MiNmos es capaz de detectar el espectro completo de porcentaje de aneuploidía presente en un embrión a la par que, a diferencia del protocolo *ReproSeq mosaic w1.1,* es capaz de gradarlo y diagnosticarlo dentro de las categorías de mosaicismo. Además permite clasificar si un embrión es euploide, aneuploide o mosaico. Por todo ello, MiNmos ha demostrado ser más preciso que los algoritmos empleados por el software de IR a la par que arroja una posibilidad de intentar comprender los efectos del mosaicismo sobre los ciclos de IVF y establecer (si existe), en un futuro, un valor umbral a partir del cual la transferencia de embriones mosaico pueda ser o no recomendable.

2.8.5 Limitaciones

La bibliografía recoge que las técnicas de detección de aneuploidías por NGS fallan en la detección de aneuploidías recíprocas (embriones que poseen células trisómicas y

células monosómicas para un mismo cromosoma) cuando se encuentran en una proporción cercana al 50%[287], pues debido a la técnica de amplificación y la detección realizada por los métodos de análisis, el suceso de trisomía quedaría enmascarado (compensado) por el de monosomía. Por este motivo y previo a la transferencia de embriones, debería realizarse siempre un proceso de consejo genético advirtiendo de la posibilidad de transferir un embrión afecto de trisomías ocultas y del riesgo que puede conllevar para la descendencia transferir un embrión mosaico con monosomía (como indican los cálculos presentados en este estudio, la monosomía es un suceso más sencillo de detectar). Debemos recordar que ciertas trisomías son viables y pueden desembocar en la tenencia final de un bebé afecto; por el contrario la transferencia de embriones mosaico para otras trisomías y/o monosomías podría desembocar en abortos espontáneos y fallos de implantación.

En relación con este fenómeno y debido a la creencia de que la mayor parte de mosaicismos derivan de la no-disyunción mitótica[288,289], *Mertzanidous et al*. realizaron diversos estudios donde esperaban que al menos uno de los embriones, analizados por disgregación y análisis de células únicas, presentase aneuploidías recíprocas para al menos un cromosoma[152,290–292]. Sin embargo ningún embrión analizado cumplió las expectativas, lo que nos permite afirmar que, si bien la limitación de detección existe, la tasa de aparición de este fenómeno es tan baja que podría considerarse despreciable.

Por último, debemos recordar que MiNmos permite la detección del porcentaje de aneuploidía y gradación del porcentaje de mosaicismo en células de trofoectodermo biopsiadas, pero en ningún momento pretende establecer el porcentaje de aneuploidía o nivel de mosaicismo a partir del cual la transferencia de un embrión mosaico resultaría o no exitosa. Para establecer dicho porcentaje deberíamos recurrir a una guía de transferencia y continuar realizando estudios intensivos sobre el tema mediante la aplicación de este nuevo algoritmo de análisis en clínica.

A modo de resumen, podemos afirmar que MiNmos es un algoritmo específicamente diseñado para la detección de bajos porcentajes de aneuploidía y determinación del nivel de mosaicismo de muestras de DGP-A por técnicas de NGS.

MiNmos es el primer algoritmo para DGP-A capaz de gradar y determinar el nivel de mosaicismo de la muestra, indicando el porcentaje de aneuploidía presente.

El *valor Z-Score del log10 de los niveles de cobertura corregidos respecto a los valores de las dos líneas base* analizado por MiNmos es un buen indicador del estado de ploidía y el nivel de mosaicismo.

MiNmos permite la detección del estado de ploidía y el nivel de mosaicismo incluso con valores de MAPD por encima de 0,3.

MiNmos presenta una mayor sensibilidad y especificidad en la distinción entre embriones euploides y mosaicos de bajo nivel respecto a las técnicas actuales del estado del arte.

MiNmos necesita un menor número de lecturas mínimas para obtener resultados fiables respecto a los algoritmos del estado del arte. Esto supone un ahorro en tiempo y dinero, pues a menor número de lecturas necesarias para un correcto diagnóstico, mayor cantidad de muestras pueden ser secuenciadas a la vez.

Capítulo 3: Dispersión de las lecturas

3.1 Estadísticos propuestos

La Figura 30 muestra que los estadísticos *Rango, Rango medio, Rango intercuartílico, Coef. de variación, Varianza y Desviación típica* presentaron perfiles muy similares al perfil arrojado por la representación del valor de MAPD de las muestras, mostrado en la Figura 19 y la Figura 20 de la sección 1.2. MAPD y número de lecturas. Las muestras pertenecientes a categorías con alto número de lecturas presentaron siempre valores inferiores de la dispersión con respecto a las muestras de categorías con bajo número de lecturas. Esto sugiere que, como venimos comentando, la dispersión de las lecturas es mayor en muestras con bajo número de lecturas.

Por el contrario, los perfiles de los estadísticos basados en Z-Score arrojaron conclusiones muy diferentes. Todas las muestras se agruparon en torno a un valor umbral común independientemente del número de lecturas que presentasen, lo que sugiere que la dispersión de las lecturas no está relacionada con el número de lecturas de la muestra.

La Figura 31a muestra el valor del estadístico Z-Score. Se observó que todas las muestras se agruparon en torno al valor $\approx -3{,}85 * 10^{-18}$, inclusive aquellas pertenecientes a muestras que no habían sido procesadas mediante el protocolo ReproSeq 5.0, el cual no incluye una fase de filtrado de artefactos. Esto es debido a que el valor del Z-Score admite valores positivos y negativos, lo que contrarresta el efecto de la dispersión entre distintas muestras y centra el valor en el promedio.

Por su parte, la Figura 31b muestra el valor absoluto de dicho Z-Score. En este caso se observó que los valores procedentes de muestras procesadas mediante el protocolo ReproSeq 5.0 si presentaron diferencias con respecto a las muestras procesadas por los algoritmos incluidos en los otros dos protocolos, aunque todas se agruparon en torno al valor 0,78.

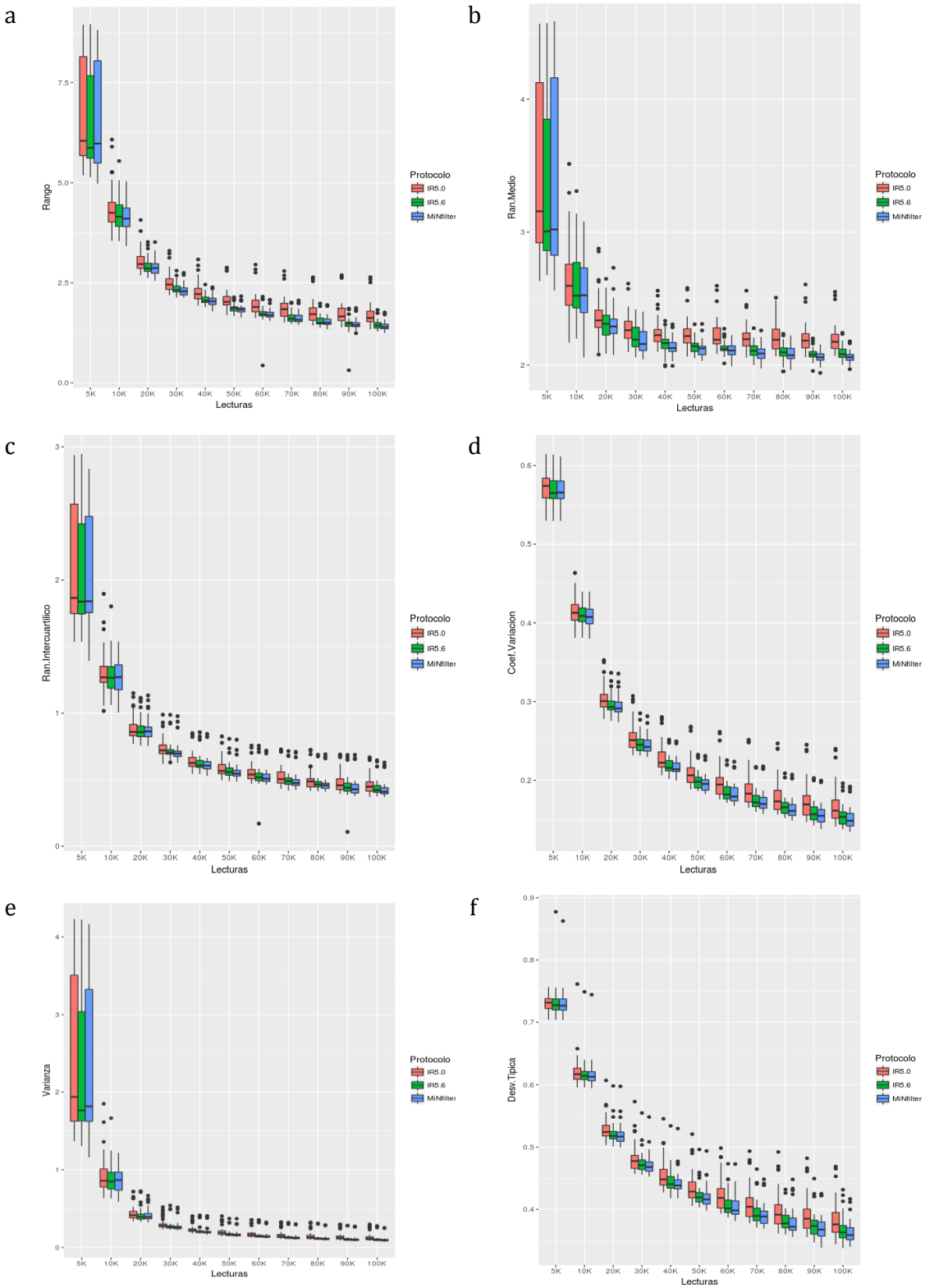

Figura 30: Gráfico de caja y bigotes del valor de a)Rango, b)Rango Medio, c)Rango Intercuartílico, d)Coeficiente de Variación, e)Varianza, f)Desviación típica de la dispersión de las lecturas en las distintas categorías muestrales del número de lecturas.

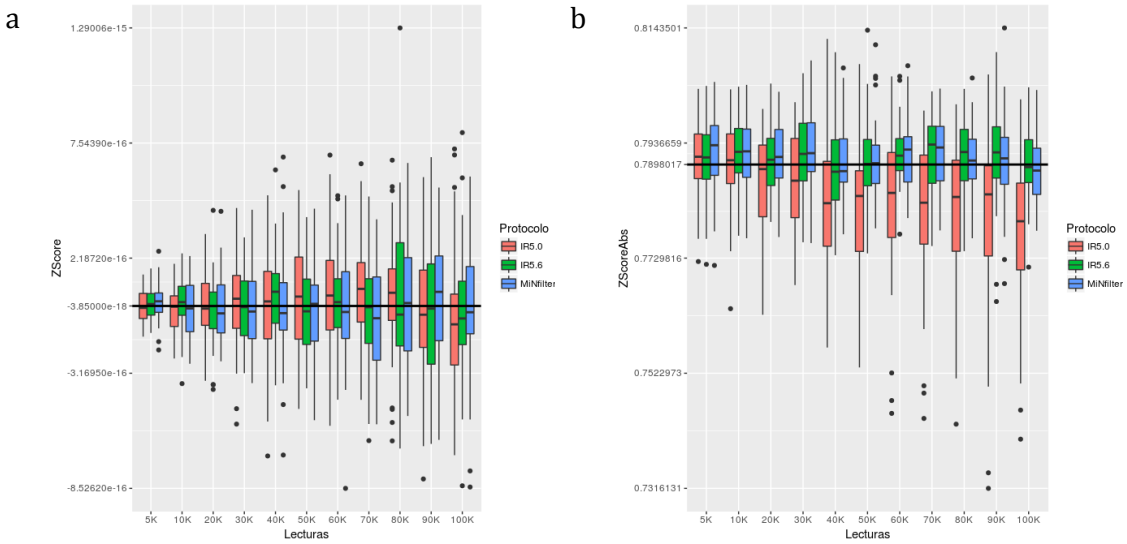

Figura 31: Gráfico de caja y bigotes del valor de a)Z-Score, b)Z-Score absoluto de la dispersión de las lecturas en las distintas categorías muestrales del número de lecturas.

Finalmente, la Figura 32 muestra este segundo tipo de perfil, pero el valor de *Z-Score* ha sido corregido gracias al reescalado introducido por el logaritmo negativo y la consideración del valor absoluto del mismo. Esta gráfica muestra cómo los valores se agrupan en torno al valor ≈0,28, y maximiza las diferencias entre las distintas categorías y protocolos permitiendo comparar la dispersión de las lecturas de los embriones con independencia del número de lecturas que presente la muestra.

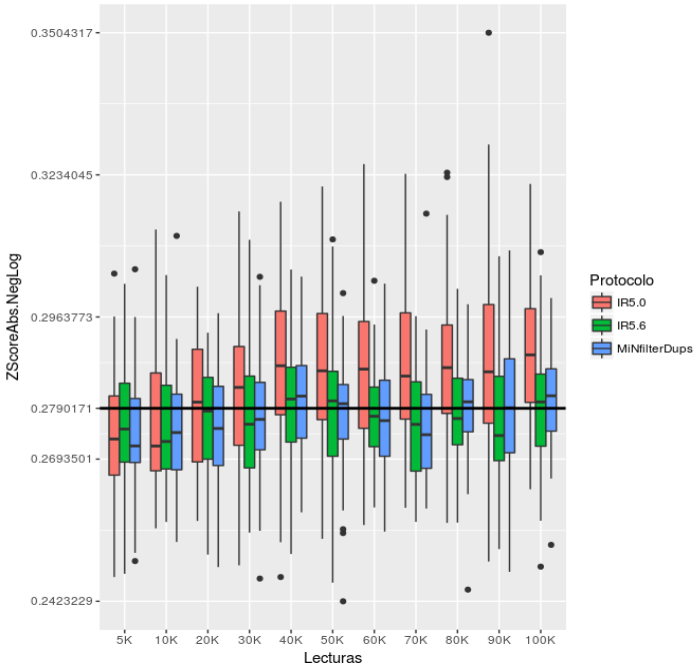

Figura 32: Gráfico de caja y bigotes del valor de logaritmo negativo del valor de Z-Score absoluto de la dispersión de las lecturas en las distintas categorías muestrales del número de lecturas.

3.2 Discusión

El MAPD es un valor inversamente relacionado con el número de lecturas. Cuando los embriones tienen menos de 40.000 lecturas el MAPD asciende por encima del valor umbral de 0,3 fijado por IRS como máximo permitido para una correcta detección de aneuploidías, y el diagnóstico deja de ser fiable. Sin embargo, cuando las lecturas son filtradas usando MiNFilterDups, este valor se alcanza con un número menor de lecturas, es decir, embriones con un número de lecturas inferior a 40.000 todavía presentan valores de MAPD por debajo del valor límite de 0.3.

De esta forma, un menor número de muestras son descartadas debido a un MAPD alto. Esto es especialmente relevante en el diagnóstico genético preimplantacional debido a la dificultad que representa el hecho de tener que repetir un análisis. Cuando un análisis es considerado como no válido o no concluyente por el motivo que sea, una opción para lograr un resultado consiste en la repetición de la biopsia para obtener más muestra, sin embargo esto no es muy recomendable debido al daño que se le provoca al embrión, no solo por la obtención de la biopsia en sí, sino también por el proceso de vitrificación y desvitrificación al que debe ser sometido. Una segunda opción, puesto que la re-amplificación no es posible, consiste en la re-secuenciación de la muestra a partir del material ya biopsiado. Si bien esta opción se antoja como la mejor a fin de garantizar la viabilidad del embrión, supone ciertos problemas técnicos debido al alargamiento del tiempo necesario hasta la obtención de resultados (un factor crítico en la mayor parte de los casos sometidos a DGP, especialmente en los ciclos 24horas, donde se biopsia en día 3 o día 5 y se pretende transferir en día 5 o 6 respectivamente, para realizar todo el proceso dentro del mismo ciclo hormonal de la madre) y al sobrecoste del proceso. Por ello resulta interesante el desarrollo de metodologías de análisis que permitan realizar un "rescate" de aquellas muestras que se habrían descartado al ser analizadas por otros protocolos.

El motivo por el cual una muestra presenta un número bajo de lecturas está estrechamente relacionado con el hecho de que, tras haber sido identificada mediante su *barcode*, se combinan para ser secuenciadas a la vez de manera que cada fragmento será identificado por medio de ese código de barras previamente añadido. Esto es lo que se conoce como *pooling*. Las librerías de DGP-A no se realizan siguiendo el mismo protocolo que otras librerías de NGS. En este caso, la concentración es determinada mediante la fluorescencia emitida, medida en equipos como el *Qubit*. Este método no tiene en cuenta el tamaño de los fragmentos de ADN, a diferencia de otros como *Bioanalyzer* o *TapeStation*, por lo que no es capaz de determinar con precisión la concentración exacta. Además, se da

la circunstancia de que las librerías destinadas a los análisis DGP-A presentan un rango de fragmentos muy amplio, que va desde los 200 a los 2000 pb, lo cual resulta mucho mayor que los fragmentos usualmente empleados. Finalmente, aunque los perfiles amplificados suelen ser similares entre muestras procedentes de embriones con la misma ploidía, esto no es necesariamente cierto en todos los casos. Por ejemplo, algunas muestras pueden presentar perfiles desviados relacionados con la calidad del ADN final analizado. Así, al equiparar las concentraciones entre las distintas muestras para realizar el *pool* o combinado que será secuenciado, puede obtenerse una mezcla que no sea realmente equimolar debido a todo lo anteriormente descrito, ocasionando que una o varias librerías queden sub y/o súper representadas.

Cuando una muestra presenta un elevado número de lecturas la dispersión se densifica debido a que las lecturas ocupan todo el espacio y "caen" homogéneamente dentro de las ventanas consideradas para el análisis de DGP-A, de manera que dichas ventanas se interpretan como si tuviesen la misma cobertura, percibiéndose como si de una línea continua se tratase, disminuyendo el valor de MAPD. Por el contrario, un archivo BAM con bajo número de lecturas será observado como un conjunto de lecturas discretas y dispersas y el valor de MAPD de la muestra ascenderá porque algunas ventanas no contendrán tantas lecturas como otras. Sin embargo, esta deriva del valor de MAPD podría ser un efecto secundario debido únicamente al número de lecturas, de manera que aporte una falsa sensación de seguridad que podría conducir a un pronóstico mal diagnosticado en la ploidía del embrión por haber quedado oculta tras el exceso de lecturas, o al descarte de embriones bien diagnosticados debida a una falsa inseguridad debida a la falta de lecturas suficientes.

Si nos fijamos en la propuesta planteada observamos que todas las muestras filtradas presentaron un valor aproximado de 0,28 en la dispersión de las lecturas. Esto resulta lógico, ya que el proceso de amplificación no es muy diferente entre unas muestras y otras y los filtros eliminan los artefactos. Por su parte, las muestras sin filtrar tampoco difieren en exceso de dicho parámetro, aunque se desvían en algunas categorías. Esto se debe nuevamente al proceso de amplificación; un proceso muy controlado en el que se introducen ciertas desviaciones pero nunca en tal exceso como para desviar completamente el perfil de la muestra, ya que en ese caso serían consideradas como un fallo de la técnica de amplificación y, probablemente, eliminadas del análisis DGP-A. Así, muestras con bajos números de lecturas podrían presentar una baja dispersión si estas se reparten homogéneamente por el genoma entre las distintas muestras, de forma que se produzca una representación equimolar al ADN original.

Empleando un estadístico del tipo planteado en esta tesis nos encontramos que cada muestra presenta un patrón de dispersión de las lecturas único pero relativamente controlado dentro de unos parámetros lógicos y que dicha dispersión no tendría que estar relacionada con la cantidad de lecturas necesariamente.

Así, el perfil real de una muestra mostrará una dispersión que será mucho mayor que la dispersión añadida por los artefactos que se hayan producido, debido al bajo porcentaje que estas lecturas duplicadas suponen con respecto al resto. Sin embargo esta dispersión añadida o ruido puede ser suficiente para que el diagnóstico emitido por el software de análisis difiera de la ploidía real de la muestra. Al eliminar dicho ruido con el filtro por medio de la eliminación de los artefactos, nos estaríamos encontrando con el valor de dispersión real de las lecturas del embrión, una dispersión que deberíamos considerar y no descartar en el análisis del modelo de ploidía presentado, pues no sería ya debido a artefactos, sino a la distribución de las lecturas entre las ventanas.

Por tanto, si todas las muestras presentan una dispersión mínima y el uso del filtro permite detectar esa dispersión por la eliminación total del ruido, nos encontramos frente a la hipótesis de que todos los embriones son comparables desde el punto de vista de la dispersión de las lecturas, con independencia del número de lecturas que presente. De esta forma, podría ser que la ecuación del MAPD estuviese introduciendo un falso efecto de categorización del valor de dispersión en función de las lecturas, mientras que empleando el estadístico *Z-ScoreAbs_NegLog* aquí propuesto podríamos descartar muestras cuya dispersión se saliese de los rangos considerados como "normales", no por su número de lecturas, sino porque dicha muestra presente un perfil efectivamente anómalo.

Simplificando el ejemplo, supongamos que el perfil real de una muestra fuese semejante a la curva del seno de x ($\sin 2x$), representado en la

Figura 33a, donde a mayor número de lecturas, mayor frecuencia de la onda debido a que las lecturas se condensan como lo hacen las ondas del seno en la gráfica. Por efecto de los artefactos introducidos durante la fase de amplificación se habría introducido un ruido que distorsionaría dicho perfil de manera que la curva del seno quedaría oculta y el modelo podría ser erróneamente clasificado como sin $4x$, representado en la

Figura 33b. Por otro lado, imaginemos un embrión cuyo perfil real sea sin $10x$ (Figura 34a), y que será identificado como sin $15x$ (Figura 34b). El objetivo del uso de algoritmos de filtrado reside en la eliminación de dicho ruido para que los programas de análisis sean capaces de detectar el modelo subyacente. Tras su aplicación nos encontraríamos con que el resultado asignado por el algoritmo de DGP-A para la primera muestra podría ser ahora sin $3x$ y para la segunda sin $12x$ (

Figura 33c y Figura 34c).

En este momento entraría en juego la medida de la dispersión para validar la confianza de dichos diagnósticos. Probablemente nos encontraríamos con que la primera muestra, cuya curva es más amplia y la onda es más dispersa, tendría valores de MAPD superiores a 0,3, mientras que la segunda muestra, cuyas curvas son más estrechas y el patrón se superpone cubriendo casi todo el espacio, tendría un valor inferior. Según esto, deberíamos descartar el "diagnóstico" de la primera muestra y considerar válido el de la segunda. Sin embargo, a simple vista no es sencillo concretar qué resultado fue más acertado en la detección del modelo subyacente.

Por tanto, la ecuación de cálculo del MAPD provoca un efecto secundario de ocultamiento que podría explicar la razón por la cual la ploidía de muestras filtradas con MiNFilterDups que presentaron valores de MAPD superiores a 0,3 continuó siendo correctamente detectada. El uso del modelo de Z-Score propuesto en esta tesis aporta una medida absoluta independiente del número de lecturas, que podría ser comparada y tenida en cuenta para comparar el nivel de confianza entre perfiles diferentes de muestras distintas.

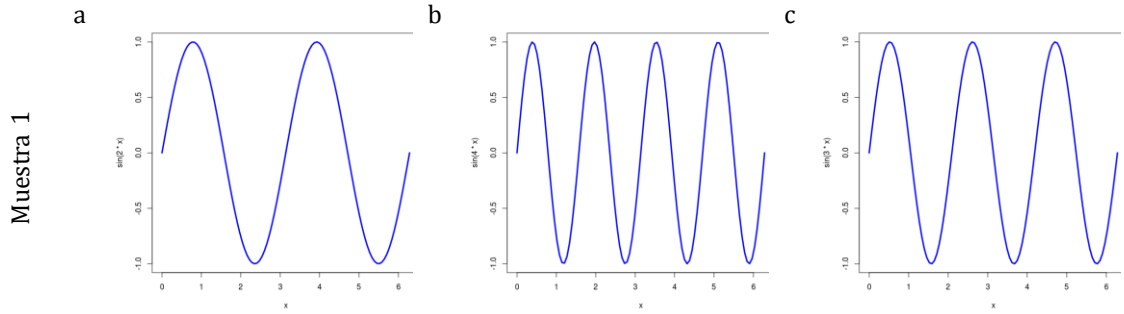

Figura 33: Perfil de a) seno(2x), b) seno(4x), c) seno(3x)

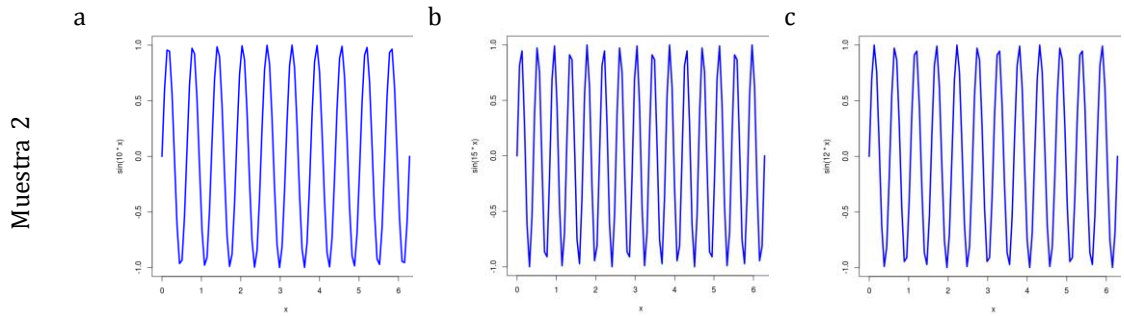

Figura 34: Perfil de a) seno(10x), b) seno(15x), c) seno(12x)

El valor de MAPD de la muestra está muy relacionado con el número de lecturas por lo que no permite la comparación entre muestras con distinto número.

La ploidía de un embrión puede ser correctamente determinada con independencia de su valor de MAPD si ha sido correctamente filtrada.

El logaritmo negativo en base 10 del valor absoluto del Z-Score de la cobertura de las muestras es una medida absoluta que permite la comparativa de la dispersión de las lecturas entre las muestras con independencia del número de lecturas que presenten.

Capítulo 4: MiNtagSNP: Algoritmo de selección de tagSNPs para la maximización de la informatividad en DGP-M

4.1 Valores óptimos para MAxP y HETrate

"Esta tesis doctoral está sometida a procesos de protección o transferencia de tecnología o de conocimiento, por lo que esta sección está inhibida en la publicación en los repositorios institucionales. Esta sección contiene, además, los siguientes elementos inhibidos:

Figura 35: Dispersión de los valores de MaxP frente a la informatividad arrojada por cada SNP de la población analizada.

Figura 36: Dispersión de los valores de HETrate frente a la informatividad arrojada por cada SNP de la población analizada.

Figura 37: Dispersión de los valores de MaxP frente a la informatividad arrojada por los SNP de la población analizada que sobrepasaron el valor umbral 0,1.

Figura 38: Dispersión de los valores de HETrate frente a la informatividad arrojada por los SNP de la población analizada que sobrepasaron el valor umbral 0,22.

Figura 39: Dispersión de los valores de MaxP y HETrate de los SNPs que componen la Matriz M.

Autorizado por la Comisión General de Doctorado de la Universidad de Murcia con fecha 24 de febrero de 2021"

4.2 Validación In silico

"Esta tesis doctoral está sometida a procesos de protección o transferencia de tecnología o de conocimiento, por lo que esta sección está inhibida en la publicación en los repositorios institucionales.

Autorizado por la Comisión General de Doctorado de la Universidad de Murcia con fecha 24 de febrero de 2021"

4.2.1 Población

"Esta tesis doctoral está sometida a procesos de protección o transferencia de tecnología o de conocimiento, por lo que esta sección está inhibida en la publicación en los repositorios institucionales. Esta sección contiene, además, los siguientes elementos inhibidos:

Tabla 11: Distribución demográfica de la población escogida.

Autorizado por la Comisión General de Doctorado de la Universidad de Murcia con fecha 24 de febrero de 2021"

4.2.2 Paneles diseñados

"Esta tesis doctoral está sometida a procesos de protección o transferencia de tecnología o de conocimiento, por lo que esta sección está inhibida en la publicación en los repositorios institucionales. Esta sección contiene, además, los siguientes elementos inhibidos:

Tabla 12: Polimorfismos registrados en la base de datos 1000GenomesDB en cada región analizada.

Tabla 13: Número de tagSNPs diseñados por cada algoritmo en cada región analizada.

Figura 40: Relación r^2 entre los distintos tagSNPs de los paneles diseñados por el algoritmo a), e), i), m) SNPinfo, b), f), j), n) MiNtagSNP, c), g), k), o) OMNI5.2 y d), h), l), p) Selección aleatoria para las regiones a), b), c), d)ATXN2, e), f), g), h) CFTR, i), j), k), l)PKD1, y m), n), o), p) VHL

Figura 41: Relación D' entre los distintos tagSNPs de los paneles diseñados por el algoritmo a), e), i), m) SNPinfo, b), f), j), n) MiNtagSNP, c), g), k), o) OMNI5.2 y d), h), l), p) Selección aleatoria para las regiones a), b), c), d)ATXN2, e), f), g), h) CFTR, i), j), k), l)PKD1, y m), n), o), p) VHL.

Autorizado por la Comisión General de Doctorado de la Universidad de Murcia con fecha 24 de febrero de 2021"

4.2.3 Tiempo de ejecución

"Esta tesis doctoral está sometida a procesos de protección o transferencia de tecnología o de conocimiento, por lo que esta sección está inhibida en la publicación en los repositorios institucionales. Esta sección contiene, además, los siguientes elementos inhibidos:

Tabla 14: Resumen de la reducción y tiempo empleados por MiNtagSNP para el diseño de los paneles.

Autorizado por la Comisión General de Doctorado de la Universidad de Murcia con fecha 24 de febrero de 2021"

4.2.4 Informatividad

"Esta tesis doctoral está sometida a procesos de protección o transferencia de tecnología o de conocimiento, por lo que esta sección está inhibida en la publicación en

los repositorios institucionales. Esta sección contiene, además, los siguientes elementos inhibidos:

Tabla 15: Porcentaje de informatividad arrojada por los tagSNP seleccionados por cada algoritmo en 300 cruces aleatorios.

Autorizado por la Comisión General de Doctorado de la Universidad de Murcia con fecha 24 de febrero de 2021"

4.2.5 Imputación

"Esta tesis doctoral está sometida a procesos de protección o transferencia de tecnología o de conocimiento, por lo que esta sección está inhibida en la publicación en los repositorios institucionales. Esta sección contiene, además, los siguientes elementos inhibidos:

Tabla 16: Precisión de imputación para cada algoritmo en cada región.

Tabla 17: Datos relativos al análisis de la zona de 14 Mb.

Autorizado por la Comisión General de Doctorado de la Universidad de Murcia con fecha 24 de febrero de 2021"

4.3 Implementación

"Esta tesis doctoral está sometida a procesos de protección o transferencia de tecnología o de conocimiento, por lo que esta sección está inhibida en la publicación en los repositorios institucionales.

Autorizado por la Comisión General de Doctorado de la Universidad de Murcia con fecha 24 de febrero de 2021"

4.3.1 Informatividad

"Esta tesis doctoral está sometida a procesos de protección o transferencia de tecnología o de conocimiento, por lo que esta sección está inhibida en la publicación en los repositorios institucionales. Esta sección contiene, además, los siguientes elementos inhibidos:

Tabla 18: Tabla resumen de casos in vitro. *Se indican con asterisco (*) las 5 parejas donde hubo el mayor porcentaje de tagSNPs secuenciados en ambos individuos; se indica con dos asteriscos (**) los 5 casos con el menor porcentaje.*

Autorizado por la Comisión General de Doctorado de la Universidad de Murcia con fecha 24 de febrero de 2021"

4.3.2 Imputación

"Esta tesis doctoral está sometida a procesos de protección o transferencia de tecnología o de conocimiento, por lo que esta sección está inhibida en la publicación en los repositorios institucionales. . Esta sección contiene, además, los siguientes elementos inhibidos:

Tabla 19: Resumen de la imputación de tagSNPs en las muestras con mayor y menor porcentaje de polimorfismos secuenciados

Autorizado por la Comisión General de Doctorado de la Universidad de Murcia con fecha 24 de febrero de 2021"

4.4 Discusión

4.4.1 Estrategia algorítmica de MiNtagSNP

Intuitivamente podríamos suponer que un análisis que permite la detección directa de la alteración causante de una enfermedad es la manera correcta de comprobar si dicha alteración está presente en la descendencia; sin embargo, debido al llamado efecto de *allele drop-out*, frecuente en DGP-M, el alelo alterado podría no amplificarse y, por tanto, no ser detectado, dando como resultado la transferencia de un embrión portador. El análisis de marcadores próximos al alelo causante permite descartar dichos embriones con mayor seguridad al desestimar no solo aquellos que muestran la mutación de forma directa, sino aquellos que, aun no mostrándola, presentan el haplotipo identificado como asociado a dicha alteración en los progenitores. Por tanto, una ventaja del método es que puede ser empleado para realizar un doble análisis mediante la detección directa, cuando fuese posible (siendo necesario tan solo conocer la posición exacta de la mutación) y una detección indirecta (a través de los polimorfismos). Además, en aquellos casos donde la detección directa no es viable, tales como grandes *indels* o variaciones del número de repeticiones, el método puede ser igualmente empleado, permitiendo descartar igualmente los haplotipos asociados al parental portador.

Clásicamente, en DGP-M el análisis indirecto se realiza mediante el estudio de STRs y secuenciación capilar. En esta tesis se propone la utilización de la técnica de secuenciación

masiva y el uso de SNPs para tal proceso. Para ello, se ha desarrollado un algoritmo de cálculo de tagSNPs que permita realizar la selección de los polimorfismos útiles en DGP-M de entre todos los existentes en una región concreta.

La densidad de polimorfismos es muy variable a lo largo del genoma[293], sin embargo, se suele considerar un promedio de aparición de un SNP cada 145 pares de bases[294], de manera que podríamos encontrar unos 27500 SNPs dentro de una región de 4Mb como la región bloque tomada por MiNtagSNP. De esta forma, aunque podría caber la posibilidad de secuenciarlos todos, dado que la mayoría forma parte de haplobloques su secuenciación es del todo innecesaria, pues tan solo aporta información redundante a la par que incrementa notablemente el coste de la prueba[295]. Por este motivo se emplean estrategias de búsqueda de tagSNPs que representen a los polimorfismos de un haplobloque. Si bien es cierto que la utilización de tagSNPs es algo que viene realizándose desde hace muchos años, las estrategias existentes no son óptimas para la aplicación propuesta en la presente tesis principalmente debido a dos motivos: (1) los SNPs seleccionados deben ser informativos con una alta probabilidad para permitir la detección de fenómenos de recombinación y el reconocimiento de las líneas de herencia en una muestra trío; y (2) la región seleccionada abarca tan solo unas 4Mb, lo que dificulta la identificación de posibles tagSNPs debido a correlaciones intrínsecas difíciles de cuantificar.

En cuanto a la región seleccionada, la región bloque se centra en unas pocas megabases alrededor del gen que presenta la alteración (2Mb a cada lado por defecto). Esta región es suficientemente grande como para permitir detectar un número suficiente de tagSNP, pero los bordes se encuentran lo suficientemente cercanos al gen como para que las probabilidades de ocurrencia de fenómenos de recombinación sean bajas, lo que facilita el análisis posterior. En este tipo de estudios la recombinación es un riesgo que puede desembocar en un diagnóstico erróneo. En cada embrión cabe la posibilidad de que se identifiquen varios polimorfismos asociados al alelo silvestre, por ejemplo, pero que luego el gen presente la alteración porque justo en ese punto se haya producido un sobrecruzamiento entre el alelo silvestre y el mutado; si bien es cierto que la frecuencia de recombinación es también un parámetro relativamente variable a lo largo del genoma, en general se considera que es de alrededor del 1,2% por cada megabase[296]. Considerando el bloque de 4MB la probabilidad de una doble recombinación se reduciría al 0,01 %, lo cual permite, junto al hecho de que se toman polimorfismos a ambos lados del gen de interés, el uso de esta estrategia en los análisis de DGP-M.

La principal diferencia entre las hipótesis de asociación y las hipótesis libres de asociación empleadas para dirigir la estrategia de selección de los polimorfismos que serán tag, reside en que el segundo tipo de estrategia permite detectar no solo aquellos haplotipos que en la población general están relacionados con una enfermedad con mayor probabilidad, sino diseñar un panel de tagSNPs que permita detectar cualquier posible haplotipo, haya sufrido o no fenómenos de recombinación con respecto al haplotipo ancestral. Esto es debido a que en cada polimorfismo cada alelo es identificado como heredado a partir de uno u otro progenitor, permitiendo reconstruir la línea de herencia y, con ello, adaptar el proceso de selección de los embriones al caso concreto presentado por la pareja que se somete al ciclo de IVF, razón por la cual el algoritmo MiNtagSNP presentado sigue esta segunda corriente.

Calcular un conjunto mínimo y óptimo de tagSNPs es un problema NP-complejo[224] es decir, no es posible calcular una solución óptima para todo el genoma en tiempo y espacio abarcable. Sin embargo, sí es posible hacerlo para regiones más pequeñas. MiNtagSNP aprovecha esta ventaja y se centra en la identificación de tagSNPs útiles en la caracterización de regiones de ascendencia común es decir, haplotipos conservados y no tanto en aquellos SNPs más alejados que pudieran estar asociados por selección natural, hibridación o procesos aleatorios de herencia. Los SNPs de una región de ascendencia común se encuentran suficientemente cercanos dentro del genoma como para impedir que se produzcan fenómenos de recombinación frecuentes entre ellos. Esto implica que es posible definir una región relativamente pequeña del genoma y emplear el algoritmo MiNtagSNP para obtener tagSNPs útiles en la caracterización de la línea de descendencia y, con ello, realizar el análisis DGP-M sobre los embriones.

La selección de tagSNPs ha sido ampliamente estudiada, dando como resultado diversas metodologías que pueden encontrarse descritas en la bibliografía[226,227,297–299]. Sin embargo, todas ellas se basan en 3 estrategias principales para realizar dicha selección:

1. Los métodos basados en el desequilibrio de ligamiento intentan identificar un conjunto de regiones tales que los SNPs de cada ventana estén en alto desequilibrio de ligamiento entre sí[223]. Estas regiones ventana se autodefinen con base en los valores que los polimorfismos muestran en ciertos parámetros que cada algoritmo considera, siendo estos parámetros variables entre los distintos algoritmos desarrollados. Lo interesante de esta estrategia es, por tanto, que son los propios polimorfismos los que se

"identifican" entre sí, mostrando las correlaciones que existen entre ellos a través del desequilibrio de ligamiento.

2. Los métodos basados en bloques se basan en el hecho de que el 50% de los fenómenos de recombinación ocurren en menos del 10% del genoma[232], lo cual permite definir, siempre *a priori*, bloques haplotípicos donde la probabilidad de recombinación es casi inexistente. De esta manera los tagSNPs se escogen de entre todos los polimorfismos pertenecientes al bloque. La desventaja de este tipo de métodos reside en la multitud de estrategias existentes para generar dichos bloques, de manera que un mismo algoritmo podría producir diferentes conjuntos de tagSNPs. Lo interesante de estas estrategias es que permiten acotar la búsqueda de tagSNPs a un número limitado de polimorfismos, definidos en este caso por su proximidad en el genoma; de esta forma el problema se hace fácilmente computable en tiempo y recursos, alcanzando rápidamente una solución óptima.
3. Las estrategias libres de bloques emplean estadísticos basados en vecindad para seleccionar tagSNPs a lo largo de todo el espacio de polimorfismos disponible, considerando como parámetro de entrada el número de tagSNPs que tendrá el panel final[237]. Nuevamente, al igual que en las estrategias basadas en LD, resulta interesante el hecho que de los polimorfismos no son previamente divididos, pero en este caso entran en juego otros parámetros que definen la correlación entre los mismos como los estadísticos de vecindad en proximidad o lejanía o la asignación de pesos estadísticos con base en el p-valor de la χ^2.

Mientras que los métodos basados en esta tercera estrategia se basan en encontrar un conjunto de tagSNPs tal que el haplotipo de una muestra desconocida pueda ser predicho con alta precisión[300], métodos basados en las dos primeras estrategias tienen como fin encontrar un conjunto mínimo y óptimo de tagSNPs para ser empleado en estudios de asociación[239]. Por su parte, los métodos basados en LD identifican tagSNPs que pueden identificar a otros polimorfismos que están separados en la secuencia genómica, mientras que los métodos de bloques se usan para escoger tagSNPs que representen polimorfismos en una región continua[236]. Sin embargo, los SNPs seleccionados mediante métodos basados en LD pueden fallar a la hora de distinguir todos los posibles haplotipos existentes para una región en la población, mientras que los seleccionados por estrategias de bloque pueden distinguirlos todos.

Así, empleando métodos basados en LD, como el desarrollado por *Quin et al.*[301], el *SNPPicker*[302] o el propio *SNPinfo*[235] evaluado en esta tesis, logramos identificar polimorfismos en alto LD con otros situados en regiones lejanas del genoma, lo que resulta ventajoso a la hora de identificar, en los embriones, el haplotipo parental portado; sin embargo, debido a que se pretende aplicar en muestras concretas y no en poblaciones, donde no tiene cabida una cierta tasa de error de estimación debido a que podría desembocar en la transferencia de un embrión enfermo, estos métodos no resultan fiables en tanto que el panel diseñado podría fallar en la distinción del haplotipo concreto presentado por el parental. Por su parte, en la práctica, el empleo de un parámetro de división fijo no resulta realmente práctico ni deseable, pues a lo largo de la evolución las tasas de recombinación y, con ello, el desequilibrio de ligamiento entre polimorfismos ha ido variando a lo largo del genoma[303] , cosa que demuestra el hecho de que el número de polimorfismos que puede cubrir un tagSNPs es completamente diferente entre unos y otros. Además, enfoques basados en bloques como el presentado por *Zhang et al.*[304] o el enfoque de *Schulze et al.*[305] consideran que dos SNPs tendrán un nivel de correlación suficiente como para ser imputados uno a partir del otro y ser tagSNP si se producen en el mismo bloque de haplotipos, es decir, si hay poca o nula evidencia de recombinación dentro de dicho bloque. Esto no es lo mismo que decir que un SNP puede ser usado para predecir otro si hay poca evidencia de recombinación entre ellos. Además, el hecho de predefinir las regiones con base en criterios de división generales para todo el genoma y no con base en la alteración concreta presentada, puede provocar que polimorfismos en distintos bloques si estén realmente correlacionados entre ellos, pero que por azar hayan caído dentro de bloques distintos y por ello no se computen, lo que provocaría cierta redundancia de información en el conjunto final. Finalmente, métodos libres de bloques como el de *Halldórsson et al*.[300] resultan muy interesantes para abordar el problema presentado por el análisis DGP-M debido a la capacidad predictiva de los tagSNPs seleccionados. Sin embargo la no división de los polimorfismos en subconjuntos convierte el problema de la selección de tagSNPs en un problema costoso computacionalmente con respecto a las otras dos estrategias, lo que incrementa el tiempo necesario para la obtención de resultados.

Por todas estas razones MiNtagSNP combina las tres estrategias en una sola, permitiendo definir un bloque basado, no en una longitud de secuencia fija por defecto para todo el genoma de forma que los bloques abarquen aquellos genes que "caigan" en su interior, sino en la bibliografía existente sobre la alteración para la que se desea emplear el método, de forma que el bloque abarque la región de interés a partir de la posición de dicha alteración. A su vez, dentro de dicho bloque, los polimorfismos son analizados en primera

instancia como en un enfoque libre de bloques a partir de las correlaciones r^2 [261] existentes entre los mismos, lo cual genera grupos con una evidente baja tasa de recombinación dentro de los cuales se analiza el desequilibrio de ligamiento *D'* que permite seleccionar los tagSNPs que compondrán el panel final.

Existen evidencias del éxito de la correlación r^2 en estudios de asociación de caso-control, donde SNPs asociados directamente con haplotipos causantes de enfermedad son reemplazados por la secuenciación de tagSNPs, asociados de forma indirecta[222]. Esto resulta interesante a la hora de seleccionar polimorfismos que permitan identificar posteriormente cualquier haplotipo presente en la pareja en estudio. Por su parte, la medida D' describe la relación genotípica existente entre dos polimorfismos. Un inconveniente de esta medida reside en su sensibilidad a valores pequeños de frecuencia alélica, como ocurre en los casos de polimorfismos raros. Así, si suponemos dos SNPs bialélicos SNP-1 cuyos alelos pertenecen al conjunto {a, b} y SNP-2, {c, d} y que la frecuencia alélica del alelo alternativo b del SNP-1 es muy baja, este polimorfismo podría presentar un desequilibrio de ligamiento completo con el SNP-2 si el alelo b se encontrase solamente presente cuando lo está el alelo c. Por su parte, la r^2 no sería tan alta, pues que aparezca el alelo c no implicaría necesariamente que el alelo b estuviese presente, ya que a veces el alelo c aparece con el alelo a. MiNtagSNP solventa este inconveniente gracias a la preselección que realiza el paso con *SPA* para maximizar la solución del problema de la informatividad antes de la selección de los tagSNPs del *SSA*, de manera que polimorfismos con frecuencias muy bajas suelen quedar excluidos de la Matriz *M*.

La ingente cantidad de polimorfismos presente en el genoma implica que, incluso en un estudio con una muestra poblacional grande y suficiente, algunos de estos polimorfismos podrían correlacionarse por casualidad a pesar de encontrarse bastante alejados. Normalmente, las correlaciones ocurren entre polimorfismos que se encuentran físicamente cerca. Resulta lógico creer que dos polimorfismos en dos cromosomas diferentes no estén realmente correlacionados. Sin embargo por efecto del azar, podrían presentar un desequilibrio de ligamiento significativo. Además, el hecho de que dos polimorfismos presenten un desequilibrio de ligamiento *D'* máximo no significa que uno de ellos pueda ser empleado para predecir el otro con alta precisión, pero una correlación r^2 máxima sí supone una predicción exacta[223,301,306–308]. Por todos estos motivos MiNtagSNP presenta una selección en dos pasos, empleando en primer lugar el análisis de la correlación r^2 para generar ventanas de SNPs que pueden predecirse unos a través de los valores de los otros y, en segundo lugar, a partir de su desequilibrio de ligamiento D' determina aquellos que actuarán mejor como tagSNPs.

La varianza de las tasas alélicas, es decir, la diferencia en los valores de dichas frecuencias, es ingente para un gran número de polimorfismos. En muchos casos esto puede deberse a que el tamaño muestral empleado en la estimación de dichas frecuencias es muy pequeño, pero incluso cuando el tamaño de muestra es suficiente, se han visto fluctuaciones importantes entre grupos étnicos o, incluso, regiones de una misma zona geográfica[309]. Debido a esto, MiNtagSNP selecciona tagSNPs a partir de una población muestral representativa de la población de interés, de forma que podamos esperar que las correlaciones descubiertas se observen, de forma generalizada, en toda la población de interés y, con ello, en la pareja en estudio.

Además, a pesar de que existen multitud de medidas y parámetros que pueden ser calculados para determinar si un SNP puede ser o no considerado útil en la predicción de otros polimorfismos, el caso concreto de la selección de SNPs útiles en DGP requiere considerar más factores que la correlación entre los SNPs de la región estudiada, pues no solo deben ser representativos del haplotipo presentado por los individuos sino identificativo y diferencial entre ellos, para permitir establecer con exactitud qué polimorfismos han sido heredados por los embriones. Así, el MAF juega un papel esencial. Para que un SNP sea informativo debe estar en heterocigosis en un miembro de la pareja y en homocigosis en el otro. La probabilidad de que esto ocurra aumenta con el valor del MAF, ya que a mayor frecuencia alélica, mayor número de individuos podrán portarlo. Sin embargo esto no es del todo cierto, pues el alelo podría presentarse siempre en heterocigosis. Esta es la razón de no emplear el MAF como marcador del suceso de éxito para el problema de la informatividad. Como comentamos anteriormente, el suceso exitoso es aquel donde un padre sea homocigoto y el otro heterocigoto, es decir, donde uno de los 4 alelos sea distinto [*p p p q*], [*q q q p*], [*p p q p*] y [*q q p q*]. Esto se cuantifica a través de la ecuación MaxP, que calcula la probabilidad de encontrar dicho suceso y se evalúa para cada polimorfismo en la población. Posteriormente, la ecuación HETrate permite cuantificar los individuos heterocigotos en la población, ya que debido a los fenómenos de ADO entre otras cosas, estos individuos pueden estar subestimados (resultados no publicados).

Finalmente debemos destacar que MiNtagSNP no realiza distinciones entre regiones codificantes y no codificantes, ya que se ha reportado que no existe significancia en la densidad de SNPs[233].

Así, MiNtagSNP puede integrarse fácilmente en los protocolos de análisis DGP-M ya que puede ser empleada a través de la línea de comandos para diseñar paneles de tagSNPs optimizados para resolver el problema de la informatividad con la mínima redundancia de

información, en una o varias poblaciones teniendo en cuenta el desequilibrio de ligamiento y otros contrastes configurados por el usuario, para ser empleados en la reconstrucción de los haplotipos progenitores a fin de seleccionar embriones potencialmente transferibles libres de alteraciones monogénicas. Para ello hemos propuesto un método que separa predicción de polimorfismos informativos (a través de la maximización de la solución el problema informativo) de la selección de tagSNPs (por medio de la cuantificación de la correlación y el desequilibrio de ligamiento).

Encontrar un set de tagSNPs es siempre un reto debido a que la información haplotípica no está siempre disponible. Es por ello conveniente que los métodos como el nuestro puedan ser empleados con datos tanto genotípicos como haplotípicos.

Por último, se estima que de forma aleatoria son necesarios al menos 15 tagSNPs para encontrar uno informativo[310], lo que supone un 6%, aunque en la presente tesis se obtuvo un valor que ronda el 3,5%. La bibliografía recoge que el Karyomapping presenta un número de tagSNPs informativos que asciende al 20%[218]. Por su parte, OMNi presentó un 35% de informatividad sobre los cruces obtenidos a partir de 1000GenomesDB, mientras que SNPinfo ascendió a 40,3%. MiNtagSNP logró tasas del 49,7%, próximas al máximo teórico alcanzable según el problema de la informatividad. Esto demuestra que MiNtagSNP es capaz de obtener polimorfismos informativos que retienen una mayor cantidad de información de la región de estudio a la vez que minimiza el número de tagSNPs necesarios para cubrir dicha región.

4.4.2 Tiempo de ejecución

En la teoría, el uso de una estrategia de búsqueda exhaustiva es la única que asegura la obtención de una solución óptima, sin embargo, su aplicación al problema de la selección de tagSNPs en todo el genoma de forma directa se hace imposible actualmente debido al costo computacional que supone analizar cada uno de los polimorfismos existentes. Dado que, intuitivamente, podemos afirmar que el desequilibrio de ligamiento ocurre entre polimorfismos que se encuentran próximos en el genoma (posiciones alejadas son susceptibles de experimentar fenómenos de recombinación entre ellos), una solución práctica consiste en la descomposición del set de polimorfismos en "grupos" tales que SNPs que se encuentren en diferentes grupos nunca presentarán un desequilibrio de ligamiento significativo. Sin embargo, esto podría provocar pérdidas de información. El establecimiento de los grupos en función de su posición genética puede generar que SNPs

de distintos grupos si se encuentren en LD, ya que los parámetros de recombinación varían a lo largo de todo el genoma en función de la región y los genes que ésta albergue. Por ello, hemos decidido acotar la región a analizar en función del gen de interés que se desee estudiar, como si de un enfoque en bloques se tratase. Formamos tan solo un único bloque, tratado internamente como en un enfoque libre de bloques, pues dentro no se clasifican sus SNPs en función de la posición genómica, sino del grado de correlación que presenten. Esta combinación de doble estrategia permite disminuir notablemente el rango de datos a analizar, haciendo que el tiempo de computación empleado por la estrategia de búsqueda exhaustiva sea perfectamente abarcable.

Por otro lado, la mayoría de los algoritmos precisan de mucho tiempo para realizar la selección de tagSNPs debido a que precisan analizar cada polimorfismo del grupo uno por uno. En cambio, la preselección que realiza MiNtagSNP para descartar aquellos SNPs no informativos, acelera considerablemente el proceso. Además, el hecho de que los polimorfismos sean ordenados según el rango funcional (capacidad de representar a otros polimorfismos del grupo) permite reducir la complejidad computacional del proceso de selección de los tagSNPs, ya que los primeros polimorfismos siempre serán mejores candidatos que los siguientes. Hay que destacar que el tiempo computacional del SSA se incrementará a medida que lo haga la cantidad de datos a procesar, siendo este el paso más costoso computacionalmente. Por este motivo es esencial la división en dos pasos con preselección de candidatos.

También se estableció el uso de una estrategia basada en métodos de LD para calcular el nivel de correlación entre SNPs. En la fase *SSA*, MiNtagSNP calcula los valores de los parámetros r^2 y *D'* para cada par de SNPs a través del uso de una modificación que hemos realizado con base en el paquete *genetics*[262] de R[263]; estas modificaciones del código del paquete no solo permiten el cálculo de los estadísticos de interés, sino acelerar el proceso de obtención de los mismos.

Todo esto supone un gran avance, pues el reducido tiempo necesario para el diseño del panel final de tagSNPs permite reducir el tiempo de espera desde la obtención de la biopsia hasta el embarazo, lo que disminuye las tasas de estrés de las pacientes, un parámetro fundamental en el éxito de una transferencia *in vitro*[311,312].

Así, se observó que gracias a la reducción del 98,8% realizada por SSA el algoritmo SPA tardó una media de 26minutos y 50 segundos en seleccionar los tagSNPs. Además, de aquí se deduce que SSA es capaz de reducir el set inicial de polimorfismos a ese,

aproximadamente, 0,1% de polimorfismos que genera la diversidad. Finalmente podemos concluir que 27 minutos es un tiempo razonable para un proceso de selección de tagSNP.

4.4.3 D', r2, HETrate y MaxP

La correlación r^2 mide la variación explicada por la regresión del modelo, es decir, la fracción por la cual la varianza de los errores es menor que la varianza de la variable dependiente. Esto indica que no se está midiendo la capacidad predictiva del modelo, sino la relación de cada uno de los datos con el modelo. Es por esta razón que la medida de correlación r^2 es empleada para predecir cuáles serán los polimorfismos que actuarán como mejores tagSNPs, pero no para indicar la capacidad de predicción de los tagSNPs escogidos en el panel final.

Por su parte, el estadístico D' mide el desequilibrio de ligamiento entre dos polimorfismos, o lo que es lo mismo, las veces que uno aparece con respecto al otro, hecho que hemos discutido el apartado *4.4.1* Estrategia algorítmica de MiNtagSNP del bloque *IV* Resultados y discusión.

El valor mínimo de r^2 y *D'* recomendado para asegurar una correlación significativa es 0,75[264]. Sin embargo, el valor umbral de ambos estadísticos depende del objetivo a cumplir. Cuando la pretensión se centra, como en este caso, en la predicción de una respuesta con precisión a partir de los datos disponibles, debemos pues tratar de seleccionar un set de datos inicial tal que la mayor parte de la información quede contenida en dicho set. Este set de datos serán los tagSNPs y la población a inferir estará compuesta por el resto de polimorfismos. Basándonos en todos estos argumentos, y sabiendo que la bibliografía considera que a partir de 0,8 se puede considerar la existencia de una correlación muy fuerte, se decidió establecer el valor umbral de r^2 en 0,9, de forma que el 90% de la varianza de los polimorfismos a los que cubre un tagSNP queda explicada por dicho tagSNP. A su vez, se estableció 0,85 como umbral de D', de forma que los tagSNPs presentasen un desequilibrio de ligamiento del 85% con los polimorfismos que cubren. Estos altos valores aseguran que los tagSNPs sean realmente restrictivos de manera que un SNP será considerado tagSNP única y exclusivamente si proporciona casi la misma información que otro polimorfismo sobre el que ha sido considerado tagSNP.

La razón de emplear valores más restrictivos para r^2 que para D' se fundamenta en dos premisas. Por un lado, la correlación r^2 es empleada en primer lugar para generar los

grupos de polimorfismos sobre los que se seleccionarán los tagSNPs. En este paso resulta interesante que los SNPs de cada grupo estén muy correlacionados, de manera que sean lo más similares. Esto facilita que, en la segunda parte del algoritmo SPA, los tagSNPs puedan ser escogidos con un valor un poco más laxo del parámetro D'. Debemos señalar también que, a pesar de que los datos no han sido incluidos en la presente tesis, pruebas realizadas con parámetros más altos seleccionaron números de tagSNPs superiores, lo que no resultó tan interesante debido a la redundancia de información y al costo de secuenciación, mientras que valores menores de estos parámetros disminuían el número de tagSNPs del panel final, lo que provocaba que, debido al alto porcentaje de polimorfismos que no llegan a ser secuenciados, no se obtuviesen suficientes posiciones informativas para realizar el análisis DGP-M.

Por su parte, el estudio realizado para la determinación de los valores umbrales de MaxP y HETrate, permitió establecer 0,01 y 0,22 como los valores óptimos para dichos parámetros en la población europea de 1000GenomesDB. Además, aunque no se publica en esta tesis, a partir de pruebas realizadas en nuestro laboratorio y aplicaciones diseñadas para clientes de la empresa, se ha comprobado que dichos valores pueden ser extrapolados a otras poblaciones obteniendo resultados igualmente exitosos.

4.4.4 Paneles de SNPs

Los valores de los SNPs para cada individuo fueron obtenidos a partir de la base de datos 1000Genomes DB. En dicha base de datos encontramos registrados 404 individuos pertenecientes a la población Europea. Estos individuos fueron separados en población referencia y población muestral. Los tagSNPs fueron calculados a partir de la población referencia. La población referencia fue empleada para la imputación de los valores de los polimorfismos a partir de los tagSNPs seleccionados. Esto es debido a que no debe emplearse como población test la misma población sobre la que se calcula un modelo, ya que los datos serán efectivamente buenos al emplear dicho modelo para inferir el resto de parámetros.

El intensivo estudio *in silico* realizado sobre los genes VHL, CFTR, ATXN2 y PKD1, mostró que MiNtagSNP emplea menor número tagSNP que las estrategias empleadas por SNPinfo y el *array* OMNI2.5 para maximizar la informatividad y precisión de imputación.

MiNtagSNP diseñó 1 SNP cada 30,5Kb para el caso donde el panel final fue de menor tamaño; según esto, los paneles diseñados por MiNtagSNP pueden ser empleados en la

detección de fenómenos de pérdida de heterocigosis, puesto que se ha establecido que la distancia mínima entre SNPs para poder analizar dicho fenómeno es de 1Mb[313] .

Además, el algoritmo proporciona dos posibilidades adicionales que permiten reducir el número de tagSNPs seleccionados, pero que pueden generar pérdidas de información debidas a que ciertos SNPs de bloque región podrían quedar sin cubrir por ninguno de los tagSNP incluidos en el panel final. La primera consiste en declarar que todo tagSNP debe estar en LD con al menos un número *X* de SNPs especificado por el usuario. Esta opción puede ser de utilidad para eliminar *singleton* tagSNPs. La segunda opción consiste en la especificación del número máximo de tagSNPs de los que puede constar el panel; debido a que los SNPs son escogidos de acuerdo a su rango funcional, tagSNPs en alto LD con un gran número de SNPs se seleccionan en primer lugar, mientras que los *singleton* tagSNPs son escogidos en último lugar. Añadiendo esta restricción, el algoritmo se detendría tras haber escogido *X* número de SNPs, en orden decreciente de rango funcional, y dejaría de añadir tagSNPs al panel final con independencia de si los tagSNPs seleccionados son suficientes o no para cubrir todo el bloque región de interés.

Como ya hemos comentado, con una estrategia adecuada el usuario puede modificar los valores umbrales de *MaxP*, *HETrate*, r^2 y *D'* para ampliar el número de tagSNPs escogidos. Sin embargo, cabe destacar en este punto que el número de polimorfismos a analizar para seleccionar los embriones potencialmente transferibles durante el análisis DGP-M se puede aumentar tras el diseño con AmpliSeq. Aunque inicialmente solo queramos secuenciar un polimorfismo es posible que dentro del amplicón diseñado (que puede tener entre 100 y 400pb) se encuentren más polimorfismos con un valor de MAF superior al 1%. Hemos denominado estos polimorfismos como *cstSNPs* (por sus siglas en inglés, *common and secondary tagSNPs*). Esta acción resulta en un incremento del número total de SNPs del panel a coste cero, ya que todo locus dentro del amplicón diseñado para secuenciar el tagSNP es efectivamente secuenciado y, por tanto, su información está disponible. Este hecho resulta muy útil, ya que el modelo obtiene los tagSNPs a partir de datos de grandes poblaciones, sin embargo, estos datos luego pueden no cumplirse totalmente en la pareja concreta sobre la que se aplicar el análisis. De esta manera, el SNP informativo podría no ser el tagSNP escogido sino el polimorfismo asociado a dicho tagSNP.

4.4.5 Informatividad y precisión de imputación

La informatividad obtenida por el panel de tagSNPs seleccionados empleando el algoritmo MiNtagSNP fue siempre superior a la informatividad mostrada por el resto de los métodos de selección. Además esta informatividad en promedio fue del 49,7% (con una desviación del 0,1%), muy cercana al 50% máximo teórico posible, según lo explicado en la sección 2.2.2 Problema de la informatividad del bloque 0 Introducción. Por su parte, el resto de métodos analizados mantuvieron unas cifras de polimorfismos informativos muy por debajo de dicho máximo. Estas cifras se mantuvieron en la validación *in vitro* realizada sobre los casos recogidos en nuestro laboratorio. Esto demuestra que MiNtagSNP es capaz de maximizar la informatividad recogida por el panel de SNPs seleccionado con respecto a los métodos del estado del arte aquí presentados.

La precisión de imputación obtenida a partir de la simulación *in silico* fue similar para todos los métodos, a pesar de que los gráficos de correlación r^2 y *D'* entre los polimorfismos de cada panel mostraron mayor independencia entre los polimorfismos escogidos con algoritmos de selección de tagSNPs que los escogidos por selección aleatoria. Además, se observó que, a pesar de que dichos gráficos mostraron una muy baja LD entre los polimorfismos seleccionados por MiNtagSNP, aún fue posible imputar más del 61% de los tagSNPs que no habían sido secuenciados correctamente a partir de los tagSNPs del panel que sí lo habían sido. Todo esto sugiere que la región sobre la que se están realizando los cálculos es en sí muy pequeña, de manera que los polimorfismos permanecen correlacionados debido a su cercanía en el genoma. Esto queda corroborado por la imputación *in silico* realizada sobre una región de 14Mb, donde los datos obtenidos fueron similares tanto para la informatividad como para la precisión de imputación. Sin embargo, a pesar de dicha correlación subyacente, MiNtagSNP es capaz de seleccionar como tagSNPs los polimorfismos con mayor probabilidad de ser informativos y, por tanto, mayor utilidad en el estudio DGP-M.

A modo de resumen, podemos afirmar que MiNtagSNP es un algoritmo específicamente diseñado para seleccionar un conjunto mínimo de tagSNPs útiles para los análisis de DGP-M por técnicas de NGS.

La aplicación de los tagSNPs seleccionados con MiNtagSNP permite incluir la detección directa de la alteración sin necesidad de realizar pruebas accesorias.

MiNtagSNP necesita seleccionar un menor número de tagSNPs que las técnicas del estado del arte y, además, estos son más informativos.

A pesar de calcular tagSNPs en regiones pequeñas, MiNtagSNP es capaz de seleccionar tagSNPs que son, en promedio, más independientes que los polimorfismos seleccionados por el estado del arte.

Capítulo 5: Comparación entre el PGT-M basado en STRs y SNPs.

5.1 Informatividad por STRs

La Tabla 20 muestra el tamaño de los fragmentos obtenidos en la PCR. El número de repeticiones a los que corresponde cada fragmento (el alelo) se ha indicado para cada individuo en los 3 árboles genealógicos que se incluyen a continuación.

		KG8	SM7	CW2	D24	D21	D99
Pareja 1	Madre	119	145	152	220|232	153|167	184
	Padre	119	137|145	150	234	154|171	185
	E1				230	167	
	E2				220	153	
	E3				232	167	
	E4				234	167	
	E5				232	154	
	E7				230	153|167	
	E9				232	153	
	E10				218|234	151|169	
	E11				232	165	
	E12				234	153|167	
	E13				-	167	
	E14				220|234	169	
Pareja 2	Madre	115|119	145|149	151	226|232	153|171	173|177
	Padre	119	145	144|151	226|234	165	175
	Abuela Pat	119	145|154	144|151	234	167	175
	Abuelo Pat	119	145	151	226|234	153|165	175
	E2			151	-		
	E4			144|151	226|234		
	E6			-	226|232		
	E7			151	-		
	E8			144|151	226|234		
	E10			151	226		
	E12			151	226		
	E14			144|151	226|232		
Pareja 3	Madre	121|128	143	138|152	233|239	153|169	173
	Padre	119	145	133|152	233	164	175
	Tia Pat1	119	145	149|152	228	153|167	182
	Tia Pat2	119	145	149|152	228	153|167	176
	Abuelo Pat	118	145	152	2299	153|167	182
	Hermano Pat	118	142|145	133|152	225|232	153|169	172
	E1			149			
	E2			138|155			
	E3			133			
	E4			137			
	E5			137			
	E6			137			
	E8			133			

Tabla 20: Tamaño de los fragmentos secuenciados para cada STR.

De forma general, en la Figura 42, la Figura 43 y la Figura 44 se ha indicado un guión alto (-) cuando no hubo señal para uno de los alelos. Los casos donde se detectó un alelo extraño no identificado en los parentales se indican con la fuente en rojo. Los alelos presentados en homocigosis en un parental pero cuya herencia a partir de uno u otro cromosoma no es determinable a ciencia cierta se presentan con la fuente negra y sin color de fondo. Los alelos localizados en el cromosoma asociado a la patología se han representado en todas las figuras con el fondo en color rojo.

En la pareja 1, representada en la Figura 42, se encontraron dos STRs informativos, el STR D24 y el D21, ambos localizados aguas arriba del gen PKD1 y separados por 1,5Mb; el STR D24 está localizado a 0,5 Mb del principio del gen. Esta pareja no presentaba ningún familiar afecto por lo que se decidió secuenciar cada embrión por la técnica de Sanger. Posteriormente, uno de los embriones con la alteración fue tomado como referencia y a partir de él se realizó el fasado de toda la familia. Este embrión fue el E2, rodeado por un marco de color rojo en la figura. Como podemos observar, todos los embriones presentaron fallos de amplificación para alguno de los alelos paternos en alguno de los STRs analizados:

- Los embriones E1 y E4 presentaron fallos de amplificación del alelo materno para el STR D24 y el E13 para el paterno también, pero todos presentaron el alelo sano para D21. Basándonos en esto podemos suponer que los tres han heredado el cromosoma materno no portador y por lo tanto serán embriones sanos, pero no podemos asegurar que esto sea cierto al 100% debido a la posibilidad de que se produzca un sobrecruzamiento entre dicho STR y el gen PKD1.
- El embrión E3 presentó fallo de amplificación de los alelos paternos de ambos STRs, pero se detectaron los alelos maternos del cromosoma materno no portador, por lo que podemos llegar a la misma conclusión que con los embriones E1, E4 y E13.
- Los embriones E5 y E9 presentaron el alelo materno asociado al cromosoma portador en D21, pero el alelo materno asociado al cromosoma sano en D24, lo que sugiere que entre estos dos STRs se produjo un sobrecruzamiento, pero habría que realizar prueba extra para confirmar el diagnóstico. Con base en estos resultados podríamos suponer que se trata de embriones sanos, ya que el STR más cercano al gen presentó el alelo sano y a que la probabilidad de observar un doble sobrecruzamiento en un región tan pequeña es muy baja.
- Los embriones E7 y E12 presentaron problemas de amplificación del STR D24, pero el STR D21 mostró el alelo materno asociado al cromosoma no portador, por lo que podríamos suponer que ambos embriones son no portadores de la alteración, pero

nuevamente podría haberse producido un fallo de amplificación entre dicho STR y el inicio del gen que no habría sido detectado.

- El embrión E10 presentó alelos desconocidos, probablemente debidos a fallos de la polimerasa. Razonando lo sucedido en este embrión, parece que se haya producido la terminación prematura del proceso de amplificación, de manera que al sumar una repetición más obtendríamos el alelo 17 en D24 y 27 en D21 asociados al cromosoma portador materno, por lo que podríamos considerar que dicho embrión es portador de la alteración causante de enfermedad en el gen PKD1, sin embargo debería repetirse la prueba para confirmar el resultado.
- El embrión E11 presentó el alelo sano en D24, por lo que podríamos presuponer que será sano no portador.
- El embrión E14 presentó un fallo de amplificación para D21, pero en D24 se detectó el alelo asociado al cromosoma patológico, por lo que cabría esperar que porte también la alteración de PKD1.

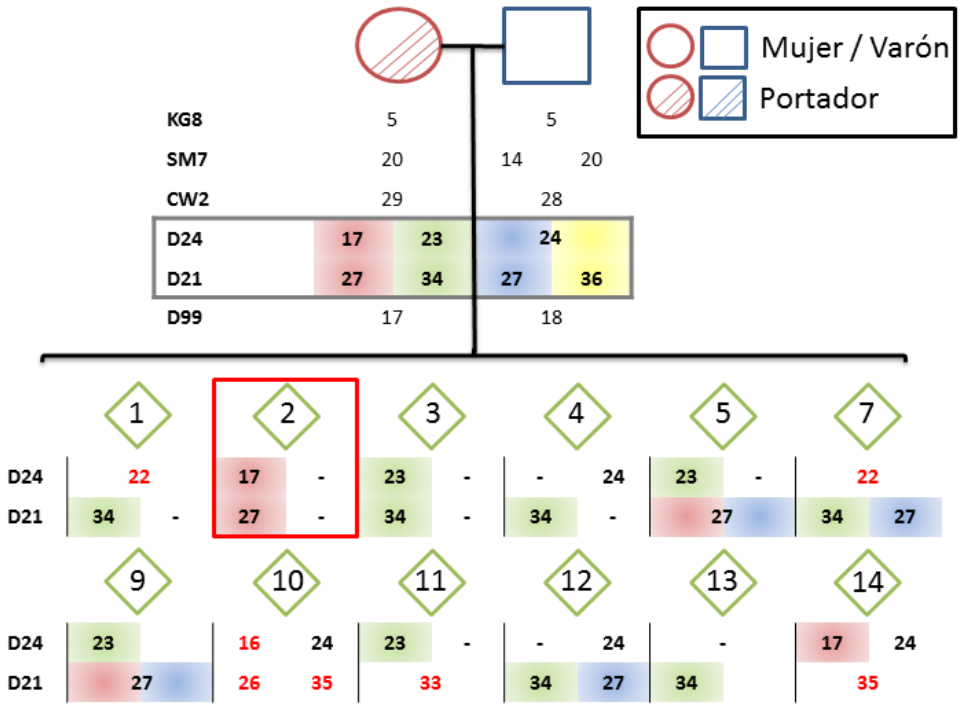

Figura 42: Árbol genealógico correspondiente al empleo de STRs sobre la pareja 1.

En la pareja 2, representada en la Figura 43, la madre del padre (abuela) era portadora de la misma alteración que su hijo. El fondo rojo identifica los alelos asociados al cromosoma causante de dicha alteración mientras que el fondo azul indica los alelos asociados al cromosoma heredado del padre sano. Tan solo 2 STRs resultaron informativos en la pareja principal, D24 y CW2, situado unas 270kb aguas arriba del gen PKD1. Se obtuvieron los siguientes resultados:

- Los embriones E2 y E7 presentaron fallos de amplificación del STR D24, mientras que mostraron el alelo asociado al cromosoma no portador en CW2, por lo que podemos suponer que son embriones sanos, aunque podría haber habido un sobrecruzamiento entre el STR y el gen.
- Los embriones E4 y E8 presentaron los alelos asociados al cromosoma afecto, por lo que podemos considerarlos afectos con casi total seguridad, ya que la probabilidad de una doble recombinación es casi inexistente en una distancia tan pequeña.
- El embrión E6 presentó fallo de amplificación del STR CW2, pero en D24 mostró el alelo asociado al cromosoma no portador por lo que podemos suponer que será sano aunque nuevamente podría haber sufrido un sobrecruzamiento no detectado.

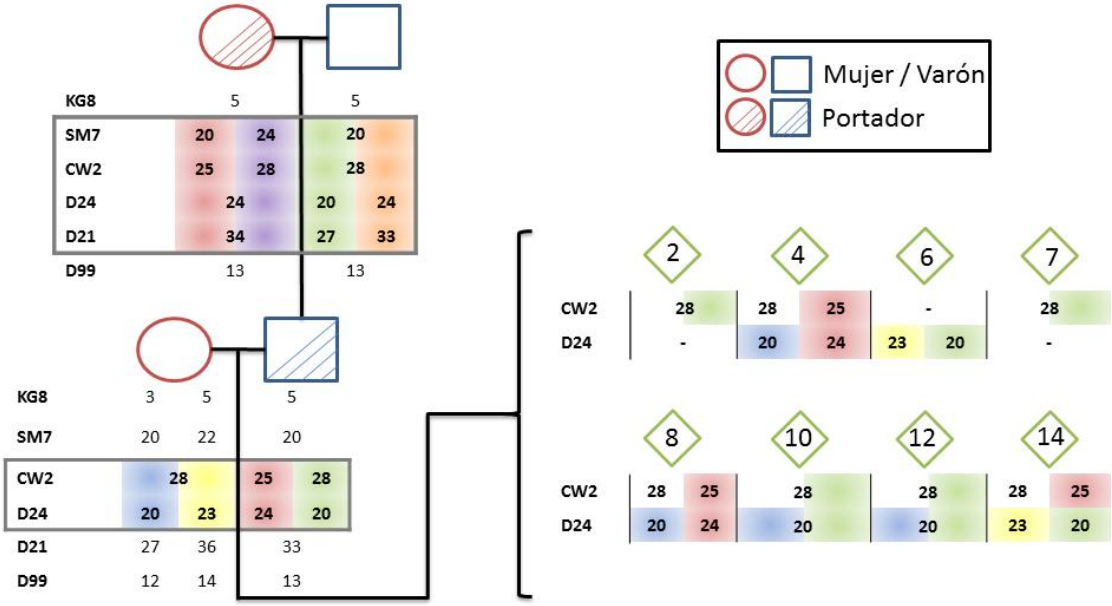

Figura 43: Árbol genealógico correspondiente al empleo de STRs sobre la pareja 2.

- Los embriones E10 y E12 presentaron los alelos asociados al cromosoma paterno no portador.

- El embrión E14 mostró el alelo asociado al cromosoma portador en el STR D24 pero el alelo asociado al cromosoma no portador en CW2, lo que indica que hubo un sobrecruzamiento. Debido a esto no es posible determinar si este embrión será sano o enfermo, pues dependerá de la posición donde dicha recombinación se produjese, de forma que será sano si se produjo entre el gen y el STR CW2, pero será enfermo si se produjo entre el D24 y el gen.

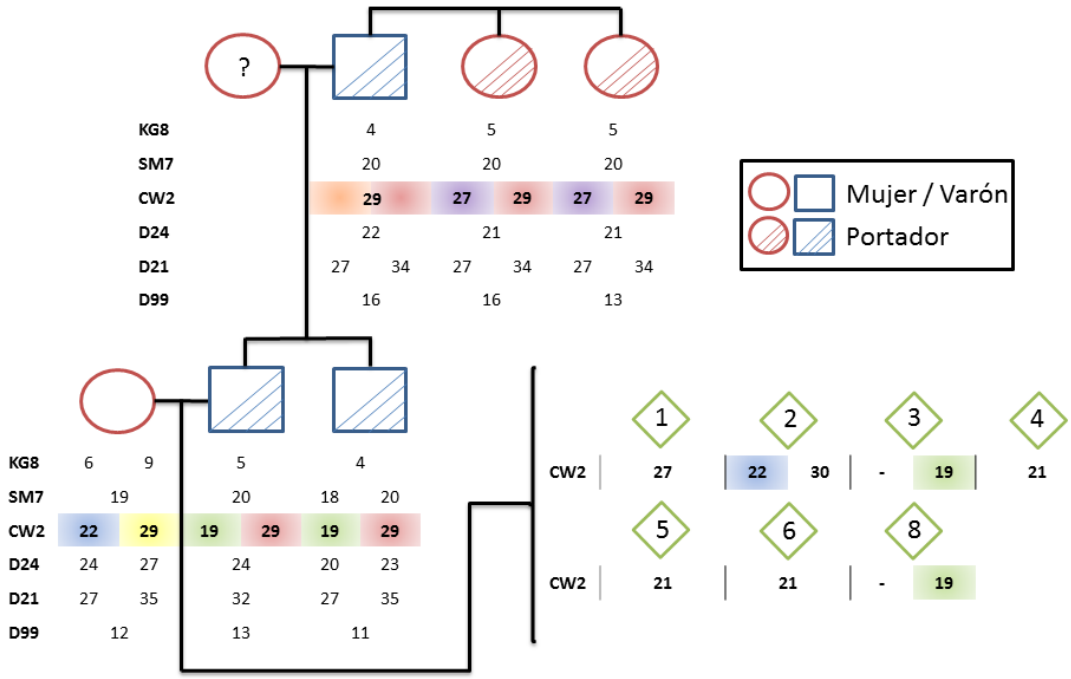

Figura 44: Árbol genealógico correspondiente al empleo de STRs sobre la pareja 3.

Gracias a la gran cantidad de familiares disponibles fue posible determinar el cromosoma portador del alelo patológico en la familia 3, representada en la Figura 44. El único STR informativo encontrado fue CW2. Se obtuvieron los siguientes resultados:

- Los embriones E1, E4, E5 y E6 presentaron alelos extraños. El embrión E1 podría ser un fallo en la amplificación del alelo 29, que procedería bien del padre, en cuyo caso sería un embrión enfermo, o de la madre con un ADO en el alelo paterno, por tanto no es posible identificar su estado. Los embriones E4,E5 y E6 son igualmente imposibles de determinar.
- El embrión E2 presentó un alelo raro que podría proceder de la amplificación errónea del alelo paterno relacionado con el cromosoma portador de patología, aunque no se pudo asegurar.

- Los embriones E3 y E8 presentaron el alelo relacionado con el cromosoma paterno sano.

5.2 Informatividad por MiNtagSNP

A continuación se adjuntan los árboles genealógicos para cada pareja analizada (Figura 45, Figura 46 y Figura 47) y el resultado del fasado. De forma general en las tablas que acompañan a cada caso (Tabla 21, Tabla 22 y Tabla 23) se ha coloreado el fondo de cada polimorfismo de acuerdo al cromosoma paterno de procedencia cuando no hubo duda alguna. La posición de la alteración se ha coloreado con un fondo gris. Los casos compatibles con un efecto ADO se muestran con el fondo sin colorear y la fuente en color rojo. Los tagSNP que fueron *key* pueden identificarse fácilmente observando cada tabla.

En la pareja 1 (Figura 45) pudieron secuenciarse correctamente 169 tagSNPs de los cuales 75 resultaron informativos (44,4%). Los alelos cosegregantes con cada cromosoma pueden consultarse en la Tabla 21. Se obtuvieron los siguientes resultados:

- El embrión E2 mostró la alteración de forma directa, lo que permitió fasar sus cromosomas y emplearlo como referencia para fasar los alelos de los cromosomas maternos y del resto de embriones.
- En los embriones E1, E5, E7, E11 y E13 no fue posible secuenciar los alelos de la posición correspondiente a la alteración, pero fueron determinados nuevamente como embriones sanos portadores al mostrar los alelos que cosegregan en el cromosoma materno no portador. En el embrión E5 se observó una posición que fue interpretada como ADO y que puede ser compatible con un fenómeno de sobrecruzamiento, siendo esa la última posición antes del punto de recombincación. Este resultado es compatible por tanto con el fenómeno de recombinación observado en el apartado anterior al emplear la técnica de STRs.
- El análisis directo de los embriones E3 y E4 no mostro que dichos embriones fuesen portadores, lo cual confirma el resultado del análisis indirecto, ya que mostraron los alelos cosegregantes con el cromosoma no portador.
- El embrión E10 mostró la alteración de forma directa así como los polimorfismos asociados al cromosoma materno portador; además, se observó una posición compatible con fenómeno de recombinación aguas arriba del gen que confirmaría el resultado de la técnica de STRs.
- La secuenciación directa de la alteración en los embriones E9, E12 y E14 tampoco fue satisfactoria. El análisis indirecto de los polimorfismos mostro la existencia de

fenómenos de recombinación aguas arriba de la posición de la alteración dentro del gen PKD1, lo que los convierte en embriones portadores de la alteración.

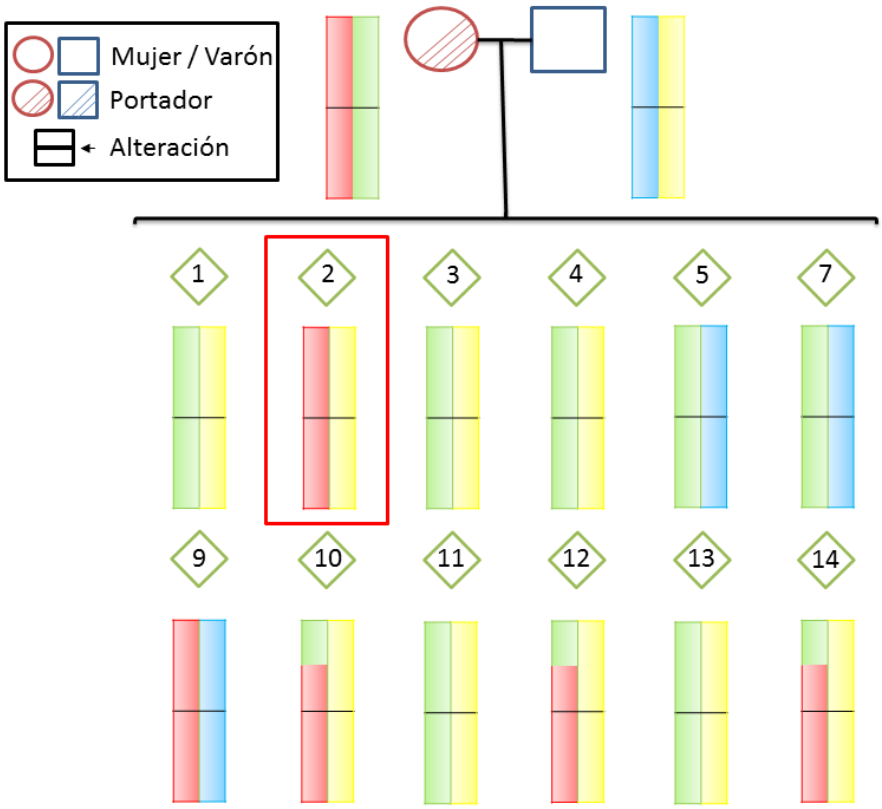

Figura 45: Árbol genealógico correspondiente al empleo de tagSNPs sobre la pareja 1.

Madre		Padre		E1		E2		E3		E4		E5		E7		E9		E10		E11		E12		E13		E14	
G	G	G	A	A	A	A	A	G	A	NC	NC	G	G	NC	NC	G	G	G	A	G	A	NC	NC	A	A	NC	NC
G	G	G	A	A	A	G	A	G	A	G	A	G	G	G	G	NC	NC	G	A	G	A	NC	NC	NC	NC	NC	NC
A	A	A	G	G	G	A	A	A	G	A	G	A	A	A	A	NC	NC	A	G	A	G	NC	NC	NC	NC	NC	NC
A	A	A	G	G	G	NC	NC	A	G	A	G	A	A	A	A	A	A	A	G	A	G	A	A	NC	NC	NC	NC
G	G	G	A	A	A	G	A	G	A	G	A	G	G	G	G	NC	NC	G	A	G	A	NC	NC	NC	NC	NC	NC
A	A	A	C	C	C	A	C	A	A	C	C	A	A	A	A	A	A	A	C	C	C	A	A	C	C	A	A
G	G	G	A	G	A	G	A	G	A	A	A	G	G	NC	NC	NC	NC	NC	NC	NC	NC	G	A	NC	NC	A	A
T	T	T	C	T	C	T	C	T	C	T	C	T	T	T	T	T	T	T	C	T	C	T	T	C	C	T	C
T	T	T	G	NC	NC	T	G	T	G	T	G	T	T	T	T	T	T	NC	NC	G	T	T	G	NC	NC	NC	NC
A	A	A	G	NC	NC	A	G	A	G	A	G	A	A	A	A	A	A	NC	NC	A	G	A	G	NC	NC	NC	NC
T	T	T	C	NC	NC	T	C	T	C	T	C	T	T	T	T	T	T	NC	NC	T	C	T	C	NC	NC	NC	NC
T	T	T	G	T	G	T	G	T	G	T	G	T	T	T	T	T	T	T	G	T	G	T	G	G	G	T	G
C	C	C	G	C	G	C	G	C	G	C	G	C	C	C	C	C	C	C	G	C	G	C	G	G	G	C	G
A	A	G	A	A	G	A	G	G	A	G	A	A	A	A	A	A	A	A	G	G	A	G	A	G	A	G	A

Continúa en la página siguiente

Madre		Padre		E1		E2		E3		E4		E5		E7		E9		E10		E11		E12		E13		E14	
T	T	T	C	T	C	T	C	T	C	T	C	T	T	T	T	T	T	T	C	T	C	T	C	C	C	T	C
A	A	A	G	G	G	A	G	A	G	A	G	A	A	A	A	A	A	A	G	A	G	A	G	G	G	A	G
A	A	A	C	NC	NC	A	C	NC	NC	NC	NC	A	A	NC	NC	A	A	NC	NC	A	A	A	A	NC	NC	A	C
T	T	T	C	NC	NC	T	C	NC	NC	NC	NC	T	T	NC	NC	T	T	NC	NC	T	T	T	T	NC	NC	T	C
A	A	A	G	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	G	G	NC	NC	NC	NC	NC	NC
G	G	G	A	NC	NC	A	A	A	A	G	A	NC	NC	G	G	NC	NC	G	A	G	A	A	A	A	A	G	A
G	G	G	C	C	C	G	C	G	G	NC	NC	G	G	NC	NC	G	G	NC	NC	G	G	G	C	C	C	C	C
C	C	C	T	C	T	C	T	C	T	C	T	NC	NC	C	C	C	C	C	T	C	C	C	T	T	T	C	T
T	T	T	C	T	C	T	C	C	C	T	T	T	T	T	T	T	T	T	C	T	C	T	C	C	C	T	C
T	T	T	G	T	G	T	-	T	-	T	T	T	T	T	T	T	T	T	-	T	T	T	-	-	-	T	-
T	T	G	T	T	T	T	T	T	T	G	T	T	G	T	G	T	G	T	T	T	T	T	T	T	T	T	T
T	T	G	T	T	T	NC	NC	NC	NC	NC	NC	NC	NC	G	G	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC
C	C	A	C	NC	NC	NC	NC	NC	NC	NC	NC	C	A	NC	NC	A	A	NC	NC	NC	NC	C	C	NC	NC	C	C
T	T	C	T	T	T	T	T	T	T	T	T	T	C	NC	NC	T	C	T	T	T	T	T	T	T	T	T	T
C	C	T	C	C	C	C	C	C	C	C	C	C	T	NC	NC	C	T	C	C	C	C	C	C	C	C	C	C
T	T	C	T	T	T	T	T	T	T	T	T	T	C	NC	NC	T	C	T	T	T	T	T	T	T	T	T	T
T	T	C	T	T	T	T	T	T	T	T	T	T	C	NC	NC	T	C	T	T	T	T	T	T	T	T	T	T
C	C	T	C	C	C	C	C	C	C	C	C	C	T	C	T	C	T	C	C	C	C	C	C	C	C	C	C
T	T	C	T	T	T	T	T	T	T	T	T	T	C	T	T	C	C	T	T	T	T	T	T	T	T	T	T
T	T	C	T	T	T	T	T	T	T	T	T	T	C	T	C	T	C	T	T	T	T	T	T	T	T	T	T
G	G	T	G	G	G	G	G	G	G	G	G	G	T	G	T	G	T	G	G	G	G	G	G	G	G	G	G
A	A	C	A	A	A	A	A	A	A	A	A	A	C	A	C	A	C	A	A	A	A	A	A	A	A	A	A
G	G	C	G	G	G	G	G	G	G	G	G	G	C	G	C	G	C	G	G	G	G	G	G	G	G	G	G
C	C	A	C	C	C	C	C	C	C	C	C	C	A	C	A	C	A	C	C	C	C	C	C	C	C	C	C
A	A	C	A	A	A	A	A	A	A	A	A	A	C	A	C	A	C	A	A	A	A	A	A	A	A	A	A
G	G	C	G	G	G	G	G	G	G	G	G	G	C	G	C	G	C	G	G	G	G	G	G	G	G	G	G
C	C	G	C	C	C	C	C	C	C	C	C	G	G	C	C	C	G	C	C	C	C	C	C	C	C	C	C
T	T	G	T	T	T	T	T	T	T	T	T	T	G	T	G	T	G	T	T	T	T	T	T	T	T	T	T
A	A	G	A	A	A	NC	NC	NC	NC	A	A	NC	NC	G	G	G	G	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC
G	G	G	A	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC
G	G	G	A	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC
A	A	A	G	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC
G	G	A	G	NC	NC	G	G	G	G	NC	NC	A	A	NC	NC	G	A	NC	NC	G	G	NC	NC	NC	NC	G	G
A	T	A	A	A	A	A	A	A	A	A	A	A	A	A	A	T	A	T	A	A	A	T	A	A	A	T	A
G	C	C	C	C	C	G	C	C	C	C	C	C	C	C	C	G	C	NC	NC	C	C	G	C	C	C	NC	NC
A	G	G	G	G	G	A	G	G	G	G	G	G	G	G	G	NC	NC	NC	NC	G	G	G	G	G	G	G	G
G	C	G	G	C	C	G	G	C	C	C	G	NC	NC	NC	NC	G	G	G	G	C	C	G	G	NC	NC	G	G
C	T	C	C	T	C	NC	NC	T	C	T	T	NC	NC	C	C	C	C	C	C	C	C	C	C	C	C	C	C

Continúa en la página siguiente

Madre		Padre		E1		E2		E3		E4		E5		E7		E9		E10		E11		E12		E13		E14	
G	T	T	T	T	T	G	T	T	T	T	T	T	T	T	T	G	T	G	T	T	T	T	T	T	T	NC	NC
G	A	G	G	G	G	G	G	A	G	A	G	A	G	A	G	G	G	G	G	A	G	G	G	G	G	G	G
A	G	G	G	NC	NC	A	G	G	G	G	G	G	G	NC	NC	G	G	A	G	G	G	A	G	G	G	A	G
G	C	C	C	NC	NC	NC	NC	C	C	NC	NC	NC	NC	C	C	C	C	NC	NC	NC	NC	C	C	NC	NC	C	C
A	G	G	G	NC	NC	A	G	G	G	G	G	NC	NC	NC	NC	NC	NC	A	G	NC	NC	NC	NC	NC	NC	NC	NC
C	T	T	T	T	T	C	T	T	T	T	T	T	T	T	T	C	C	NC	NC	T	T	C	T	NC	NC	C	T
C	T	C	C	C	C	C	C	NC	NC	C	C	C	C	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	C	C
T	C	C	C	C	C	T	C	NC	NC	C	C	C	C	NC	NC	NC	NC	NC	NC	NC	NC	T	T	NC	NC	T	C
C	T	T	T	T	T	C	T	T	T	T	T	T	T	T	T	NC	NC	NC	NC	T	T	C	T	T	T	C	T
A	T	T	T	T	T	A	T	T	T	T	T	T	T	T	T	NC	NC	NC	NC	T	T	A	T	T	T	A	T
C	T	T	T	T	T	C	T	T	T	T	T	T	T	T	T	NC	NC	NC	NC	T	T	C	T	T	T	C	T
G	C	C	C	C	C	G	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	G	C	C	C	G	C
G	A	A	A	A	A	G	A	A	A	A	A	A	A	A	A	G	A	G	A	A	A	A	A	A	A	G	G
C	G	G	G	G	G	C	G	G	G	G	G	G	G	G	G	C	G	C	G	G	G	G	G	G	G	C	C
G	A	A	A	A	A	G	A	A	A	A	A	A	A	NC	NC	G	A	NC	NC	A	A	G	A	A	A	G	A
C	G	G	G	G	G	C	G	G	G	G	G	G	G	NC	NC	C	G	C	G	G	G	C	G	G	G	C	G
T	C	C	C	C	C	T	C	C	C	C	C	C	C	C	C	T	C	NC	NC	C	C	T	C	C	C	T	C
T	C	T	T	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	T	T	NC	NC
A	G	G	G	NC	NC	A	G	G	G	G	G	G	G	G	G	A	G	-	G	NC	NC	A	G	G	G	A	A
G	A	A	A	NC	NC	G	A	NC	NC	A	A	A	A	A	A	G	A	G	A	A	A	G	A	NC	NC	A	G
G	A	A	A	A	A	G	A	A	A	A	A	A	A	A	A	G	A	G	A	A	A	G	A	A	A	A	G
C	T	T	T	T	T	C	T	T	T	T	T	T	T	T	T	C	T	C	T	T	T	C	T	T	T	T	C
C	T	T	T	T	T	C	T	T	T	NC	NC	NC	NC	T	T	C	T	NC	NC	T	T	C	T	T	T	T	C
T	C	C	C	C	C	T	C	C	C	C	C	C	C	C	C	T	C	T	C	C	C	C	C	C	C	C	T

Tabla 21: Alelos cosegregantes con cada cromosoma de la pareja1.

Para la pareja 2 no se realizó análisis directo de la mutación pues los cromosomas fueron fasados a partir de los familiares disponibles. Para la pareja de abuelos se secuenciaron con éxito 130 polimorfismos de los cuales 70 tagSNP (53,4%) fueron informativos. Para la pareja principal se pudieron secuenciar 158 polimorfismos de los cuales 70 (44,3%) fueron informativos. Los polimorfismos ara cada cromosoma quedan reflejados en la Tabla 22. Se obtuvieron los siguientes resultados:

- Los embriones E2 y E7 presentaron polimorfismos compatibles con fenómenos de recombinación, sin embargo ambos pueden ser determinados como sanos no portadores debido a que el sobrecruzamiento con el cromosoma paterno portador tuvo lugar aguas arriba del gen PKD1.

- El embrión E14, presentó polimorfismos compatibles con una recombinación aguas arriba del gen, convirtiéndolo en portador de la alteración.
- Los embriones E4 y E8 mostraron los polimorfismos segregantes con el cromosoma portador, por lo que pueden considerarse portadores de la alteración.
- Los embriones E6, E10 y E11 sin embargo mostraron los polimorfismos asociados al cromosoma sano no portador, aunque se observó una posición extraña en medio de la secuencia del embrión E6 que no concuerda con el resto del cromosoma. Esta posición podría ser resultado de una mutación espontánea o bien un fallo de la secuenciación.

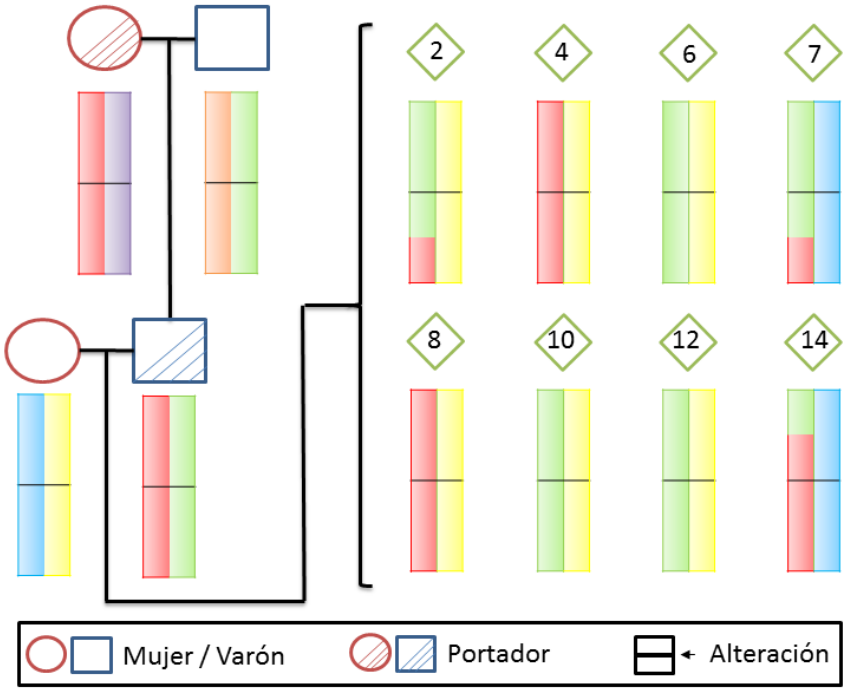

Figura 46: Árbol genealógico correspondiente al empleo de tagSNPs sobre la pareja 2.

Abuela		Abuelo		Padre		Madre		E2		E3		E4		E5		E6		E7		E8		E9	
C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
NC	NC	NC	NC	G	G	A	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
C	T	T	C	C	C	T	C	C	C	C	C	C	C	C	C	NC	NC	NC	NC	C	C	C	T
G	A	G	G	G	G	G	A	NC	NC	G	A	G	A	G	A	G	A	G	G	G	A	G	G
G	A	G	G	G	G	G	A	NC	NC	G	A	G	A	G	A	G	A	G	G	G	A	G	G
T	G	T	T	T	T	T	G	NC	NC	G	G	T	G	G	G	NC	NC	NC	NC	T	G	T	T

Continúa en la página siguiente

Abuela		Abuelo		Padre		Madre		E2		E3		E4		E5		E6		E7		E8		E9	
C	T	C	C	C	C	C	T	NC	NC	NC	NC	C	T	NC	NC	C	T	C	C	C	T	C	C
NC	NC	NC	NC	T	C	T	C	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC
G	G	G	G	G	G	G	T	NC	NC	G	G	G	T	G	G	G	T	G	G	G	T	G	G
G	C	G	G	G	G	C	G	NC	NC	G	G	G	G	G	G	G	G	G	C	G	G	G	C
A	G	A	A	A	A	A	G	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC
C	C	C	C	C	C	C	C	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	C	C	NC	NC
C	C	C	C	C	C	C	C	NC	NC	NC	NC	C	C	C	C	NC	NC	C	C	C	C	C	C
G	G	G	G	G	G	G	G	NC	NC	NC	NC	G	G	G	G	NC	NC	G	G	G	G	G	G
C	C	C	C	C	C	C	T	NC	NC	NC	NC	NC	NC	T	T	NC	NC	NC	NC	NC	NC	NC	NC
G	G	G	G	G	G	G	C	NC	NC	G	C	G	C	G	C	G	C	G	G	G	C	G	G
A	A	A	A	A	A	A	G	NC	NC	NC	NC	A	G	A	G	A	G	NC	NC	NC	NC	A	A
C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
G	A	A	A	G	A	A	G	G	G	A	G	A	G	A	G	A	G	A	A	G	G	A	A
G	G	G	A	G	A	A	G	G	G	A	G	A	G	A	G	A	G	A	A	G	G	A	A
G	A	A	A	G	A	A	G	G	G	A	G	A	G	A	G	A	G	A	A	G	G	A	A
A	G	G	G	A	G	G	A	NC	NC	G	A	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	G	G
G	C	C	C	G	C	C	G	NC	NC	NC	NC	C	G	NC	NC	NC	NC	NC	NC	G	G	NC	NC
C	A	NC	NC	C	A	C	C	C	C	A	C	A	C	A	C	NC	NC	A	C	C	C	NC	NC
C	T	T	T	C	T	C	T	T	T	T	C	T	C	T	C	T	C	NC	NC	C	C	T	T
C	C	C	T	C	T	T	C	NC	NC	T	C	C	C	T	C	T	C	T	T	C	C	T	T
T	T	NC	NC	T	C	C	T	NC	NC	C	T	T	T	C	T	C	T	C	C	T	T	C	C
T	T	T	C	T	C	C	T	C	T	C	T	T	T	C	T	C	T	C	C	T	T	C	C
T	T	T	C	T	C	C	T	NC	NC	C	T	T	T	C	T	C	T	C	C	T	T	C	C
C	C	C	T	C	T	T	C	T	T	T	C	C	C	T	C	T	C	T	T	C	C	T	T
C	C	C	C	C	C	C	C	NC	NC	C	C	C	C	C	C	C	C	C	C	C	C	C	C
A	A	G	G	A	G	G	A	NC	NC	G	A	A	A	G	A	G	A	G	G	A	A	G	G
T	T	C	C	T	C	C	T	NC	NC	NC	NC	T	T	NC	NC	C	T	C	C	T	T	C	C
A	A	NC	G	A	G	G	A	NC	NC	G	A	A	A	G	A	G	A	G	G	A	A	G	G
T	T	C	C	T	C	C	T	NC	NC	C	T	T	T	C	T	C	T	C	C	T	T	C	C
C	C	G	G	C	G	G	C	NC	NC	G	C	C	C	G	C	G	C	G	G	C	C	G	G
A	G	G	G	A	G	A	A	G	A	G	A	A	A	G	A	G	A	G	A	A	A	G	A
G	A	A	G	G	G	G	G	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	G	G	NC	NC
C	T	T	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
T	C	C	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
C	T	T	C	C	C	C	C	C	C	C	C	C	C	NC	NC	NC	NC	C	C	C	C	C	C
T	T	NC	NC	T	T	T	T	NC	NC	T	T	T	T	T	T	NC	NC	T	T	T	T	T	T
C	G	G	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
T	C	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T

Continúa en la página siguiente

Abuela		Abuelo		Padre		Madre		E2		E3		E4		E5		E6		E7		E8		E9	
C	T	T	C	C	C	C	T	C	T	C	T	C	T	C	T	C	T	C	C	C	T	C	C
T	C	NC	NC	C	T	C	T	NC	NC	T	T	T	C	T	T	T	T	T	C	T	C	T	C
C	C	C	C	C	C	T	C	C	C	C	C	C	C	C	C	C	C	C	T	C	C	C	T
C	T	T	C	C	C	T	T	NC	NC	C	T	C	T	C	T	C	T	C	T	C	T	C	T
G	A	G	G	G	G	G	A	G	A	G	A	G	A	G	A	G	A	G	G	G	A	G	G
G	A	A	G	G	G	A	A	NC	NC	G	G	G	A	A	G	A	G	A	G	A	G	A	G
T	T	T	C	T	C	T	C	NC	NC	C	C	T	C	NC	NC	NC	NC	NC	NC	T	C	T	C
C	T	C	C	C	C	T	C	NC	NC	NC	NC	C	C	C	C	C	C	C	T	C	C	C	T
T	T	C	C	T	C	C	C	C	C	C	C	T	C	C	C	C	C	C	C	T	C	C	C
A	A	G	G	A	G	A	A	A	A	G	A	A	A	G	A	G	A	G	A	A	A	G	A
C	C	C	C	C	C	C	C	T	T	C	C	C	C	C	C	C	C	C	C	C	C	C	C
G	G	G	A	G	A	A	G	NC	NC	NC	NC	G	A	NC	NC	NC	NC	A	G	A	G	A	G
T	T	C	C	T	C	T	T	NC	NC	C	T	T	T	C	T	C	T	C	T	T	T	C	T
A	T	T	T	A	T	T	T	NC	NC	T	T	A	T	T	T	NC	NC	T	T	A	T	T	T
T	G	G	G	T	G	G	G	NC	NC	G	G	T	G	G	G	NC	NC	NC	NC	T	G	G	G
G	A	A	A	G	A	A	A	NC	NC	A	A	G	A	A	A	A	A	A	A	G	A	A	A
C	C	NC	NC	G	C	G	C	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC
G	G	G	G	G	G	G	G	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	G	G	NC	NC	NC	NC
T	C	C	C	T	C	C	T	NC	NC	C	T	T	T	C	T	C	T	C	C	T	T	C	C
A	G	NC	NC	G	A	A	G	NC	NC	NC	NC	G	G	NC	NC	NC	NC	A	A	G	G	A	A
T	T	T	T	T	T	T	T	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC
G	-	G	G	G	G	G	G	NC	NC	G	G	G	G	G	G	NC	NC	G	G	G	G	G	G
T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
C	T	T	T	C	T	T	T	T	T	T	T	C	T	T	T	T	T	T	T	C	T	T	T
G	G	G	G	G	G	A	G	NC	NC	G	A	G	A	G	A	G	A	G	G	G	A	G	G
A	A	A	A	A	A	A	G	A	A	A	G	A	G	A	G	A	G	A	A	A	G	A	A
T	T	T	T	T	T	T	G	NC	NC	T	G	T	G	T	G	T	G	T	T	T	G	T	T
G	C	NC	NC	G	C	C	G	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC
G	G	G	G	G	G	A	G	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC
C	C	C	C	C	C	A	C	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	C	A
NC	NC	G	G	G	G	C	G	NC	NC	G	G	G	G	NC	NC	NC	NC	G	C	G	G	G	C
C	C	C	C	C	C	G	C	NC	NC	C	C	C	C	C	C	C	C	C	G	C	C	C	G
G	G	G	A	G	A	A	G	NC	NC	G	A	G	G	G	A	G	A	A	A	G	G	A	A
G	G	G	G	G	G	T	G	NC	NC	G	G	G	G	NC	NC	NC	NC	T	T	G	G	NC	NC
C	C	C	T	C	T	C	C	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC
G	G	G	G	G	G	G	G	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC
T	T	C	C	T	C	T	C	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC
T	T	C	C	T	C	C	T	NC	NC	T	C	T	T	T	C	T	C	C	C	T	T	T	C

Continúa en la página siguiente

Abuela	Abuelo	Padre	Madre	E2	E3	E4	E5	E6	E7	E8	E9
G G	G G	G G	T G	NC NC	G G	G G	NC NC	G G	NC NC	NC NC	G T
G G	G G	G G	A G	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC	G A
C C	C T	C T	T C	NC NC	T C	C C	T C	T C	T T	C C	C C
G G	G G	G G	G G	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC
NC NC	NC NC	G C	G G	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC
C G	C C	C C	C C	C C	C C	C C	C C	C C	C C	C C	C C
C T	C C	C C	C C	NC NC	C C	C C	C C	C C	C C	C C	C C
G A	G G	G G	G G	NC NC	G G	G G	G G	G G	G G	NC NC	G G
C T	C C	C C	C C	NC NC	C C	C C	C C	C C	C C	C C	C C
T C	T T	T T	T T	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC
G A	G G	G G	G G	NC NC	G G	G G	G G	NC NC	NC NC	G G	NC NC
T T	T C	T C	C C	C C	C C	T C	C C	C C	C C	T C	T C
G G	G A	G A	A A	NC NC	A A	G A	A A	G A	A A	G A	G A
T T	T G	T G	G G	NC NC	G G	T G	G G	G G	G G	T G	T G
C C	NC NC	C C	G G	G G	G G	C G	G G	G G	NC NC	G C	C C
G G	G T	G T	T G	G G	T G	G G	T G	G T	T T	G G	G T
T T	C T	T T	T T	NC NC	NC NC	T T	NC NC	NC NC	T T	T T	T T
C T	T T	C T	T T	NC NC	T T	C T	T T	T T	T T	C T	C T
T T	T T	T T	T T	NC NC	NC NC	T T	NC NC	NC NC	NC NC	T T	T T
G G	G G	G G	G G	NC NC	G G	G G	G G	NC NC	NC NC	G G	G G
C C	C C	C C	C C	C C	C C	C C	C C	C C	C C	C C	C C
G G	G G	G G	G G	G G	NC NC	G G	NC NC	NC NC	NC NC	G G	NC NC
A A	A C	A C	A A	C C	C A	A A	C A	C A	A A	A A	A A
T T	T C	T C	T T	C C	C T	T T	C T	C T	T T	T T	T T
G G	G G	G G	G G	NC NC	G G	G G	NC NC	G G	A G	G G	G G
A A	A T	A T	A A	NC NC	T A	A A	T A	T A	A A	A A	A A
G G	G A	G A	G G	NC NC	A G	G G	A G	A G	G G	G G	G G
G A	G G	G G	A A	NC NC	G A	G A	G A	A G	A G	A G	A G
A G	A A	A A	G G	NC NC	A G	A G	A G	G A	G A	G A	G A
G A	G A	G A	A G	NC NC	G A	G G	A A	NC NC	G A	G G	G A
NC NC	NC NC	C T	C C	NC NC	T C	C C	T C	T C	C C	C C	C C
NC NC	NC NC	G A	G A	NC NC	A A	G A	A A	A A	G G	G A	G G
T T	T C	T C	T C	C C	C C	T C	C C	C C	T T	T C	T T
C T	T T	C T	C T	NC NC	T C	C C	T C	T C	T C	C C	T C
C T	NC NC	C T	C T	NC NC	T T	T T	NC NC	NC NC	C T	T T	NC NC
NC NC	NC NC	A C	C C	C C	C C	A C	C C	NC NC	A C	A C	A C
G C	C C	G C	C C	C C	C C	G C	C C	C C	G C	G C	G C

Continúa en la página siguiente

Abuela	Abuelo	Padre	Madre	E2	E3	E4	E5	E6	E7	E8	E9
C T	T T	C T	T T	NC NC	T T	C T	T T	T T	C T	C T	C T
C T	T T	C T	T T	T T	T T	C T	T T	T T	C T	C T	C T
C T	NC NC	C T	C C	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC
C T	T T	C T	T T	NC NC	T T	C T	T T	T T	C T	C T	C T
C G	G G	C G	G G	NC NC	G G	C G	NC NC	NC NC	C G	C G	C G
T C	C C	T C	C C	NC NC	C C	T C	C C	C C	T C	T C	T C
T T	G G	T G	G T	NC NC	G T	T T	G T	T G	T G	T T	T G
A G	G G	A G	G A	NC NC	G A	A A	NC NC	A G	NC NC	A A	A G
T C	C C	T C	C T	NC NC	C C	NC NC	NC NC	NC NC	NC NC	NC NC	C C
T C	NC NC	C T	T C	NC NC	T C	C C	T C	T C	T C	C C	T C
NC NC	A A	G A	A G	NC NC	A A	G G	A G	NC NC	NC NC	G G	A G
A A	G G	A G	G A	NC NC	NC NC	A A	NC NC	A G	A G	NC NC	A G
NC NC	NC NC	A G	G A	NC NC	A G	A A	A G	A G	A G	A A	A G
C C	T T	C T	T C	NC NC	C T	C C	C T	C T	C T	C C	C T
T C	T C	T C	T T	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC	T C
A G	G G	A G	G G	G G	G G	A G	G G	G G	A G	A G	A G
G C	C C	G C	C C	NC NC	NC NC	G C	C C	NC NC	NC NC	NC NC	G C
A G	G G	A G	G G	A G	G G	A G	G G	G G	A G	NC NC	A G
G A	NC NC	A G	G G	G G	G G	A G	G G	NC NC	A G	A G	A G
T G	G G	T G	G G	T G	G G	T G	G G	G G	T G	T G	T G
T C	C C	T C	C C	C C	C C	T C	C C	C C	T C	T C	T C
A G	G A	A A	G G	A A	A G	A G	A G	G A	A G	A G	G A
T G	NC NC	G T	T T	NC NC	NC NC	G T	NC NC	NC NC	G T	NC NC	T G
A C	C C	A C	C C	A C	C C	A C	C C	C C	A C	A C	A C
T T	T T	T T	T T	T T	T T	T T	T T	T T	T T	T T	T T
A G	NC NC	G A	A A	A A	A A	G A	A A	NC NC	G A	G A	G A
C G	G G	C G	G G	C C	G G	C G	G G	G G	C G	C G	C G
A G	NC NC	A A	G G	NC NC	G G	A G	G G	NC NC	A G	A G	G A
C C	C T	C T	C C	NC NC	T C	NC NC	T C	NC NC	NC NC	NC NC	NC NC
C T	NC NC	T C	C C	NC NC	C C	T C	C C	C C	T C	T C	T C
T C	C C	T C	C C	NC NC	C C	T T	C C	C C	T C	T T	T C
G A	NC NC	G G	A A	NC NC	G G	G G	G G	NC NC	G A	G G	G A
A A	A A	A A	A A	NC NC	A A	A A	A A	A A	A A	A A	A A
NC NC	NC NC	C C	C C	NC NC	C C	C C	NC NC	NC NC	C C	C C	C C
G A	A G	G G	A G	NC NC	G G	G G	G G	G G	G A	G G	G A
G G	C G	G G	C C	NC NC	G C	G C	G C	C G	G C	C G	C G
T T	NC NC	T T	C C	T T	T C	C T	T C	C T	NC NC	C T	C T
G G	G A	G A	G A	NC NC	A A	A G	NC NC	NC NC	G G	A G	G G
G C	NC NC	G C	C G	NC NC	NC NC	G G	G G	NC NC	C C	G C	C C

Tabla 22: Alelos cosegregantes con cada cromosoma de la pareja2.

Los alelos de la pareja 3 fueron fasados a partir de los familiares disponibles. La Tabla 23 muestra los alelos segregantes con cada cromosoma. Los resultados obtenidos fueron los siguientes:

- Los embriones E2, E3, E4, E6 y E8 presentaron los alelos asociados al cromosoma paterno no portador. El embrión E8 mostró resultados compatibles con un fenómeno de recombinación de los cromosomas materno.
- El embrión E5 mostró resultados compatibles con una recombinación de los cromosomas paternos, pero aguas arriba de la posición de la alteración por lo que no es portador de la misma.
- El embrión E1 mostró los alelos que cosegregan con el cromosoma portador, por lo que puede considerarse afecto por la patología. La Tabla 23 muestra el patrón de alelos segregantes.

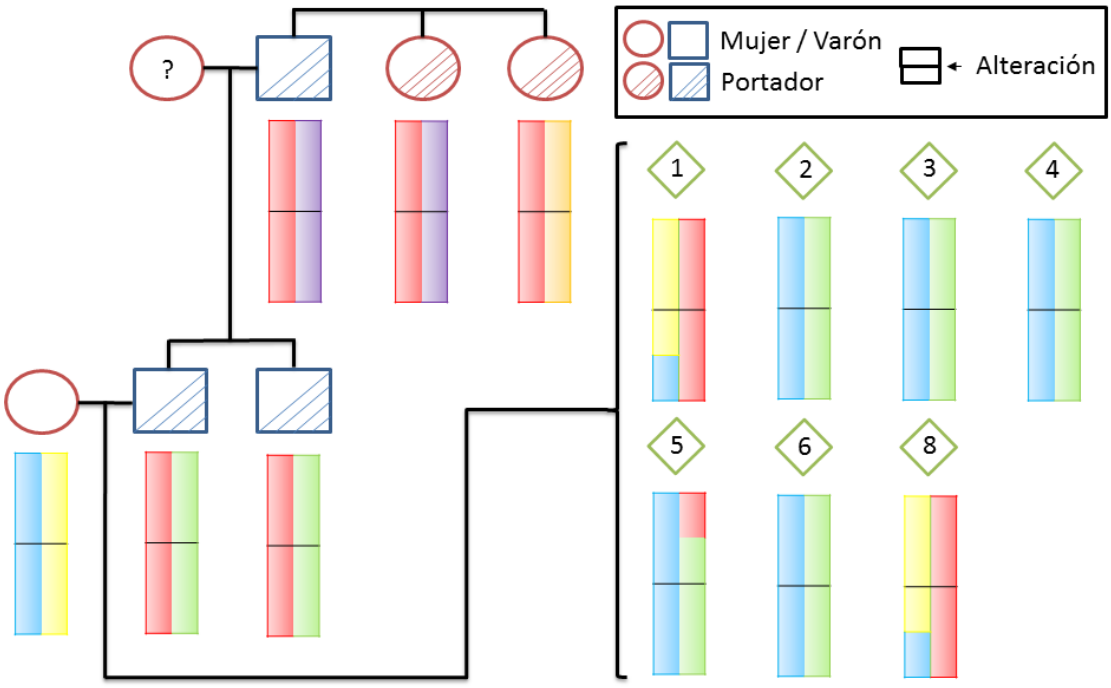

Figura 47: Árbol genealógico correspondiente al empleo de tagSNPs sobre la pareja 3.

Abuelo	Hermano	Tia1	Tia2	Madre	Padre	e1	e2	e3	e4	e5	e6	e8
G G	G C	G G	G G	C C	G C	NC NC	NC NC	C C	C C	NC NC	C C	C C
NC NC	A G	A A	A A	G G	A G	G A	G G	G G	G G	G G	G G	G G
C T	C C	C T	C T	T C	C C	NC NC	C C	NC NC	C C	C C	C C	T C
NC NC	C C	C C	C C	C A	C C	C C	A C	C C	A C	A A	A C	C C
G A	G G	G A	G A	A G	G G	A G	G G	G G	G G	G G	G G	A G

Continúa en la página siguiente

Abuelo	Hermano	Tia1	Tia2	Madre	Padre	e1	e2	e3	e4	e5	e6	e8
G A	G G	G A	G G	A G	G G	A G	G G	G G	G G	G G	G G	A G
G G	G T	G G	G G	G T	G T	NC NC	T T	NC NC	T T	T T	NC NC	NC NC
T T	T C	T T	T T	T C	T C	NC NC	C C	C C	C C	C C	C C	NC NC
T C	T C	T C	T C	T T	C T	NC NC	T T	NC NC	NC NC	NC NC	NC NC	NC NC
NC NC	C A	C A	C A	A C	C A	C A	C A	A A	C A	C C	C C	A A
T T	T G	T T	T G	T G	T G	G T	NC NC	G G	NC NC	T T	T G	G G
C C	C G	C C	C G	G G	C G	NC NC	G G	NC NC	G G	G G	G G	G G
G A	G A	G G	G A	A A	G A	NC NC	A A	NC NC	NC NC	NC NC	NC NC	NC NC
A A	A G	A A	A G	G G	A G	NC NC	NC NC	G G	G G	NC NC	NC NC	NC NC
G G	G A	G G	G A	A A	G A	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC
A G	A G	A G	A A	A A	A G	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC
G A	G A	G A	G A	A A	G A	NC NC	NC NC	NC NC	NC NC	A A	NC NC	NC NC
A G	A A	A G	A A	A A	A A	A A	NC NC	NC NC	A A	A A	A A	A A
A G	A G	A G	A A	A G	A G	A A	G G	G G	G G	G G	G G	A A
A G	A A	A G	A A	A A	A A	A A	A A	A A	A A	A A	A A	A A
G A	G G	G A	G G	G G	G G	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC
C G	C C	C G	C C	C C	C C	NC NC	NC NC	NC NC	NC NC	C C	NC NC	NC NC
A C	A A	A C	A C	A A	A A	NC NC	A A	A A	A A	NC NC	NC NC	NC NC
T C	T T	T C	T T	T T	T T	NC NC	T T	T T	NC NC	T T	NC NC	NC NC
C C	C T	C C	C T	T T	C T	T C	T T	T T	T T	T T	T T	T T
T T	T C	T T	T C	C C	T C	C T	C C	C C	C C	C C	C C	C C
T T	T C	T T	T C	C C	T C	C T	C C	NC NC	C C	C C	C C	NC NC
C C	C T	C C	C T	T T	C T	T C	T T	NC NC	T T	T T	T T	T T
G A	G G	G A	G G	G G	G G	NC NC	G G	G G	G G	G G	G G	G G
C T	C C	C T	C C	C C	C C	NC NC	C C	NC NC	C C	C C	NC NC	C C
A A	A G	A A	A G	G G	A G	NC NC	G G	G G	G G	G G	NC NC	NC NC
C T	C C	C T	C C	C C	C C	C C	C C	C C	C C	C C	C C	C C
G C	G G	G C	G G	G G	G G	G G	G G	G G	G G	G G	G G	G G
A G	A G	A G	A G	A G	A G	A G	G G	G G	G G	G G	G G	A A
G A	G A	G A	G A	G A	G A	NC NC	NC NC	NC NC	G G	NC NC	NC NC	G G
C T	C T	C T	C T	C T	C T	C T	C T	C T	C T	T T	C C	C C
C C	C T	C C	C T	C C	C T	C C	C T	NC NC	C C	C C	C T	C T
C T	C T	C T	C T	C T	C T	NC NC	C T	NC NC	C T	T T	T T	NC NC
NC NC	T T	T C	T T	T C	T T	NC NC	C T	C T	C C	NC NC	C T	NC NC
C T	C C	C T	C C	C T	C C	NC NC	T C	T C	C C	T T	T T	NC NC
C C	C C	C C	C C	C G	C C	C C	G C	NC NC	G C	G G	G C	C C
A A	A A	A A	A A	A G	A A	A A	G A	NC NC	G A	G G	G A	A A
T C	T C	T C	T C	C T	C T	C C	T T	T T	T T	T T	T T	C C

Continúa en la página siguiente

Abuelo	Hermano	Tia1	Tia2	Madre	Padre	e1	e2	e3	e4	e5	e6	e8
T T	T T	T T	T T	C C	T T	NC NC	C T	T T	C T	C C	C T	C T
C C	C C	C C	C C	C T	C C	C C	T C	NC NC	T T	T T	T T	C C
C C	C C	C C	C T	C C	C C	C C	C C	C C	C C	C C	C C	C C
G G	G A	G G	G A	A G	G A	A G	A G	A A	G G	G G	A G	A A
G G	G A	G G	G A	G A	A G	G G	NC NC	NC NC	A G	G G	G G	NC NC
C C	C T	C C	C C	T T	C T	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC
C C	C C	C C	C C	T T	C C	C C	NC NC	NC NC	NC NC	T T	NC NC	NC NC
NC NC	G G	G G	G A	G G	G G	NC NC	NC NC	NC NC	NC NC	NC NC	G G	NC NC
A A	A G	A A	A G	A A	G A	A A	A A	NC NC	A A	A A	A A	A G
T C	T C	T C	T C	T T	C T	NC NC	C C	NC NC	T C	NC NC	NC NC	NC NC
A A	A G	A A	A A	G A	A G	G G	NC NC	G G	A G	A A	NC NC	G G
T T	T C	T T	T T	T T	T C	T T	T C	C C	T C	T T	T C	C C
A T	A T	A T	A T	T T	T A	NC NC	T A	NC NC	NC NC	NC NC	T T	NC NC
G T	G G	G T	G G	G G	G G	G G	G G	NC NC	NC NC	NC NC	NC NC	NC NC
A G	A A	A G	A G	A A	A A	A A	A A	A A	A A	A A	A A	A A
A G	A A	A G	A A	A G	A A	A A	A A	G G	G A	G G	G G	A A
C C	C C	C C	G C	G C	C C	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC
G A	G G	G A	G G	G A	G G	NC NC	NC NC	NC NC	A A	NC NC	NC NC	NC NC
T T	T T	T T	C T	C T	T T	NC NC	T T	NC NC	T T	T T	T T	C T
G G	G G	G G	A G	A G	G G	NC NC	G G	NC NC	G G	G G	G G	G G
T C	NC NC	T C	T C	T T	T T	NC NC	NC NC	NC NC	T T	NC NC	NC NC	T T
T C	T C	T C	C C	T T	T C	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC
A G	A G	A G	G G	A A	A G	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC
A G	A G	A G	A A	G G	A G	G G	G G	NC NC	G G	G G	G G	T T
G T	G T	G T	G T	T T	G T	NC NC	T T	NC NC	T T	T T	T T	T T
T C	T C	T C	T C	C T	C T	C C	T T	NC NC	T T	T T	T T	C T
G A	G G	G A	G G	G A	G G	G G	A G	G G	A G	G G	A G	G G
C T	C T	C T	C C	T T	C T	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC
A G	A A	A G	A A	A G	A A	A A	A A	G G	G A	G G	G A	G A
T G	T G	T G	T T	T G	T G	NC NC	G G	NC NC	G G	G G	G G	T T
C G	C G	NC NC	NC NC	C G	C G	NC NC	NC NC	NC NC	C G	C G	C G	NC NC
NC NC	A C	NC NC	A C	C A	C A	A C	A A	NC NC	A A	A A	NC NC	NC NC
C C	C C	C C	C G	G C	C C	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC	C C
G G	G G	G G	G G	G A	G G	NC NC	NC NC	NC NC	G G	NC NC	NC NC	NC NC
C A	C A	C A	C A	C A	C A	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC	A A
G C	G C	G C	G C	C G	C G	NC NC	G G	G G	G G	NC NC	NC NC	NC NC
C G	C G	C G	C G	G C	G C	NC NC	C C	C C	C C	C C	C C	G C
A G	A G	A G	A A	A G	A G	A A	G G	NC NC	G G	NC NC	G G	A G

Continúa en la página siguiente

Abuelo	Hermano	Tia1	Tia2	Madre	Padre	e1	e2	e3	e4	e5	e6	e8
G T	G T	G T	G T	G T	G T	T T	NC NC	NC NC	G G	NC NC	NC NC	NC NC
C C	C C	C C	C T	C T	C C	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC
T C	T T	T C	T T	T T	T C	NC NC	NC NC	NC NC	T C	T C	NC NC	NC NC
T C	T C	T C	T T	T T	T C	T T	C C	C C	T T	T T	T C	T C
T G	T G	T G	T T	T G	T G	NC NC	G G	G G	NC NC	G G	NC NC	NC NC
A G	A G	A G	A A	A G	A G	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC
T C	T C	T C	T T	T C	T C	T C	C C	NC NC	C C	C C	C C	T T
NC NC	G G	G G	G G	A G	G G	NC NC	G G	G G	G G	NC NC	NC NC	NC NC
C C	C C	C C	C C	G G	C C	G C	G C	NC NC	G G	G G	G C	NC NC
C C	C C	C C	C C	T T	C C	T C	T C	C C	T T	T T	T C	T T
G G	G G	G G	G G	A A	G G	NC NC	NC NC	NC NC	A G	A A	A A	NC NC
C C	C C	C C	C C	T T	C C	NC NC	T T	C C	T C	T T	NC NC	NC NC
T T	T C	T T	T T	C C	T C	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC
G G	G G	G G	G G	A A	G G	NC NC	NC NC	G G	NC NC	NC NC	NC NC	NC NC
NC NC	G A	NC NC	NC NC	G G	G A	G A	G G	NC NC	G G	G G	G G	NC NC
C T	C T	C T	C C	T T	C T	NC NC	NC NC	NC NC	T T	NC NC	T T	T T
A G	A G	A G	A A	G G	A G	NC NC	G G	NC NC	G G	NC NC	G G	G G
G T	G T	G T	G G	T G	G T	T G	G T	NC NC	G T	G G	T T	T T
G A	G A	G A	G G	A G	G A	A G	A A	A A	NC NC	NC NC	G G	NC NC
T G	T G	T G	T T	G G	T G	G T	G G	G G	G G	G G	G G	G G
T T	T T	T T	T T	C T	T T	NC NC	T T	NC NC	NC NC	NC NC	NC NC	NC NC
T T	T T	T T	T T	C T	T T	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC
C T	C T	C T	C T	T T	T C	T T	T T	NC NC	T T	T T	T T	T T
A G	A G	A G	A G	G G	A G	NC NC	G G	NC NC	G G	G G	G G	NC NC
T A	T A	T A	T A	A A	T A	A T	A A	NC NC	A A	A A	A A	A A
G A	G G	G A	G A	A G	G G	A G	G G	G G	G G	G G	G G	A G
A G	A A	A G	A G	G A	A A	NC NC	A A	NC NC	A A	A A	A A	NC NC
C C	C C	C C	C C	T C	C C	T C	C C	C C	C C	C C	C C	T C
NC NC	G G	G G	G G	A A	G G	A G	A C	G G	A A	A A	G A	G A
T T	T T	T T	T T	C C	T T	C T	C C	NC NC	C C	C C	T C	T T
NC NC	T T	NC NC	T A	T T	T T	NC NC	NC NC	NC NC	T T	NC NC	NC NC	NC NC
NC NC	G A	G G	G G	G A	G A	G G	A A	NC NC	A A	NC NC	A A	G G
C C	C T	C C	C C	C T	C T	C C	T T	NC NC	NC NC	T T	NC NC	NC NC
NC NC	C A	C C	C C	C A	C A	NC NC	A A	NC NC	A A	A A	A A	NC NC
T C	T C	T C	T C	T C	T C	T T	C C	NC NC	C C	C C	C C	T C
T C	T C	T C	T C	T C	T C	T T	C C	NC NC	C C	C C	C C	T C
NC NC	NC NC	A T	A T	A T	A T	NC NC	T T	NC NC	T T	NC NC	NC NC	NC NC

Continúa en la página siguiente

Abuelo	Hermano	Tia1	Tia2	Madre	Padre	e1	e2	e3	e4	e5	e6	e8
C T	C T	C T	C T	C T	C T	C C	T T	NC NC	T T	T T	T T	C T
T T	T T	T T	T T	G T	T T	G G	T T	NC NC	T T	T T	T T	G T
G A	G G	G A	G A	A G	G G	A A	G G	NC NC	G G	G G	G G	G G
NC NC	A G	A G	A G	G G	A G	G G	G G	NC NC	G G	G G	G G	G G
C T	C T	C T	C T	T T	C T	T C	T T	NC NC	T T	T T	T C	T T
T C	T C	T C	T C	T T	T C	NC NC	NC NC	NC NC	NC NC	T T	NC NC	NC NC
NC NC	A A	NC NC	A A	T T	A A	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC
NC NC	NC NC	T T	NC NC	T C	T C	NC NC	C C	NC NC	C C	NC NC	NC NC	NC NC
NC NC	A G	NC NC	A A	A A	A G	NC NC	NC NC	G G	NC NC	NC NC	NC NC	NC NC
G G	G A	G G	G G	G G	G A	G G	G A	A A	G A	G G	G A	G A
C C	C G	C C	C C	C C	C G	NC NC	NC NC	C C	NC NC	NC NC	NC NC	NC NC
G G	G A	G G	G G	G G	G A	G G	G A	A A	G A	G G	G A	G A
G G	G A	G G	G G	G G	G A	NC NC	G A	A A	G G	NC NC	NC NC	G A
G G	G T	G G	G G	G G	G T	G G	G T	T T	G T	G G	G T	G T
T C	T T	T C	T C	T T	T T	T T	T T	T T	T T	T T	T T	T T
A G	A A	A G	A G	A A	A A	A A	A A	A A	A A	A A	A A	A A
T T	T G	T T	T T	T T	T G	NC NC	NC NC	NC NC	G G	NC NC	NC NC	G G
C C	C A	C C	C C	C C	C A	NC NC	C A	C C	C A	C C	NC NC	C A
C T	C T	C T	C T	C C	C T	C C	C T	T T	C T	C C	C T	C T
T T	T T	T T	T T	T A	T T	T T	A T	T T	A T	A A	A T	T T
NC NC	T T	T T	T T	G T	T G	T T	T G	NC NC	T G	NC NC	NC NC	T T
T C	T T	T C	T C	C T	T T	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC
NC NC	T C	NC NC	T T	T T	T C	NC NC	NC NC	C C	T C	NC NC	NC NC	NC NC
G A	G A	G A	G A	G A	A G	A A	A A	G G	G G	A A	A A	G G
G A	G A	G A	G A	A G	A G	NC NC	G A	G G	NC NC	A A	NC NC	G G
A A	A T	A A	A A	A A	A T	NC NC	A A	NC NC	A T	A A	A T	A T
T G	T G	T G	T G	T T	T G	NC NC	T G	NC NC	T G	T T	T T	NC NC
NC NC	G A	NC NC	NC NC	G A	G A	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC	NC NC
G A	G G	G A	G A	G A	G G	NC NC	A G	NC NC	A G	NC NC	A G	NC NC
C T	C C	C T	C T	C C	C C	NC NC	C C	NC NC	NC NC	C C	NC NC	NC NC
T A	T T	T A	T A	T T	T T	NC NC	T T	NC NC	T T	NC NC	T T	NC NC
G A	G A	G A	G A	A A	A G	A A	A G	NC NC	A G	A G	A G	A G
C G	C G	C G	C G	G C	G C	G G	NC NC	NC NC	C C	NC NC	NC NC	C C
C T	C T	C T	C T	T C	T C	T T	C C	NC NC	C C	NC NC	C C	NC NC
NC NC	A A	A A	A A	G G	A A	G A	G G	NC NC	G A	NC NC	NC NC	NC NC

Tabla 23: Alelos cosegregantes con cada cromosoma de la pareja3.

5.3 Comparativa

La Tabla 24 muestra los resultados de ambas técnicas para cada uno de los embriones. Se ha indicado con fuente roja los casos donde hubo incongruencias entre los resultados o cuando no hubo seguridad sobre el resultado obtenido.

En general, se observó que la técnica aquí desarrollada aporta mayor seguridad y certeza sobre los resultados obtenidos. Además, en un caso un embrión fue determinado de manera diferente a la técnica de análisis por STRs.

Pareja 1			Pareja 2			Pareja 3		
Embrión	STR	SNPs	Embrión	STR	SNPs	Embrión	STR	SNPs
E1	−	−	E2	?	−	E1	?	+
E2	?	+	E4	+	+	E2	?	−
E3	−	−	E6	−	−	E3	−	−
E4	−	−	E7	?	−	E4	?	−
E5	?	−	E8	+	+	E5	?	−
E7	+	−	E10	−	−	E6	?	−
E9	?	+	E12	−	−	E8	−	−
E10	?	+	E14	?	+			
E11	−	−						
E12	SANO	+						
E13	−	−						
E14	−	−						

Tabla 24: Comparativa de resutados obtenidos por la técnica de análisis con STRs frente a los resultados obtenidos con tagSNPs- El símbolo + indica portador, el símbolo -, no portador. El ? indica que no hubo información. En rojo se indican los embriones en las que las técnicas no fueron concordantes. En verde los casos en que una de las técnicas no resultó informativa. Cabe destacar que en todos los casos la técnica basada en tagSNP fue informativa.

5.4 Discusión

La existencia de un mayor número de marcadores disponibles con respecto a la técnica de análisis por STRs hace posible obtener resultados con mayor grado de evidencia y seguridad. En estas parejas, el análisis mediante STRs ha sido insuficiente, puesto que el número de polimorfismos informativos no ha sido suficiente para dar resultados con garantías, según las recomendaciones del PGDIS. Sin embargo, con el empleo de SNPs, frente a fenómenos de ADO aún se dispone de suficientes posiciones para cubrir la zona a analizar, de manera que es posible obtener resultados en un mayor número de muestras. Además, a diferencia de la técnica por STRs, donde el cálculo del alelo a partir del tamaño de fragmento es complicado y puede generar errores, la secuenciación del alelo del polimorfismo es directa. Frente a fenómenos de sobrecruzamiento a veces indetectables por STRs, como por ejemplo en la pareja 1 donde tan solo los STRs aguas arriba del gen fueron informativos, los SNPs proporcionan un mayor número de loci repartidos a lo largo de toda la región de análisis, lo que permite acotar el punto donde la recombinación tuvo lugar. Esto se ve reflejado en el embrión E12 de la pareja 1, cuyos STRs indicaban que había heredado el cromosoma sano y por tanto era susceptible de ser transferido; sin embargo, gracias a la técnica con tagSNPs pudimos comprobar que se había producido un sobrecruzamiento aguas abajo del último STR informativo obtenido pero antes del gen, convirtiendo al embrión en un embrión portador no transferible. Por otro lado, el embrión E5, determinado como portador en la técnica de STRs, resultó mostrar los alelos asociados al cromosoma no portador en la zona de la alteración, debido a un sobrecruzamiento no detectado por la primera técnica. Además, nuestros resultados siempre concordaron con los resultados obtenidos por la secuenciación directa de la alteración lo que confirma que se trata de una técnica precisa útil en análisis DGP-M.

Es importante resaltar el caso de parejas que se someten a procesos de reproducción asistida sin la posibilidad de disponer de material de familiares afectos. A diferencia de la técnica de STRs, el método aquí propuesto permite incluir la detección directa de la alteración en los embriones, de manera que no es necesario recurrir a técnicas accesorias, evitando el consiguiente incremento del coste económico y el tiempo necesario para obtener resultados, y permitiendo además la realización de transferencias en fresco.

Por último, algunas parejas acuden a ciclos de reproducción asistida debido a que uno de los parentales presenta una traslocación balanceada, es decir, sin pérdida ni ganancia de material genético. El 50% de la descendencia de estos progenitores presentará traslocaciones desbalanceadas, detectables por técnicas de DGP-A debido al cambio del

número de copia que conllevan. Sin embargo, aunque un 25% de la descendencia será normal, no es posible distinguirla del 25% que heredará la misma anomalía que el parental mediante las técnicas actualmente disponibles, presentando también una traslocación balanceada. Empleando esta técnica es posible mapear los cromosomas paternos e identificar el portador de la traslocación, de forma que será posible seleccionar los embriones no portadores de traslocación de aquellos que sí la porten, de entre los que no presenten alteraciones del número de copia en el análisis DGP-A.

La combinación del fasado de SNPs con los tagSNPs seleccionados mediante MiNtagSNP permite realizar un análisis indirecto de la segregación de la alteración evitando el riesgo de realizar una determinación errónea debido a fenómenos de ADO.

La combinación del fasado de SNPs con los tagSNPs seleccionados mediante MiNtagSNP facilita el análisis DGP-M y determinación de embriones aptos para ser transferidos ya que los alelos son determinados por el secuenciador, por lo que no depende de la apreciación del individuo.

La combinación del fasado de SNPs con los tagSNPs seleccionados mediante MiNtagSNP permite disponer de un número de marcadores mucho más elevado que los disponibles con la aplicación de otras técnicas del estado del arte, lo que permite identificar los cromosomas portados por cada embrión a pesar del riesgo de ADO.

En casos donde no se disponga de un familiar, la combinación del fasado de SNPs con los tagSNPs seleccionados mediante MiNtagSNP permite realizar el fasado a partir del estudio directo de la alteración sin la necesidad de recurrir a pruebas accesorias.

V. Discusión general

- **Filtrado de secuencias**

A la hora de aplicar las técnicas del estado del arte al análisis DGP-A encontramos una serie de debilidades clave, como la distorsión generada por la dispersión de las lecturas, lo cual genera la emisión de resultados erróneos en la determinación de la ploidía de los embriones de los que proceden las células biopsiadas. Además, las técnicas de filtrado no están diseñadas para filtrar los archivos BAM procedentes de la secuenciación de librerías DGP-A, ya que las secuencias no se disponen en escalera y la cobertura es del 0,01X aproximadamente. En cambio, dichos algoritmos han sido diseñados para filtrado de secuenciaciones en escalera con coberturas entre el 30 y el 1000X. Finalmente, una debilidad clave en las técnicas del estado del arte consiste en la incapacidad de distinguir duplicados de PCR no idénticos, es decir, aquellos surgidos por la amplificación de fragmentos previamente amplificados y cuya secuenciación finaliza antes de haber reproducido todo el fragmento.

Todos estos aspectos fueron tenidos en cuenta en este trabajo para desarrollar una nueva herramienta de filtrado de artefactos y duplicados de PCR denominada MiNFilterDups, que permite disminuir la dispersión de las lecturas. Además, MiNFilterDups ha sido específicamente diseñado para procesar muestras secuenciadas a muy baja cobertura, eliminando las secuencias con base en la posición genómica, gracias a que la secuenciación a baja cobertura genera regiones independientes de lecturas que se distribuyen por todo el genoma, cubriéndolo homogéneamente sin solaparse entre sí. Finalmente, a fin de evitar la influencia de posibles errores introducidos durante la creación de la librería, MiNFilterDups emplea las posiciones genómicas no sólo para eliminar aquellas secuencias con posiciones coincidentes sino aquellas secuencias reconocidas como procedentes de duplicación de la misma, incluso aunque presenten diferentes tamaños.

- **Detección de la ploidía y el nivel de mosaicismo**

Si bien es cierto que la llegada del NGS para DGP-A[195,196] convirtió a la bioinformática en una herramienta útil en pleno desarrollo exponencial, los algoritmos implementados

hasta hoy no permiten determinar con exactitud el nivel de mosacisimo[197]. Por ello, hemos desarrollado un algoritmo llamado MiNmos, que permite identificar tanto aneuploidías completas como porcentajes de aneuploidía en casos de mosaicismo, mejorando los niveles de confianza y reduciendo el riesgo de emitir un resultado erróneo incluso con mosaicismos de bajo porcentaje. Este nuevo método de calcular el porcentaje de mosaicismo con precisión permite también ser más consistente con el diagnóstico final emitido[84,142,314], a la par que permite realizar el análisis en un tiempo aceptable. Además, la técnica permite el estudio de los 23 pares de cromosomas en un proceso económico que admite secuenciación paralela de varias muestras a la vez mediante el uso de secuencias del tipo código de barras. Por otro lado, este nuevo algoritmo podría, en un futuro, permitir el estudio y determinación de la significancia de los embriones mosaicos en los ciclos de IVF, así como debatir la conveniencia de la transferencia de los distintos porcentajes de mosaicismo. Por último, gracias a nuestro algoritmo podemos estar seguros de la ploidía real del embrión, diseñando una selección con base en el embrión más adecuado para ser transferido, evitando la transferencia de varios embriones y, con ello, de embarazo múltiple.

- **Estudio de la dispersión de las lecturas**

El MAPD es un valor inversamente relacionado con el número de lecturas. Cuando una muestra presenta un elevado número de lecturas la dispersión se densifica debido a que las lecturas ocupan todo el espacio y se distribuyen homogéneamente dentro de las ventanas consideradas para el análisis de DGP-A disminuyendo el valor de MAPD. Por el contrario, un archivo BAM con bajo número de lecturas será observado como un conjunto de lecturas discretas y dispersas y el valor de MAPD de la muestra ascenderá porque algunas ventanas no contendrán tantas lecturas como otras. Esto sugiere que no es posible realizar la comparación de la dispersión de las lecturas a través de los valores de MAPD entre muestras con distinto número de lecturas.

Tras realizar el filtrado de los artefactos de PCR de la muestra resulta lógico pensar que nos encontraríamos frente a la hipótesis de que todas las muestras son comparables, en cuanto a dispersión de lecturas se refiere, con independencia del número de lecturas que presente. De esta forma, podría ser que la ecuación del MAPD estuviese introduciendo un falso efecto de categorización del valor de dispersión en función de las lecturas. En cambio, el estadístico *Z-ScoreAbs_NegLog* aquí propuesto podríamos descartar muestras cuya

dispersión se saliese de los rangos considerados como "normales", no por su número de lecturas, sino por presentar un perfil efectivamente anómalo.

- **Selección de tagSNPs útiles en el análisis DGP-M**

Las distintas técnicas de selección de tagSNPs que podemos encontrar en el estado del arte calculan conjuntos mínimos de tagSNPs que no pueden ser empleados en los análisis DGP-M, ya que no permiten determinar con exactitud los alelos segregantes en cada cromosoma.

La presente tesis desarrolla un nuevo método de selección denominado MiNtagSNP que pretende seleccionar un conjunto mínimo de tagSNPs útiles en DGP-M al solventar el problema de la informatividad. Así, MiNtagSNP realiza la selección con base en polimorfismos potencialmente informativos, es decir, aquellos loci donde uno de los parentales es homocigoto y el otro heterocigoto. El uso de SNPs informativos permite establecer los haplotipos ancestrales heredados por cada embrión y descartar para la transferencia aquellos que muestren los polimorfismos cosegregantes con el cromosoma paterno asociado a la alteración.

- **Fasado de polimorfismos y determinación de embriones aptos para transferencia**

Al no estar contemplado el empleo de los tagSNPs en los análisis DGP-M, el fasado de los mismos hasta la fecha consistía en el estudio de los alelos presentados por los STRs estudiados. Sin embargo, la aplicación de paneles de tagSNPs provoca que el número de marcadores se incremente notablemente, complicando el análisis de informatividad a realizar, ya que son muchas las consideraciones a tener en cuenta para detectar los alelos que segregan con cada cromosoma y detectando posibles fenómenos de recombinación.

La utilización de un método de fasado de SNPs permite realizar el fasado de dichos polimorfismos (predicción de dicha segregación) empleando como marcadores los paneles de tagSNPs, para determinar los embriones libres de portar la alteración.

- **Integración de los algoritmos propuestos**

La presente tesis desarrolla los algoritmos necesarios para el análisis de los datos obtenidos por técnicas de NGS para la realización de DGP-A y DGP-M con una única biopsia, lo que supone un ahorro respecto a las técnicas del estado del arte en términos económicos y de tiempo necesario hasta la obtención de resultados. Para aunar ambas técnicas en primer lugar se realiza una librería para DGP-A; una ventaja respecto a las técnicas de cariomapeado como el *Karyomapping* consiste en la posibilidad de adaptar la técnica al número de muestras disponible, mientras que los métodos basados en arrays presentan números prefijados, de manera que la maximización de resultados tan solo se obtiene con múltiplos de dicho número. Una alícuota del ADN amplificado resultante es enriquecido con los tagSNPs diseñados por MiNtagSNP para poder obtener la librería para DGP-M. Ambas librerías se combinan y se secuencian. El empleo de distintos protocolos para la elaboración de las librerías y el uso de códigos de barras diferentes permite al secuenciador arrojar los resultados para DGP-A y DGP-M por separado en cada muestra a partir de una única carrera de secuenciación. Así, cada análisis se realiza siguiendo el método correspondiente de los aquí propuestos, permitiendo la obtención de resultados en menor tiempo y costo. Los archivos procedentes de las librerías DGP-A serán filtrados empleando MiNFilterDups y posteriormente evaluados por MiNmos para conocer la ploidía de los embriones de los que proceden las biopsias. Por su parte, los ficheros correspondientes a las librerías DGP-M son fasados. Este proceso es mucho más rápido que el empleo de otras tecnologías como el karyomapping, pues tan solo añade las 4 horas necesarias para la preparación de la librería de DGP-M al tiempo necesario para la preparación de la librería de DGP-A, pudiendo realizarse todo el proceso en 12horas. Esto resulta ideal para procesos de transferencia en fresco, pero también de transferencias en diferido, ya que el protocolo puede pararse en diversos puntos.

- **Trabajo futuro**

Actualmente MiNFilterDups permite la eliminación de duplicados y artefactos de PCR para muestras de DGP-A. Uno de los principales hitos futuros consiste en la investigación de su aplicación a otras plataforma-as de secuenciación. Lo mismo ocurre con el resto de algoritmos desarrollados en el marco de esta tesis, pues aunque han sido validados y desarrollados a partir de las plataformas disponibles en nuestro laboratorio, su aplicación a otras plataformas resultaría muy beneficiosa.

MiNmos ha sido diseñado para determinar la ploidía en función de tamaños de ventana prefijados. Resultaría interesante por tanto conocer el alcance de sensibilidad del algoritmo al disminuir el tamaño de dichas ventanas, determinando así el tamaño mínimo de aneuploidía detectable en mosaicismo. Otro hito ya comentado consistiría en el establecimiento, si existe, de un valor umbral del nivel de mosaicismo a partir del cual sea recomendable o no realizar la transferencia en parejas que no dispongan de embriones euploides, así como las consecuencias de dicha transferencia. Esto actualmente se realiza, pero sin control del valor exacto de mosaicismo presente en el embrión transferido. Esto también permitiría dar respuesta a la controversia generada por la transferencia de embriones mosaico y estudiar el fenómeno del rescate embriónico en mayor profundidad.

Por su parte, MiNtagSNP podría ser ampliado para permitir la selección de polimorfismos teniendo en cuenta un mayor número de parámetros, como la probabilidad de secuenciación. Como hemos visto, gran parte de los tagSNP seleccionados no eran finalmente secuenciados con éxito en las muestras, lo que reduce el número final de polimorfismos disponibles en el análisis. La mejora, tanto de las técnicas de laboratorio, como de diseño bioinformático de esta selección, mediante la consideración de su capacidad de secuenciación permitiría disminuir aún más el número de polimorfismos necesarios para realizar el análisis DGP-M y, con ello, reducir el coste de la técnica permitiendo que más parejas accedan a ella. Por otro lado, debemos tener en cuenta las limitaciones "naturales" del propio genoma. Regiones teloméricas y regiones muy cercanas al centrómero suelen experimentar tasas de recombinación muy bajas en comparación con el resto del genoma, lo que provoca la existencia de una menor diversidad alélica y por tanto, mayor frecuencia de homocigotos. En estas zonas el algoritmo permite la selección de aquellos polimorfismos con mayor probabilidad de resultar informativos, pero su aplicación en el estudio de segregación de una pareja concreta de dicha población puede finalmente aportar un menor número de informativos de los deseables para asegurar la significancia de los resultados del estudio de informatividad. Esto podría ocurrir también en una zona intercromosómica con una alta densidad de genes esenciales o con muy baja tasa de recombinación. La estrategia más recomendable en estos casos pasa por ampliar la región bloque o modificar el valor de los parámetros de selección. Por último, parejas con alto grado de consanguinidad podrían también dificultar la aplicación. MiNtagSNP saca los tagSNPs más probables de ser informativos en una población, pero eso no implica que lo vayan a ser en el 100% de las parejas de esa población. Si en esa zona ambos padres son homocigotos, los tagSNPs no serán informativos.

Finalmente, el fasado de SNPs está actualmente preparado para considerar el fasado de familias relacionadas, pero podría complementarse con la información de bases de datos externas sobre variantes permitiendo emplear datos sobre las frecuencias alélicas de los polimorfismos y/o de las fases haplotípicas presentadas por otras familias para completar el análisis DGP-M. Esto podría ser beneficioso en aquellas parejas donde no se dispone de familiares y la detección directa de la alteración sea complicada. Además, actualmente solo se automatiza el fasado de familiares en primer y segundo grado. Una mejora consiste en ampliar el rango de parentescos posibles, introduciendo la posibilidad de casos donde existan saltos generacionales, es decir, donde no exista la posibilidad de obtención de muestras de los individuos pertenecientes a una generación.

VI. Producción científica

- **Repercusión del proyecto desarrollado**

La presente tesis ha sido desarrollada bajo el marco de la *Ayuda para la formación de doctores en empresas, "Doctorados industriales"*, englobada en el *Programa Estatal de Promoción del Talento y su Empleabilidad en I+D+I* otorgada por el, entonces *Ministerio de Economía, Industria y Competitividad*, actual *Ministerio de Ciencia, Innovación y Universidades* y cofinanciada por el *Banco Europeo de Inversiones*, a la empresa Bioarray SL con CIF:B54363049 para la contratación de Natalia Castejón Fernández como doctorando bajo el proyecto específico DI-14-06922 titulado "Herramientas Bioinformáticas para el Diagnóstico Genético Preimplantacional mediante técnicas de Secuenciación Masiva".

Probablemente, dada la connotación económica que un doctorado industrial conlleva subyacente al desarrollar su trabajo íntegramente en una empresa con independencia de la tutela de la Universidad, el rendimiento más reseñable de la presente tesis consiste en la **presentación de la patente** titulada "*MÉTODO PARA EL ESTUDIO DE MUTACIONES EN EMBRIONES EN PROCESOS DE REPRODUCCIÓN IN VITRO*", solicitada en la fecha 20/07/2018con número de solicitud 201830731, y publicada el 20/01/2020 con número de publicación ES2738176 y clasificación internacional de patentes G16B 20/20, C12Q 1/6827. https://www.patentes-y-marcas.com/patente/metodo-para-el-estudio-de-mutaciones-en-embriones-en-procesos-de-reproduccion-in-vitro-p201830731

Publicaciones directamente relacionadas con la tesis

Durante los tres años que ha durado el presente proyecto, que dio comienzo el 12 de enero de 2016, las ideas y herramientas desarrolladas han repercutido en forma de:

Artículos en revistas

- Castejón-Fernandez, N, Amoros, D., Gonzalez-Reig, S., Blanca, H., Penacho, V., Galán, F., Alcaraz, L.A., 2018. A novel algorithm for determining the level of mosaicism in preimplantation genetic screening (PGS) with next-generation sequencing (NGS). Reproductive BioMedicine Online 36, e16.. 2017 16th Conference on PGDIS. https://doi.org/10.1016/j.rbmo.2017.10.038

Comunicaciones en congresos

- Castejón- Fernández N. "Herramientas bioinformáticas para el diagnóstico genético preimplantacional mediante secuenciación masiva". 2018 Conferencia antiguos alumnos.

Master Interuniversitario. Universidad de Murcia y Universidad Politécnica de Cartagena. http://edit.um.es/campusdigital/conferencias-de-antiguos-alumnos-del-master-en-bioinformatica-de-la-facultad-de-biologia-de-la-umu/

- Castejon-Fernandez N, Amoros D, Gonzalez-Reig S, Blanca H, Penacho V, Lopez-Huedo A, Galan F, Fernandez MA, Fernández Breis JT Alcaraz LA. "A novel Algorithm for determining the level of mosaicism in PGS with NGS". 2017 III Jornadas Doctorales de la Universidad de Murcia.

- Castejon-Fernandez N, Amoros D, Gonzalez-Reig S, Blanca H, Penacho V, LopezHuedo A, Galan F, Fernandez MA, Alcaraz LA. IX Congreso ASEBIR. IX Congreso ASEBIR 2017. "Algoritmo de Maximización de la Informatividad De TagSNP y su aplicación en el Diagnóstico Genético Preimplantacional" (Póster)

Otras publicaciones

Artículos en revista

- Blanca, H., González-Reig, S., Penacho, V., Castejón-Fernández, N., Amoros, D., Galán, F., Alcaraz, L.A., 2018. "Detection limit of partial insertions and deletions for PGS in terms of NGS by analyzing 242 embryos of couples with balanced translocations". Reproductive BioMedicine Online 36, e17.. 2017 32th Annual meeting of ESHRE. https://doi.org/10.1016/j.rbmo.2017.10.041

- Penacho, V., Amoros, D., González-Reig, S., Galán, F., Blanca, H., Castejón, N., Alcaraz, L.A., 2018. From prenatal diagnosis of fetal abnormality to preimplantation genetic diagnosis for skeletal dysplasia using next-generation-sequencing technologies. Reproductive BioMedicine Online 36, e9–e10.. 2017 16th Conference on PGDIS. https://doi.org/10.1016/j.rbmo.2017.10.023

- González-Reig, S., Penacho, V., Amorós, D., Castejón-Fernández, N., Blanca, H., Galán, F., Alcaraz, L.A., 2018. New all-in-one protocol for 24-chromosome aneuploidies and monogenic diseases detection by next- generation sequencing: first-year experience. Reproductive BioMedicine Online 36, e34–e35. https://doi.org/10.1016/j.rbmo.2017.10.083

- Payá, Gloria, Vanesa Bautista, Mónica Camacho, Natalia Castejón-Fernández, Luís A. Alcaraz, María José Bonete, and Julia Esclapez. 2018. "Small RNAs of Haloferax Mediterranei: Identification and Potential Involvement in Nitrogen Metabolism." Genes (2). https://doi.org/10.3390/genes9020083

- Peter Frandsen, Claudia Fontsere-Alemany, Svend Vendelbo Nielsen, Natalia Castejon-Fernandez, David Hughes, Jessica Hernandez Rodriguez, Frands Carlsen, Hans Redlef Siegismund, Thomas Mailund, Tomas Marques-Bonet, Christina Hvilsom. "Targeted genetic

conservation of the endangered chimpanzee". Heredity 125, 15–27 (2020). https://doi.org/10.1038/s41437-020-0313-0

Comunicaciones en congresos

- Luís A Alcaraz, Vanessa Penacho, Santiago Gonzalez-Reig, Natalia CastejónFernández, Francisco Galán, Diego Amorós, Helena Blanca, Leonardo Díaz, Miguel Fernández. 2017 ASEBIR course. "Diagnostico Genético Preimplantacional Mediante Secuenciación Masiva combinando detección de Aneuploidías y trastornos monogénicos".

- Gonzalez-Reig S, Penacho V, Amoros D, Blanca H, Castejon-Fernandez N, Lopez-Huedo A, Galan F, Fernandez MA, Alcaraz LA. IX Congreso ASEBIR 2017. "Analisis Retrospectivo de los resultados de estudios preimplantacionales dobre parejas portadoras de translocaciones equilibradas".

- Blanca H, Gonzalez-Reig S, Penacho V, Castejon-Fernandez N, Amoros D, Galan F, Alcaraz LA. 2017 16th Conference on PGDIS. "Detection limit of partial insertions and deletions for PGS in terms of NGS by analyzing 242 embryos of couples with balanced translocations".

- 2016 31th Annual meeting of ESHRE. "Fastest benchtop NGS workflow for Preimplantation Genetic Screening with Ion reproSeq". Alcaraz LA, Penacho V, Gonzalez-Reig S, Amoros D, Castejon N, Ramos B, Enciso M, Fernandez M, Aizpurua J.

- 2016 31th Annual meeting of ESHRE. "A powerfull tagging SNP method in terms of NGS that combines PGD, informativity testing and Aneuploidy screening". Penacho V, Alcaraz LA, Gonzalez- Reig S, Amoros D, Castejon N.

- 2016 15th Conference on PGDIS. "Powerful NGS workflow that combines Preimplantation Genetic Diagnosis, Informativity Testing and Chromosoma Abnormality Screening" Penacho V, Alcaraz LA, Gonzalez-Reig S, Amoros D, Castejon N.

- 2017 FEBS3+, XL SEBBM Congress. "sRNA in haloarchaea: torwards a new model of nitrogen assimilation pathway regulation". G Payá, V Bautista, M Camacho, N Castejón-Fernández, LA Alcaraz, MJ Bonete, J Esclapez (Póster)

VII. Conclusiones

232| CONCLUSIONES

- **Evaluación de las hipótesis de trabajo**

Filtrado de duplicados de PCR:

- El filtrado de duplicados realizado por las técnicas del estado del arte no era suficientemente eficiente. MiNFilterDups es un algoritmo específicamente diseñado para muestras de DGP-A por técnicas de NGS.

- El filtrado con MiNFilterDups permite disminuir el valor de MAPD y aumentar la confianza de los resultados de las muestras filtradas con respecto al estado del arte.

- El filtrado de duplicados realizado por MiNFilterDups permite que muestras con menor número de lecturas muestren valores de MAPD inferiores a 0,3, mientras que el filtrado de dichas muestras con los algoritmos disponibles en el estado del arte muestra valores por encima de dicho umbral.

- El filtrado con MiNFilterDups permite que el diagnóstico de muestras con valores de MAPD superiores a 0,3 aún sea fiable.

- El filtrado con MiNFilterDups reconoce duplicados idénticos y secuencias formadas a partir de fragmentos previamente amplificados.

Determinación de la ploidía y el nivel de mosaicismo:

- MiNmos es un algoritmo específicamente diseñado para la detección de bajos porcentajes de aneuploidía y determinación del nivel de mosaicismo de muestras de DGP-A por técnicas de NGS.

- MiNmos es el primer algoritmo para DGP-A capaz de gradar y determinar el nivel de mosaicismo de la muestra, indicando el porcentaje de aneuploidía presente.

- El *valor Z-Score del log10 de los niveles de cobertura corregidos respecto a los valores de las dos líneas base* analizado por MiNmos es un buen indicador del estado de ploidía y el nivel de mosaicismo.

- MiNmos permite la detección del estado de ploidía y el nivel de mosaicismo incluso con valores de MAPD por encima de 0,3.

- MiNmos presenta una mayor sensibilidad y especificadad en la distinción entre embriones euploides y mosaicos de bajo nivel respecto a las técnicas actuales del estado del arte.

- MiNmos necesita un menor número de lecturas mínimas para obtener resultados fiables respecto a los algoritmos del estado del arte. Esto supone un ahorro en tiempo y dinero, pues a menor número de lecturas necesarias para un correcto diagnóstico, mayor cantidad de muestras pueden ser secuenciadas a la vez.

Estudio de la dispersión de las lecturas:

- El valor de MAPD de la muestra está muy relacionado con el número de lecturas por lo que no permite la comparación entre muestras con distinto número.

- La ploidía de un embrión puede ser correctamente determinada con independencia de su valor de MAPD si ha sido correctamente filtrada.

- El logaritmo negativo en base 10 del valor absoluto de Z-Score de la cobertura de las muestras es una medida absoluta que permite la comparativa de la dispersión de las lecturas entre las muestras con independencia del número de lecturas que presenten.

Selección de tagSNPs útiles en análisis DGP-M:

- MiNtagSNP es un algoritmo específicamente diseñado para seleccionar un conjunto mínimo de tagSNPs útiles para los análisis de DGP-M por técnicas de NGS.

- La aplicación de los tagSNPs seleccionados con MiNtagSNP permite incluir la detección directa de la alteración sin necesidad de realizar pruebas accesorias.

- MiNtagSNP necesita seleccionar un menor número de tagSNPs que las técnicas del estado del arte y, además, estos son más informativos.

- A pesar de calcular tagSNPs en regiones pequeñas, MiNtagSNP es capaz de seleccionar tagSNPs que son, en promedio, más independientes que los polimorfismos seleccionados por el estado del arte.

Fasado de SNPs y determinación de embriones aptos para transferencia:

- La metodología aquí mostrada permite distinguir los embriones portadores no por presentar la alteración, sino por mostrar los polimorfismos cosegregantes con el cromosoma paterno identificado como portador.

- La combinación del fasado de SNPs con los tagSNPs seleccionados mediante MiNtagSNP permite realizar un análisis indirecto de la segregación de la alteración evitando el riesgo de realizar una determinación errónea debido a fenómenos de ADO.

- El número de marcadores mucho más elevado que los disponibles con la aplicación de otras técnicas del estado del arte, lo que permite identificar los cromosomas portados por cada embrión a pesar del riesgo de ADO.

- En casos donde no se disponga de un familiar, es posible realizar el fasado a partir del estudio directo de la alteración sin la necesidad de recurrir a pruebas accesorias.

• Conclusiones generales:

- La presente tesis ha desarrollado los algoritmos necesarios para implantar un método de análisis rápido y eficaz capaz de combinar DGP-A y DGP-M a partir de una sola biopsia empleando técnicas de secuenciación masiva.

- La aplicación de los algoritmos desarrollados mejora la confianza y fiabilidad de los resultados obtenidos con respecto a la aplicación de las técnicas del estado del arte.

236| CONCLUSIONES

- De forma genérica podemos afirmar que todos los algoritmos presentaron tiempos de ejecución asumibles que, en combinación, permiten obtener resultados más fiables en menor tiempo que otras técnicas del estado del arte.

- **Hypothesis evaluation**

Duplicates and PCR artifacts filter:

- State of the art techniques are not efficient enough for filtering duplicates and PCR artifacts. MiNFilterDups was specifically designed for PGT-A samples sequenced by NGS techniques.
- MiNFilterDups decreases the MAPD value while increasing the confidence value of the filtered simples in respect to the use of the state of the art.
- Filtering duplicates and PCR artifacts by MiNFilterDups shows values of MAPD lower than 0.3 when used with samples of low number of reads while the use of the state-of-the-art shows MAPD values above this threshold.
- Even when sample shows MAPD values greater than 0.3, the use of MiNFilterDups makes the diagnosis still reliable.
- MiNFilterDups algorithm allows the identification of identical duplicates and sequences formed that comes from previously amplified fragments.

Detection of low percentages of aneuploidy and mosaicism level determination:

- MiNmos has been specifically designed for the detection of low percentages of aneuploidy and the determination of the level of mosaicism in PGT-A samples sequenced by NGS techniques.
- MiNmos is the First algorithm capable of grading and determining the level of mosaicism in the sample, indicating the percentage of aneuploidy.
- The analysis of the *Z-Score value of the log10 coverage levels*, corrected by using two baselines is a good indicator of the ploidy status and the level of mosaicism.
- MiNmos allows the detection of the ploidy state and the level of mosaicism even when using with samples that show MAPD values above 0.3.
- MiNmos shows a greater sensitivity and specificity values in the distinction between euploid embryos and low-level mosaics with respect to the use of the currently state-of-the-art techniques.
- MiNmos requires a fewer minimum reads number to obtain reliable results compared to the state-of-the-art algorithms. This fact saves time and money, because the fewer the number of reads required for a correct diagnosis, the more samples can be sequenced at the same time.

Reads dispersion study:

- The MAPD values is closely related to the number of reads presented, because of that, the raw MAPD value does not allow the comparison between simples with different number of reads.
- The ploidy of an embryo can be correctly determined independently of the MAPD value when duplicates and PCR artifacts are correctly removed.

- Negative log10 of the absolute value of the Z-score of the coverage of the sample is an absolute measure that allows the comparison of the dispersion of the reads between several simples regardless of the number of reads.

TagSNP selection for PGT-M analysis:

- MiNtagSNP is a specifically designed algorithm to select a minimum set of useful tagSNP for PGT-M analysis of simples sequenced by NGS techniques.
- The analysis of the selected tagSNPs allows the direct detection of the mutation avoiding the need of accessory tests.
- MiNtagSNP allows the selection of a few tagSNP set that the state-of-the-art techniques. Those tagSNPs are also more informative.
- Despite of using small regions to calculate the minimum tagSNP set, MiNtagSNP can select tagSNPs that are, on average, more independent than the polymorphisms selected by the state of the art methods.

SNP phasing and suitable embryos determination:

- The methodology shown allows us to distinguish the carrier embryos not because of exhibiting the alteration, but because they show the cosegregating polymorphisms with the paternal chromosome identified as the carrier.

- The combination of the SNP phasing with the tagSNPs selected by MiNtagSNP allows an indirect analysis of the segregation of the alteration, avoiding the risk of making an erroneous determination due to ADO phenomenon.
- The number of markers is much higher than the other state-of-the-art techniques, this fact makes the identification of the chromosomes carried by each embryo possible, despite the risk of ADO.
- In cases where relatives are not available, it is possible to perform the analysis from the direct study of the alteration without the need to resort to accessory tests.

General conclusions:

- This thesis provides the necessary algorithms to implement a fast and efficient analysis method combining PGT-A and PGT-M from a single biopsy using massive sequencing techniques.
- Our algorithms improve the confidence value and reliability of the results obtained with respect to the application of the techniques of the state of the art.
- In a generic way, we can affirm that all the algorithms presented acceptable execution times that, in combination, allow for obtaining more reliable results in less time than other techniques of the state of the art.

BIBLIOGRAFÍA

1. Wells, D., Sherlock, J. K., Handyside, A. H. & Delhanty, J. D. Detailed chromosomal and molecular genetic analysis of single cells by whole genome amplification and comparative genomic hybridisation. *Nucleic acids research* **27**, 1214–8 (1999).
2. Geraedts, J. & De Wert, G. Preimplantation genetic diagnosis. *Clinical Genetics* **76**, 315–325 (2009).
3. Schoolcraft, W. B. *et al.* Clinical application of comprehensive chromosomal screening at the blastocyst stage. *Fertility and sterility* **94**, 1700–6 (2010).
4. Haddad, G. *et al.* Mosaic pregnancy after transfer of a "euploid" blastocyst screened by DNA microarray. *Journal of Ovarian Research* **6**, 70 (2013).
5. Pennisi, E. BREAKTHROUGH OF THE YEAR: Human Genetic Variation. *Science* **318**, 1842–1843 (2007).
6. Ignacio Arroyo Carrera. Classification of genetic alterations. *Pediatr Integral* **XIV(8)**, (2010).
7. Singh, R. S. Polymorphism. in *Encyclopedia of Genetics* 1507–1509 (Elsevier, 2001). doi:10.1006/rwgn.2001.1012.
8. Laurence Loewe (School of Biological Sciences, University of Edinburgh, Scotland, UK.). Genetic Mutation. *Nature Education* **1(1):113**, (2008).
9. Dwivedi, S., Singh, S., Chauhan, U. K. & Tiwari, M. K. Inter and intraspecific genetic diversity (RAPD) among three most frequent species of macrofungi (Ganoderma lucidum , Leucoagricus sp. and Lentinus sp.) of Tropical forest of Central India. *Journal of Genetic Engineering and Biotechnology* **16**, 133–141 (2018).
10. Ren, Y., Qiao, J. & Yan, L. [Advance in the methods of preimplantation genetic diagnosis for single gene diseases]. *Zhonghua yi xue yi chuan xue za zhi = Zhonghua yixue yichuanxue zazhi = Chinese journal of medical genetics* **34**, 443–447 (2017).
11. Ocak, Z., Özlü, T. & Ozyurt, O. Association of recurrent pregnancy loss with chromosomal abnormalities and hereditary thrombophilias. *African health sciences* **13**, 447–52 (2013).
12. Tech, V. How Mitochondrial Disease Is Passed Down From Mother To Child: Predicting Severity. *ScienceDaily*.
13. Grønbaek, K. & Guldberg, P. [Acquired mutations--basic cancer biology]. *Ugeskrift for laeger* **168**, 2335–8 (2006).

14. Fitzpatrick, M. A. *et al.* Identification of chromosomal alterations important in the development of cervical intraepithelial neoplasia and invasive carcinoma using alignment of DNA microarray data. *Gynecologic oncology* **103**, 458–62 (2006).

15. Nomenclature., L. G. S. M. S. J. M.-J. I. S. C. on H. C. *ISCN : an international system for human cytogenomic nomenclature (2016)*. (2016).

16. URMC. Medical Genetics: How Chromosome Abnormalities Happen. in *Health Encyclopedia* (2018).

17. Ljunger, E., Cnattingius, S., Lundin, C. & Annerén, G. Chromosomal anomalies in first-trimester miscarriages. *Acta obstetricia et gynecologica Scandinavica* **84**, 1103–7 (2005).

18. Kim, J. W. *et al.* Chromosomal abnormalities in spontaneous abortion after assisted reproductive treatment. *BMC Medical Genetics* **11**, 153 (2010).

19. Harton, G. L. & Tempest, H. G. Chromosomal disorders and male infertility. *Asian journal of andrology* **14**, 32–9 (2012).

20. Kushnir, V. A. & Frattarelli, J. L. Aneuploidy in abortuses following IVF and ICSI. *Journal of Assisted Reproduction and Genetics* **26**, 93–97 (2009).

21. SIMPSON, J. L. Causes of Fetal Wastage. *Clinical Obstetrics and Gynecology* **50**, 10–30 (2007).

22. Hassold, T. J. A cytogenetic study of repeated spontaneous abortions. *American journal of human genetics* **32**, 723–30 (1980).

23. Warburton, D. *et al.* Does the karyotype of a spontaneous abortion predict the karyotype of a subsequent abortion? Evidence from 273 women with two karyotyped spontaneous abortions. *Am J Hum Genet* **41**, 465–483 (1987).

24. Shen, J. *et al.* Chromosomal copy number analysis on chorionic villus samples from early spontaneous miscarriages by high throughput genetic technology. *Molecular Cytogenetics* **9**, 7 (2016).

25. Health., D. of; D. of C. Appendix HChromosomal Abnormalities. in *Understanding Genetics: A District of Columbia Guide for Patients and Health Professionals* (ed. Allianc, G.) Bookshelf ID: NBK132149 (2010).

26. Scholz, N. B. *et al.* Triploidy--Observations in 154 Diandric Cases. *PloS one* **10**, e0142545 (2015).

27. Huang, B., Prensky, L., Thangavelu, M., Main, D. & Wang, S. Three consecutive triploidy pregnancies in a woman: genetic predisposition? *European Journal of Human Genetics* **12**, 985–986 (2004).

28. Kolarski, M. *et al.* Genetic Counseling and Prenatal Diagnosis of Triploidy During the Second Trimester of Pregnancy. *Medical archives (Sarajevo, Bosnia and Herzegovina)* **71**, 144–147 (2017).

29. Compton, D. A. Mechanisms of aneuploidy. *Current opinion in cell biology* **23**, 109–13 (2011).

30. Griffiths AJF, Miller JH, S. D. Aneuploidy. in *An Introduction to Genetic Analysis* (ed. Freeman, N. Y. W. H.) (2000).

31. Lebedev, I. N., Ostroverkhova, N. V, Nikitina, T. V, Sukhanova, N. N. & Nazarenko, S. A. Features of chromosomal abnormalities in spontaneous abortion cell culture failures detected by interphase FISH analysis. *European Journal of Human Genetics* **12**, 513–520 (2004).

32. Hassold, T. & Hunt, P. To err (meiotically) is human: the genesis of human aneuploidy. *Nature Reviews Genetics* **2**, 280–291 (2001).

33. Curry, C. J. Autosomal Trisomies. in *Reference Module in Biomedical Sciences* (Elsevier, 2014). doi:10.1016/B978-0-12-801238-3.05516-1.

34. Brewer, C. M. Survival in trisomy 13 and trisomy 18 cases ascertained from population based registers. *Journal of Medical Genetics* **39**, 54e–554 (2002).

35. Cereda, A. & Carey, J. C. The trisomy 18 syndrome. *Orphanet Journal of Rare Diseases* **7**, 81 (2012).

36. Patterson, D. & Costa, A. C. S. Down syndrome and genetics — a case of linked histories. *Nature Reviews Genetics* **6**, 137–147 (2005).

37. Akbas, E. *et al.* Rare Types of Turner Syndrome: Clinical Presentation and Cytogenetics in Five Cases. *Laboratory Medicine* **43**, 197–204 (2012).

38. Lee, J. Y. A. & J. T. X Chromosome: X Inactivation. *Nature Education* **1(1)**, 24 (2008).

39. Sun, B. K. & Tsao, H. X-Chromosome Inactivation and Skin Disease. *Journal of Investigative Dermatology* **128**, 2753–2759 (2008).

40. Otter, M., Schrander-Stumpel, C. T. & Curfs, L. M. Triple X syndrome: a review of the literature. *European Journal of Human Genetics* **18**, 265–271 (2010).

41. JACOBS, P. A. & STRONG, J. A. A Case of Human Intersexuality Having a Possible XXY Sex-Determining Mechanism. *Nature* **183**, 302–303 (1959).

42. Klinefelter, H. F. Klinefelter's syndrome: historical background and development. *Southern medical journal* **79**, 1089–93 (1986).

43. Stochholm, K., Juul, S. & Gravholt, C. H. Diagnosis and mortality in 47,XYY persons: a registry study. *Orphanet Journal of Rare Diseases* **5**, 15 (2010).

44. D. Health. Appendix F Chromosomal Abnormalities. in *Understanding Genetics: A District of Columbia Guide for Patients and Health Professionals* (ed. Allianc, G.) Bookshelf ID: NBK132149 (2010).

45. JACOBS, P. A., MELVILLE, M., RATCLIFFE, S., KEAY, A. J. & SYME, J. A cytogenetic survey of 11,680 newborn infants. *Annals of Human Genetics* **37**, 359–376 (1974).

46. Van Dyke, D. L., Weiss, L., Roberson, J. R. & Babu, V. R. The frequency and mutation rate of balanced autosomal rearrangements in man estimated from prenatal genetic studies for advanced maternal age. *American journal of human genetics* **35**, 301–8 (1983).

47. Greaves, M. F. & Wiemels, J. Origins of chromosome translocations in childhood leukaemia. *Nature reviews. Cancer* **3**, 639–49 (2003).

48. Nielsen, J. & Wohlert, M. Chromosome abnormalities found among 34,910 newborn children: results from a 13-year incidence study in Arhus, Denmark. *Human genetics* **87**, 81–3 (1991).

49. Bandyopadhyay, R. *et al.* Parental origin and timing of de novo Robertsonian translocation formation. *American journal of human genetics* **71**, 1456–62 (2002).

50. Ye, Y., Qian, Y., Xu, C. & Jin, F. Meiotic segregation analysis of embryos from reciprocal translocation carriers in PGD cycles. *Reproductive BioMedicine Online* **24**, 83–90 (2012).

51. Suzanne Clancy, & K. M. S. DNA Deletion and Duplication and the Associated Genetic Disorders. *Nature Education* **1(1):23**, (2008).

52. Ferguson-Smith, M. A. Isochromosome. in *Encyclopedia of Genetics* 1054 (Elsevier, 2001). doi:10.1006/rwgn.2001.0717.

53. Puig, M., Casillas, S., Villatoro, S. & Cáceres, M. Human inversions and their functional consequences. *Briefings in functional genomics* **14**, 369–79 (2015).

54. Mozdarani, H., Meybodi, A. M. & Karimi, H. Impact of pericentric inversion of Chromosome 9 [inv (9) (p11q12)] on infertility. *Indian journal of human genetics* **13**, 26–9 (2007).

55. Janse, C. J. & Mons, B. Deletion, insertion and translocation of DNA sequences contribute to chromosome size polymorphism in Plasmodium berghei. *Memorias do Instituto Oswaldo Cruz* **87 Suppl 3**, 95–100 (1992).

56. Yip, M.-Y. Autosomal ring chromosomes in human genetic disorders. *Translational pediatrics* **4**, 164–74 (2015).

57. Campana, M., Serra, A., Neri, G. & Reynolds, J. F. Role of chromosome aberrations in recurrent abortion: A study of 269 balanced translocations. *American Journal of Medical Genetics* **24**, 341–356 (1986).

58. Klupa, T., Skupien, J. & Malecki, M. T. Monogenic models: what have the single gene disorders taught us? *Current diabetes reports* **12**, 659–66 (2012).

59. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, M. Online Mendelian Inheritance in Man, OMIM (TM).

60. Castro-García, S. M. L. P. B. CAPÍTULO 21: Enfermedades monogénicas. in *Biología Molecular. Fundamentos y aplicaciones en las ciencias de la salud, 2e*.

61. Simple ClinVar: an interactive web server to explore and retrieve gene and disease variants aggregated in ClinVar database | Nucleic Acids Research | Oxford Academic. https://academic.oup.com/nar/article/47/W1/W99/5494761.

62. Joshua, D. C. & Bhatia, C. R. Alteration of reproductive effort by a monogenic, recessive mutation in jute (Corchorus capsularis L.). *Theoretical and Applied Genetics* **82**, (1991).

63. Lin, M.-T. *et al.* Relation of an interleukin-10 promoter polymorphism to graft-versus-host disease and survival after hematopoietic-cell transplantation. *The New England journal of medicine* **349**, 2201–10 (2003).

64. Betticher, D. C. *et al.* Alternate splicing produces a novel cyclin D1 transcript. *Oncogene* **11**, 1005–11 (1995).

65. Services., T. N. Y.-M.-A. C. for G. and N. S. *APPENDIX EINHERITANCE PATTERNS*. *Understanding Genetics: A New York, Mid-Atlantic Guide for Patients and Health Professionals.* (2009).

66. Miko, I. Phenotype variability: penetrance and expressivity. *Nature Education* **1(1):137**,.

67. Hoppe, B. *et al.* A vertical (pseudodominant) pattern of inheritance in the autosomal recessive disease primary hyperoxaluria type 1: lack of relationship between genotype, enzymic

phenotype, and disease severity. *American journal of kidney diseases : the official journal of the National Kidney Foundation* **29**, 36–44 (1997).

68. Plastino, E. M., Guimarães, M., Matioli, S. R. & Oliveira, E. C. Codominant inheritance of polymorphic color variants of Gracilaria domingensis (Gracilariales, Rhodophyta). *Genetics and Molecular Biology* **22**, 105–108 (1999).
69. Vogler, A. J. *et al.* Effect of repeat copy number on variable-number tandem repeat mutations in Escherichia coli O157:H7. *Journal of bacteriology* **188**, 4253–63 (2006).
70. Fan, H. & Chu, J.-Y. A brief review of short tandem repeat mutation. *Genomics, proteomics & bioinformatics* **5**, 7–14 (2007).
71. Weber, J. L. & Wong, C. Mutation of human short tandem repeats. *Human molecular genetics* **2**, 1123–8 (1993).
72. Gray, I. C. Single nucleotide polymorphisms as tools in human genetics. *Human Molecular Genetics* **9**, 2403–2408 (2000).
73. Caratachea, M. A. C. Polimorfismos genéticos: Importancia y aplicaciones. *Rev. Inst. Nal. Enf. Resp. Mex* **20–3**, 213–221 (2007).
74. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–33 (2001).
75. K, T. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
76. Garrod, A. E. The incidence of alkaptonuria: a study in chemical individuality. 1902. *Molecular medicine (Cambridge, Mass.)* **2**, 274–82 (1996).
77. Garcia-Herrero, S. *et al.* Genetic Analysis of Human Preimplantation Embryos. *Current topics in developmental biology* **120**, 421–47 (2016).
78. Handyside, A. H., Kontogianni, E. H., Hardy, K. & Winston, R. M. Pregnancies from biopsied human preimplantation embryos sexed by Y-specific DNA amplification. *Nature* **344**, 768–70 (1990).
79. Vitek, W. S. *et al.* Management of the first in vitro fertilization cycle for unexplained infertility: a cost-effectiveness analysis of split in vitro fertilization-intracytoplasmic sperm injection. *Fertility and sterility* **100**, 1381–8 (2013).
80. Shi, Y. *et al.* Transfer of Fresh versus Frozen Embryos in Ovulatory Women. *New England Journal of Medicine* **378**, 126–136 (2018).

81. JOINT SOGC-CFAS. Guidelines for the number of embryos to transfer following in vitro fertilization No. 182, September 2006. *International journal of gynaecology and obstetrics: the official organ of the International Federation of Gynaecology and Obstetrics* **102**, 203–16 (2008).

82. Simpson, J. L. Preimplantation genetic diagnosis to improve pregnancy outcomes in subfertility. *Best Practice & Research Clinical Obstetrics & Gynaecology* **26**, 805–815 (2012).

83. Wilton, L., Thornhill, A., Traeger-Synodinos, J., Sermon, K. D. & Harper, J. C. The causes of misdiagnosis and adverse outcomes in PGD. *Human Reproduction* **24**, 1221–1228 (2009).

84. Capalbo, A. *et al.* Consistent and reproducible outcomes of blastocyst biopsy and aneuploidy screening across different biopsy practitioners: a multicentre study involving 2586 embryo biopsies. *Human Reproduction* **31**, 199–208 (2016).

85. Geraedts, J. *et al.* What next for preimplantation genetic screening? A polar body approach! *Human Reproduction* **25**, 575–577 (2010).

86. Montag, M., Köster, M., Strowitzki, T. & Toth, B. Polar body biopsy. *Fertility and Sterility* **100**, 603–607 (2013).

87. Zakharova, E. E., Zaletova, V. V & Krivokharchenko, A. S. Biopsy of human morula-stage embryos: outcome of 215 IVF/ICSI cycles with PGS. *PloS one* **9**, e106433 (2014).

88. De Vos, A. & Van Steirteghem, A. Aspects of biopsy procedures prior to preimplantation genetic diagnosis. *Prenatal Diagnosis* **21**, 767–780 (2001).

89. Cimadomo, D. *et al.* The Impact of Biopsy on Human Embryo Developmental Potential during Preimplantation Genetic Diagnosis. *BioMed research international* **2016**, 7193075 (2016).

90. Taylor, T. H. *et al.* The origin, mechanisms, incidence and clinical consequences of chromosomal mosaicism in humans. *Human Reproduction Update* **20**, 571–581 (2014).

91. Scott, K. L., Hong, K. H. & Scott, R. T. Selecting the optimal time to perform biopsy for preimplantation genetic testing. *Fertility and Sterility* **100**, 608–614 (2013).

92. Schoolcraft, W. B. & Katz-Jaffe, M. G. Comprehensive chromosome screening of trophectoderm with vitrification facilitates elective single-embryo transfer for infertile women with advanced maternal age. *Fertility and Sterility* **100**, 615–619 (2013).

93. Dokras, A., Sargent, I. L., Ross, C., Gardner, R. L. & Barlow, D. H. Trophectoderm biopsy in human blastocysts. *Human reproduction (Oxford, England)* **5**, 821–5 (1990).

94. MOZDARANI, H. & MORADI, S. Z. Effect of vitrification on viability and chromosome abnormalities in 8-cell mouse embryos at various storage durations. *Biological Research* **40**, (2007).

95. Gleicher, N. *et al.* A single trophectoderm biopsy at blastocyst stage is mathematically unable to determine embryo ploidy accurately enough for clinical use. *Reproductive biology and endocrinology : RB&E* **15**, 33 (2017).

96. Capalbo, A. *et al.* FISH reanalysis of inner cell mass and trophectoderm samples of previously array-CGH screened blastocysts shows high accuracy of diagnosis and no major diagnostic impact of mosaicism at the blastocyst stage. *Human Reproduction* **28**, 2298–2307 (2013).

97. TH Chan Harvard University. THE PUBLIC AND GENETIC EDITING, TESTING, AND THERAPY.

98. DARSHAK M. SANGHAVI, M. D. Wanting Babies Like Themselves, Some Parents Choose Genetic Defects.

99. Clancy, T. A clinical perspective on ethical arguments around prenatal diagnosis and preimplantation genetic diagnosis for later onset inherited cancer predispositions. *Familial Cancer* **9**, 9–14 (2010).

100. Peterson, B. *et al.* An introduction to infertility counseling: a guide for mental health and medical professionals. *Journal of assisted reproduction and genetics* **29**, 243–8 (2012).

101. Sable, D. IVF and Infertility be the numbers. *Pharma and Healthcare*.

102. Victoria Clay Wright, Jeani Chang, Gary Jeng, M. M. Assisted Reproductive Technology Surveillance-- US, 2003. www.cdc.gov/mmwr/preview/mmwrhtml/ss5504a1.html.

103. Wyns, C. *et al.* ART in Europe, 2016: results generated from European registries by ESHRE†. *Hum Reprod Open* **2020**, (2020).

104. Klitzman, R. Deciding how many embryos to transfer: ongoing challenges and dilemmas. *Reproductive biomedicine & society online* **3**, 1–15 (2016).

105. Masschaele, T., Gerris, J., Vandekerckhove, F. & De Sutter, P. Does transferring three or more embryos make sense for a well-defined population of infertility patients undergoing IVF/ICSI? *Facts, views & vision in ObGyn* **4**, 51–8 (2012).

106. Stylianou, C., Critchlow, D., Brison, D. R. & Roberts, S. A. Embryo morphology as a predictor of IVF success: An evaluation of the proposed UK ACE grading scheme for cleavage stage embryos. *Human Fertility* **15**, 11–17 (2012).

107. Dennis, S. J., Thomas, M. A., Williams, D. B. & Robins, J. C. Embryo morphology score on day 3 is predictive of implantation and live birth rates. *Journal of Assisted Reproduction and Genetics* **23**, 171–175 (2006).

108. Chen, X. *et al.* Trophectoderm morphology predicts outcomes of pregnancy in vitrified-warmed single-blastocyst transfer cycle in a Chinese population. *Journal of Assisted Reproduction and Genetics* **31**, 1475–1481 (2014).

109. Biezinová, J. *et al.* [Embryo quality evaluation according to the speed of the first cleavage after conventional IVF]. *Ceska gynekologie* **71**, 105–10 (2006).

110. Kupka, M. S. *et al.* Assisted reproductive technology in Europe, 2010: results generated from European registers by ESHRE. *Human Reproduction* **29**, 2099–2113 (2014).

111. Lee, M.-J., Lee, R. K.-K., Lin, M.-H. & Hwu, Y.-M. Cleavage speed and implantation potential of early-cleavage embryos in IVF or ICSI cycles. *Journal of Assisted Reproduction and Genetics* **29**, 745–750 (2012).

112. Zegers-Hochschild, F. *et al.* The International Glossary on Infertility and Fertility Care, 2017. *Fertility and Sterility* **108**, 393–406 (2017).

113. Borriello, F., Weinberg, D. S. & Mutter, G. L. Evaluation of gene deletions by quantitative polymerase chain reaction. Experience with the alpha-thalassemia model. *Diagnostic molecular pathology : the American journal of surgical pathology, part B* **3**, 246–54 (1994).

114. Fischer, J., Colls, P., Escudero, T. & Munné, S. Preimplantation genetic diagnosis (PGD) improves pregnancy outcome for translocation carriers with a history of recurrent losses. *Fertility and Sterility* **94**, 283–289 (2010).

115. Brugo-Olmedo, S. DEFINICIÓN Y CAUSAS DE LA INFERTILIDAD. 22.

116. Verlinsky, Y. *et al.* Preimplantation testing for chromosomal disorders improves reproductive outcome of poor-prognosis patients. *Reproductive biomedicine online* **11**, 219–25 (2005).

117. Kuliev, A., Zlatopolsky, Z., Kirillova, I., Spivakova, J. & Cieslak Janzen, J. Meiosis errors in over 20,000 oocytes studied in the practice of preimplantation aneuploidy testing. *Reproductive BioMedicine Online* **22**, 2–8 (2011).

118. Rubio, C. *et al.* Incidence of sperm chromosomal abnormalities in a risk population: relationship with sperm quality and ICSI outcome. *Human reproduction (Oxford, England)* **16**, 2084–92 (2001).

119. Rai, R. & Regan, L. Recurrent miscarriage. *Lancet (London, England)* **368**, 601–11 (2006).

120. Hyde, K. J. & Schust, D. J. Genetic considerations in recurrent pregnancy loss. *Cold Spring Harbor perspectives in medicine* **5**, a023119 (2015).

121. Robinson, W. P., McFadden, D. E. & Stephenson, M. D. The origin of abnormalities in recurrent aneuploidy/polyploidy. *American journal of human genetics* **69**, 1245–54 (2001).

122. Rubio, C. *et al.* Use of array comparative genomic hybridization (array-CGH) for embryo assessment: clinical results. *Fertility and Sterility* **99**, 1044–1048 (2013).

123. Kahraman, S. *et al.* Healthy births and ongoing pregnancies obtained by preimplantation genetic diagnosis in patients with advanced maternal age and recurrent implantation failure. *Human reproduction (Oxford, England)* **15**, 2003–7 (2000).

124. Gianaroli, L. *et al.* The role of preimplantation diagnosis for aneuploidies. *Reproductive BioMedicine Online* **4**, 31–36 (2002).

125. Milán, M. *et al.* Redefining advanced maternal age as an indication for preimplantation genetic screening. *Reproductive biomedicine online* **21**, 649–57 (2010).

126. Moayeri, M., Saeidi, H., Modarresi, M. H. & Hashemi, M. The Effect of Preimplantation Genetic Screening on Implantation Rate in Women over 35 Years of Age. *Cell journal* **18**, 13–20 (2016).

127. Kuliev, A. & Rechitsky, S. Preimplantation genetic testing: current challenges and future prospects. *Expert review of molecular diagnostics* **17**, 1071–1088 (2017).

128. Sermon, K. Novel technologies emerging for preimplantation genetic diagnosis and preimplantation genetic testing for aneuploidy. *Expert review of molecular diagnostics* **17**, 71–82 (2017).

129. Blockeel, C. *et al.* Prospectively randomized controlled trial of PGS in IVF/ICSI patients with poor implantation. *Reproductive biomedicine online* **17**, 848–54 (2008).

130. Hardarson, T. *et al.* Preimplantation genetic screening in women of advanced maternal age caused a decrease in clinical pregnancy rate: a randomized controlled trial. *Human reproduction (Oxford, England)* **23**, 2806–12 (2008).

131. Mir, P. *et al.* Improving FISH diagnosis for preimplantation genetic aneuploidy screening. *Human reproduction (Oxford, England)* **25**, 1812–7 (2010).

132. Rubio, C. *et al.* The importance of good practice in preimplantation genetic screening: critical viewpoints. *Human reproduction (Oxford, England)* **24**, 2045–7 (2009).

133. Simpson, J. L. What next for preimplantation genetic screening? Randomized clinical trial in assessing PGS: necessary but not sufficient. *Human reproduction (Oxford, England)* **23**, 2179–81 (2008).

134. Serrao, E., Cherepanov, P. & Engelman, A. N. Amplification, Next-generation Sequencing, and Genomic DNA Mapping of Retroviral Integration Sites. *Journal of visualized experiments : JoVE* (2016) doi:10.3791/53840.

135. Random Primers. http://www.thermofisher.com/order/catalog/product/48190011.

136. Song, K., Li, L. & Zhang, G. Coverage recommendation for genotyping analysis of highly heterologous species using next-generation sequencing technology. *Scientific Reports* **6**, 35736 (2016).

137. Gilfillan, G. D. *et al.* Limitations and possibilities of low cell number ChIP-seq. *BMC Genomics* **13**, 645 (2012).

138. Schweyen, H., Rozenberg, A. & Leese, F. Detection and removal of PCR duplicates in population genomic ddRAD studies by addition of a degenerate base region (DBR) in sequencing adapters. *Biol. Bull.* **227**, 146–160 (2014).

139. Kanagawa, T. Bias and artifacts in multitemplate polymerase chain reactions (PCR). *Journal of Bioscience and Bioengineering* **96**, 317–323 (2003).

140. Howson, E. L. A. *et al.* Defining the relative performance of isothermal assays that can be used for rapid and sensitive detection of foot-and-mouth disease virus. *Journal of Virological Methods* **249**, 102–110 (2017).

141. Parkinson, N. J. *et al.* Preparation of high-quality next-generation sequencing libraries from picogram quantities of target DNA. *Genome research* **22**, 125–33 (2012).

142. Zanoli, L. M. & Spoto, G. Isothermal amplification methods for the detection of nucleic acids in microfluidic devices. *Biosensors* **3**, 18–43 (2013).

143. Isothermal Amplification Ion torrent. https://tools.thermofisher.com/content/sfs/manuals/MAN0009347_IonTemplate_IA_500Kit_UG.pdf.

144. Bansal, V. A computational method for estimating the PCR duplication rate in DNA and RNA-seq experiments. *BMC bioinformatics* **18**, 43 (2017).

145. Munné, S. & Wells, D. Detection of mosaicism at blastocyst stage with the use of high-resolution next-generation sequencing. *Fertility and sterility* **107**, 1085–1091 (2017).

146. Freed, D., Stevens, E. L. & Pevsner, J. Somatic mosaicism in the human genome. *Genes* **5**, 1064–94 (2014).

147. Hassold, T. & Hunt, P. Maternal age and chromosomally abnormal pregnancies: what we know and what we wish we knew. *Current opinion in pediatrics* **21**, 703–8 (2009).

148. Fragouli, E. *et al.* Cytogenetic analysis of human blastocysts with the use of FISH, CGH and aCGH: scientific data and technical evaluation. *Human reproduction (Oxford, England)* **26**, 480–90 (2011).

149. Hegele, R. A. Copy-Number Variations and Human Disease. *The American Journal of Human Genetics* **81**, 414–415 (2007).

150. van Karnebeek, C. D. M. *et al.* Etiology of mental retardation in children referred to a tertiary care center: a prospective study. *American journal of mental retardation : AJMR* **110**, 253–67 (2005).

151. Kushnir, V. A., Darmon, S. K., Barad, D. H. & Gleicher, N. Degree of mosaicism in trophectoderm does not predict pregnancy potential: a corrected analysis of pregnancy outcomes following transfer of mosaic embryos. *Reprod Biol Endocrinol* **16**, (2018).

152. Bazrgar, M., Gourabi, H., Valojerdi, M. R., Yazdi, P. E. & Baharvand, H. Self-correction of chromosomal abnormalities in human preimplantation embryos and embryonic stem cells. *Stem cells and development* **22**, 2449–56 (2013).

153. Greco, E., Minasi, M. G. & Fiorentino, F. Healthy Babies after Intrauterine Transfer of Mosaic Aneuploid Blastocysts. *New England Journal of Medicine* **373**, 2089–2090 (2015).

154. Bolton, H. *et al.* Mouse model of chromosome mosaicism reveals lineage-specific depletion of aneuploid cells and normal developmental potential. *Nature Communications* **7**, 11165 (2016).

155. Scott, R. T. *et al.* Comprehensive chromosome screening is highly predictive of the reproductive potential of human embryos: a prospective, blinded, nonselection study. *Fertility and sterility* **97**, 870–5 (2012).

156. Preimplantation Genetic Diagnosis International Society (PGDIS). http://www.pgdis.org.

157. PGDIS Position Statement on the Transfer of Mosaic Embryos 2019 - Reproductive BioMedicine Online. https://www.rbmojournal.com/article/S1472-6483(19)30599-1/fulltext.

158. Capalbo, A., Ubaldi, F. M., Rienzi, L., Scott, R. & Treff, N. Detecting mosaicism in trophectoderm biopsies: current challenges and future possibilities. *Human reproduction (Oxford, England)* **32**, 492–498 (2017).

159. Daar, J. *et al.* Use of preimplantation genetic testing for monogenic defects (PGT-M) for adult-onset conditions: an Ethics Committee opinion. *Fertility and Sterility* **109**, 989–992 (2018).

160. Blanco, L. *et al.* Highly efficient DNA synthesis by the phage phi 29 DNA polymerase. Symmetrical mode of DNA replication. *The Journal of biological chemistry* **264**, 8935–40 (1989).

161. Findlay, I., Ray, P., Quirke, P., Rutherford, A. & Lilford, R. Allelic drop-out and preferential amplification in single cells and human blastomeres: implications for preimplantation diagnosis of sex and cystic fibrosis. *Human reproduction (Oxford, England)* **10**, 1609–18 (1995).

162. Rechitsky, S. *et al.* Allele dropout in polar bodies and blastomeres. *Journal of assisted reproduction and genetics* **15**, 253–7 (1998).

163. Blais, J. *et al.* Risk of Misdiagnosis Due to Allele Dropout and False-Positive PCR Artifacts in Molecular Diagnostics. *The Journal of Molecular Diagnostics* **17**, 505–514 (2015).

164. Hahn, S., Garvin, A. M., Di Naro, E. & Holzgreve, W. Allele drop-out can occur in alleles differing by a single nucleotide and is not alleviated by preamplification or minor template increments. *Genetic testing* **2**, 351–5 (1998).

165. De Vos, A. *et al.* Pregnancy after preimplantation genetic diagnosis for Charcot-Marie-Tooth disease type 1A. *Molecular human reproduction* **4**, 978–84 (1998).

166. Ao, A. *et al.* Clinical experience with preimplantation genetic diagnosis of cystic fibrosis (delta F508). *Prenatal diagnosis* **16**, 137–42 (1996).

167. Naehrlich, L., Bagheri-Behrouzi, A. & German CF quality assurance group. Misdiagnosis of cystic fibrosis: experience from Germany. *Journal of cystic fibrosis : official journal of the European Cystic Fibrosis Society* **12**, 68–73 (2013).

168. Slatkin, M. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nature reviews. Genetics* **9**, 477–85 (2008).

169. Cornélis, F. *et al.* New susceptibility locus for rheumatoid arthritis suggested by a genome-wide linkage study. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 10746–50 (1998).

170. Jawaheer, D. *et al.* A genomewide screen in multiplex rheumatoid arthritis families suggests genetic overlap with other autoimmune diseases. *American journal of human genetics* **68**, 927–36 (2001).

171. MacKay, K. *et al.* Whole-genome linkage analysis of rheumatoid arthritis susceptibility loci in 252 affected sibling pairs in the United Kingdom. *Arthritis and rheumatism* **46**, 632–9 (2002).

172. Goldfeld, A. E. *et al.* Association of an HLA-DQ allele with clinical tuberculosis. *JAMA* **279**, 226–8 (1998).

173. Langer-Safer, P. R., Levine, M. & Ward, D. C. Immunological method for mapping genes on Drosophila polytene chromosomes. *Proc Natl Acad Sci U S A* **79**, 4381–4385 (1982).

174. Harper, J. C. *et al.* The ESHRE PGD Consortium: 10 years of data collection. *Human Reproduction Update* **18**, 234–247 (2012).

175. Hulten, M., Dhanjal, S. & Pertl, B. *Rapid and simple prenatal diagnosis of common chromosome disorders: Advantages and disadvantages of the molecular methods FISH and QF-PCR*. vol. 126 (2003).

176. Gutiérrez-Mateo, C. *et al.* Validation of microarray comparative genomic hybridization for comprehensive chromosome analysis of embryos. *Fertility and Sterility* **95**, 953–958 (2011).

177. van Uum, C. M. J. *et al.* SNP array-based copy number and genotype analyses for preimplantation genetic diagnosis of human unbalanced translocations. *European journal of human genetics : EJHG* **20**, 938–44 (2012).

178. Fiorentino, F. *et al.* Application of next-generation sequencing technology for comprehensive aneuploidy screening of blastocysts in clinical preimplantation genetic screening cycles. *Human Reproduction* **29**, 2802–2813 (2014).

179. Vera-Rodríguez, M. *et al.* Distribution patterns of segmental aneuploidies in human blastocysts identified by next-generation sequencing. *Fertility and Sterility* **105**, 1047-1055.e2 (2016).

180. Yan, L. *et al.* Live births after simultaneous avoidance of monogenic diseases and chromosome abnormality by next-generation sequencing with linkage analyses. *Proceedings of the National Academy of Sciences* **112**, 15964–15969 (2015).

181. Ruttanajit, T. *et al.* Detection and quantitation of chromosomal mosaicism in human blastocysts using copy number variation sequencing. *Prenatal Diagnosis* **36**, 154–162 (2016).

182. Yin, X. *et al.* Massively parallel sequencing for chromosomal abnormality testing in trophectoderm cells of human blastocysts. *Biology of reproduction* **88**, 69 (2013).

183. Zong, C., Lu, S., Chapman, A. R. & Xie, X. S. Genome-Wide Detection of Single-Nucleotide and Copy-Number Variations of a Single Human Cell. *Science* **338**, 1622–1626 (2012).

184. Wang, H., Nettleton, D. & Ying, K. Copy number variation detection using next generation sequencing read counts. *BMC bioinformatics* **15**, 109 (2014).

185. Huang, J. *et al.* Validation of multiple annealing and looping-based amplification cycle sequencing for 24-chromosome aneuploidy screening of cleavage-stage embryos. *Fertility and sterility* **102**, 1685–91 (2014).

186. Faust, G. G. & Hall, I. M. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics (Oxford, England)* **30**, 2503–5 (2014).

187. Pireddu, L., Leo, S. & Zanetti, G. SEAL: a distributed short read mapping and duplicate removal tool. *Bioinformatics (Oxford, England)* **27**, 2159–60 (2011).

188. Xu, H. *et al.* FastUniq: A Fast De Novo Duplicates Removal Tool for Paired Short Reads. *PLoS ONE* **7**, e52249 (2012).

189. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25**, 2078–9 (2009).

190. Herzeel, C., Costanza, P., Decap, D., Fostier, J. & Reumers, J. elPrep: High-Performance Preparation of Sequence Alignment/Map Files for Variant Calling. *PLOS ONE* **10**, e0132868 (2015).

191. Picard. https://broadinstitute.github.io/picard/.

192. Ebbert, M. T. W. *et al.* Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. *BMC bioinformatics* **17 Suppl 7**, 239 (2016).

193. FilterDuplicates plugin IRS. http://129.130.90.13/ion-docs/GUID-F7AC38C1-D50C-46D7-989A-F36061D3C2C4.html.

194. Mosaicism.

195. Zheng, H., Jin, H., Liu, L., Liu, J. & Wang, W.-H. Application of next-generation sequencing for 24-chromosome aneuploidy screening of human preimplantation embryos. *Molecular Cytogenetics* **8**, 38 (2015).

196. Yang, Z. *et al.* Randomized comparison of next-generation sequencing and array comparative genomic hybridization for preimplantation genetic screening: a pilot study. *BMC medical genomics* **8**, 30 (2015).

197. Jia, C.-W. *et al.* Aneuploidy in Early Miscarriage and its Related Factors. *Chinese medical journal* **128**, 2772–6 (2015).

198. Affymetrix. Median of the absolute values of all pairwise differences and quality control on Affymetrix genome-wide human SNP array 6.0. Affymetrix White Paper. (2008).

199. Yim, S.-H. *et al.* Copy number variations in East-Asian population and their evolutionary and functional implications. *Human molecular genetics* **19**, 1001–8 (2010).

200. Ning, L. *et al.* Quantitative assessment of single-cell whole genome amplification methods for detecting copy number variation using hippocampal neurons. *Scientific reports* **5**, 11415 (2015).

201. MAPD. https://ionreporter.thermofisher.com/ionreporter/help/GUID-98E69C88-0170-4857-9AB0-3DD9023D895C1.html.

202. Cai, X. *et al.* Single-cell, genome-wide sequencing identifies clonal somatic copy-number variation in the human brain. *Cell reports* **8**, 1280–9 (2014).

203. Harper, J. C. *et al.* Identification of the sex of human preimplantation embryos in two hours using an improved spreading method and fluorescent in-situ hybridization (FISH) using directly labelled probes. *Human reproduction (Oxford, England)* **9**, 721–4 (1994).

204. Griffin, D. K., Handyside, A. H., Penketh, R. J., Winston, R. M. & Delhanty, J. D. Fluorescent in-situ hybridization to interphase nuclei of human preimplantation embryos with X and Y chromosome specific probes. *Human reproduction (Oxford, England)* **6**, 101–5 (1991).

205. Rechitsky, S. *et al.* First systematic experience of preimplantation genetic diagnosis for single-gene disorders, and/or preimplantation human leukocyte antigen typing, combined with 24-chromosome aneuploidy testing. *Fertility and sterility* **103**, 503–12 (2015).

206. Spits, C. *et al.* Preimplantation genetic diagnosis for cancer predisposition syndromes. *Prenatal diagnosis* **27**, 447–56 (2007).

207. De Rycke, M. *et al.* ESHRE PGD Consortium data collection XIII: cycles from January to December 2010 with pregnancy follow-up to October 2011. *Human reproduction (Oxford, England)* **30**, 1763–89 (2015).

208. Verlinsky, Y. *et al.* Preimplantation diagnosis for immunodeficiencies. *Reproductive biomedicine online* **14**, 214–23 (2007).

209. Fiorentino, F. *et al.* The minisequencing method: an alternative strategy for preimplantation genetic diagnosis of single gene disorders. *Molecular human reproduction* **9**, 399–410 (2003).

210. Pastinen, T., Kurg, A., Metspalu, A., Peltonen, L. & Syvänen, A. C. Minisequencing: a specific tool for DNA analysis and diagnostics on oligonucleotide arrays. *Genome Res.* **7**, 606–614 (1997).

211. Fiorentino, F. *et al.* Strategies and clinical outcome of 250 cycles of Preimplantation Genetic Diagnosis for single gene disorders. *Human reproduction (Oxford, England)* **21**, 670–84 (2006).

212. Little, S. Amplification-refractory mutation system (ARMS) analysis of point mutations. *Current protocols in human genetics* **Chapter 9**, Unit 9.8 (2001).

213. Altarescu, G. *et al.* Familial haplotyping and embryo analysis for Preimplantation genetic diagnosis (PGD) using DNA microarrays: a proof of principle study. *Journal of assisted reproduction and genetics* **30**, 1595–603 (2013).

214. Dreesen, J. C. *et al.* Multiplex PCR of polymorphic markers flanking the CFTR gene; a general approach for preimplantation genetic diagnosis of cystic fibrosis. *Molecular human reproduction* **6**, 391–6 (2000).

215. Lu, Y. *et al.* Preimplantation Genetic Diagnosis for a Chinese Family with Autosomal Recessive Meckel-Gruber Syndrome Type 3 (MKS3). *PLoS ONE* **8**, e73245 (2013).

216. Renwick, P., Trussler, J., Lashwood, A., Braude, P. & Ogilvie, C. M. Preimplantation genetic haplotyping: 127 diagnostic cycles demonstrating a robust, efficient alternative to direct mutation testing on single cells. *Reproductive biomedicine online* **20**, 470–6 (2010).

217. Spits, C. *et al.* Optimization and evaluation of single-cell whole-genome multiple displacement amplification. *Human mutation* **27**, 496–503 (2006).

218. Handyside, A. H. *et al.* Karyomapping: a universal method for genome wide analysis of genetic disease based on mapping crossovers between parental haplotypes. *Journal of Medical Genetics* **47**, 651–658 (2010).

219. Navin, N. & Hicks, J. Future medical applications of single-cell sequencing in cancer. *Genome medicine* **3**, 31 (2011).

220. Poli, M. *et al.* Past, Present, and Future Strategies for Enhanced Assessment of Embryo's Genome and Reproductive Competence in Women of Advanced Reproductive Age. *Front Endocrinol (Lausanne)* **10**, (2019).

221. Wells, D. & Sherlock, J. K. Strategies for preimplantation genetic diagnosis of single gene disorders by DNA amplification. *Prenatal diagnosis* **18**, 1389–401 (1998).

222. Pritchard, J. K. & Przeworski, M. Linkage disequilibrium in humans: models and data. *American journal of human genetics* **69**, 1–14 (2001).

223. Carlson, C. S. *et al.* Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *American journal of human genetics* **74**, 106–20 (2004).

224. Bafna, V., Halldorsson, B. V., Schwartz, R., Clark, A. G. & Istrail, S. Haplotypes and informative SNP selection algorithms. in *Proceedings of the seventh annual international conference on Computational molecular biology - RECOMB '03* 19–27 (ACM Press, 2003). doi:10.1145/640075.640078.

225. Quinlan, A. R. & Marth, G. T. Primer-site SNPs mask mutations. *Nature methods* **4**, 192 (2007).

226. Chen, W.-P., Hung, C.-L., Tsai, S.-J. J. & Lin, Y.-L. Novel and efficient tag SNPs selection algorithms. *Bio-medical materials and engineering* **24**, 1383–9 (2014).

227. Liu, G., Wang, Y. & Wong, L. FastTagger: An efficient algorithm for genome-wide tag SNP selection using multi-marker linkage disequilibrium. *BMC Bioinformatics* **11**, 66 (2010).

228. Zhang, K., Deng, M., Chen, T., Waterman, M. S. & Sun, F. A dynamic programming algorithm for haplotype block partitioning. *Proceedings of the National Academy of Sciences* **99**, 7335–7339 (2002).

229. Stram, D. O. *et al.* Choosing haplotype-tagging SNPS based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study. *Human heredity* **55**, 27–36 (2003).

230. Weale, M. E. *et al.* Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene SCN1A: implications for linkage-disequilibrium gene mapping. *American journal of human genetics* **73**, 551–65 (2003).

231. Rodriguez, S., Gaunt, T. R. & Day, I. N. M. Hardy-Weinberg Equilibrium Testing of Biological Ascertainment for Mendelian Randomization Studies. *American Journal of Epidemiology* **169**, 505–514 (2009).

232. McVean, G. A. T. *et al.* The fine-scale structure of recombination rate variation in the human genome. *Science (New York, N.Y.)* **304**, 581–4 (2004).

233. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature genetics* **22**, 231–8 (1999).

234. Chang, C.-J., Huang, Y.-T. & Chao, K.-M. A greedier approach for finding tag SNPs. *Bioinformatics (Oxford, England)* **22**, 685–91 (2006).

235. Xu, Z. & Taylor, J. A. SNPinfo: integrating GWAS and candidate gene information into functional SNP selection for genetic association studies. *Nucleic Acids Research* **37**, W600–W605 (2009).

236. Hinds, D. A. *et al.* Whole-genome patterns of common DNA variation in three human populations. *Science (New York, N.Y.)* **307**, 1072–9 (2005).

237. Hao, K. Genome-wide selection of tag SNPs using multiple-marker correlation. *Bioinformatics* **23**, 3178–3184 (2007).

238. Liu, L., Wu, Y., Lonardi, S. & Jiang, T. Efficient algorithms for genome-wide tagSNP selection across populations via the linkage disequilibrium criterion. *Computational systems bioinformatics. Computational Systems Bioinformatics Conference* **6**, 67–78 (2007).

239. Crawford, D. C. & Nickerson, D. A. Definition and clinical importance of haplotypes. *Annual review of medicine* **56**, 303–20 (2005).

240. Handyside, A. H. Live births following karyomapping – a "key" milestone in the development of preimplantation genetic diagnosis. *Reproductive BioMedicine Online* **31**, 307–308 (2015).

241. World Medical Association Declaration of Helsinki. *JAMA* **310**, 2191 (2013).

242. ASEBIR. *ASEBIR. Cuadernos de Embriología Clínica. Criterios ASEBIR de valoración morfológica de Oocitos, Embriones Tempranos y Blastocistos humanos*. (2015).

243. Rubino, P., Viganò, P., Luddi, A. & Piomboni, P. The ICSI procedure from past to future: a systematic review of the more controversial aspects. *Human Reproduction Update* dmv050 (2015) doi:10.1093/humupd/dmv050.

244. Ingerslev, H. J., Højgaard, A., Hindkjaer, J. & Kesmodel, U. A randomized study comparing IVF in the unstimulated cycle with IVF following clomiphene citrate. *Human reproduction (Oxford, England)* **16**, 696–702 (2001).

245. Orvieto, R. *et al.* HMG improves IVF outcome in patients with high basal FSH/LH ratio: a preliminary study. *Reproductive biomedicine online* **18**, 205–8 (2009).

246. Shahrokh Tehraninejad, E., Farshbaf Taghinejad, M., Hossein Rashidi, B. & Haghollahi, F. Controlled ovarian stimulation with r-FSH plus r-LH vs. HMG plus r-FSH in patients candidate for IVF/ICSI cycles: An RCT. *International journal of reproductive biomedicine (Yazd, Iran)* **15**, 435–440 (2017).

247. Kumar, P., Sait, S. F., Sharma, A. & Kumar, M. Ovarian hyperstimulation syndrome. *Journal of human reproductive sciences* **4**, 70–5 (2011).

248. Andersen, A. G., Als-Nielsen, B., Hornnes, P. J. & Franch Andersen, L. Time interval from human chorionic gonadotrophin (HCG) injection to follicular rupture. *Human reproduction (Oxford, England)* **10**, 3202–5 (1995).

249. Katayama, K. P. *et al.* Ultrasound-guided transvaginal needle aspiration of follicles for in vitro fertilization. *Obstetrics and gynecology* **72**, 271–4 (1988).

250. Carrell, D. T. *et al.* A randomized, prospective analysis of five sperm preparation techniques before intrauterine insemination of husband sperm. *Fertility and sterility* **69**, 122–6 (1998).

251. Referencia Genoma Humano NCBI. https://www.ncbi.nlm.nih.gov/grc/human.

252. Perl languaje. https://www.perl.org.

253. Bash is part of the GNU project. http://www.gnu.org/software/bash/manual.

254. John B. Caouette, Edward I. Altman, P. N. *Managing credit risk: the next great financial challenge*. (1998).

255. Forstmeier, W., Wagenmakers, E.-J. & Parker, T. H. Detecting and avoiding likely false-positive findings - a practical guide. *Biological Reviews* **92**, 1941–1968 (2017).

256. Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic acids research* **40**, e72 (2012).

257. Fawcett, T. An introduction to ROC analysis. *Pattern Recognition Letters* **27**, 861–874 (2006).

258. Escrig-Sos, J., Martínez-Ramos, D. & Manuel Miralles-Tena, J. Pruebas diagnósticas: nociones básicas para su correcta interpretación y uso. *Cirugía Española* **79**, 267–273 (2006).

259. Ruopp, M. D., Perkins, N. J., Whitcomb, B. W. & Schisterman, E. F. Youden Index and Optimal Cut-Point Estimated from Observations Affected by a Lower Limit of Detection. *Biometrical Journal* **50**, 419–430 (2008).

260. Shete, S., Tiwari, H. & Elston, R. C. On estimating the heterozygosity and polymorphism information content value. *Theoretical population biology* **57**, 265–71 (2000).

261. Weir, B. S. & Cockerham, C. C. Estimation of linkage disequilibrium in randomly mating populations. *Heredity* **42**, 105–111 (1979).

262. Gregory Wames, with contributions from Gregor Gorjanc, F. L. and M. M. genetics: Population Genetics. (2013).

263. R Core Team. R Foundation for Statistical Computing. vienna, A. R:A Language and Environment for Statistical Computing. Reference Index. (2014).

264. Gaunt, T. R., Rodríguez, S. & Day, I. N. Cubic exact solutions for the estimation of pairwise haplotype frequencies: implications for linkage disequilibrium analyses and a web tool 'CubeX'. *BMC Bioinformatics* **8**, 428 (2007).

265. Gibbs, R. A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

266. McEvoy, B. P. *et al.* Geographical structure and differential natural selection among North European populations. *Genome research* **19**, 804–14 (2009).

267. arraySNP GWAS Omni2.5. Illumina. https://www.illumina.com/Documents/products/datasheet_gwas_roadmap.pdf.

268. MÄGI, R., KAPLINSKI, L. & REMM, M. THE WHOLE GENOME TAGSNP SELECTION AND TRANSFERABILITY AMONG HAPMAP POPULATIONS. in *Biocomputing 2006* 535–543 (WORLD SCIENTIFIC, 2005). doi:10.1142/9789812701626_0049.

269. Xu, Z., Kaplan, N. L. & Taylor, J. A. Tag SNP selection for candidate gene association studies using HapMap and gene resequencing data. *European Journal Of Human Genetics* **15**, 1063 (2007).

270. Xing, J. *et al.* HapMap tagSNP transferability in multiple populations: general guidelines. *Genomics* **92**, 41–51 (2008).

271. Willer, C. J. *et al.* Tag SNP selection for Finnish individuals based on the CEPH Utah HapMap database. *Genetic epidemiology* **30**, 180–90 (2006).

272. NCBI retiring HapMap. https://www.ncbi.nlm.nih.gov/variation/news/NCBI_retiring_HapMap/.

273. Browning, S. R. & Browning, B. L. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. *The American Journal of Human Genetics* **81**, 1084–1097 (2007).

274. Browning, B. L. & Browning, S. R. Genotype Imputation with Millions of Reference Samples. *The American Journal of Human Genetics* **98**, 116–126 (2016).

275. Rossetti, S. *et al.* Incompletely penetrant PKD1 alleles suggest a role for gene dosage in cyst initiation in polycystic kidney disease. *Kidney International* **75**, 848–855 (2009).

276. Rossetti, S. *et al.* Mutation Analysis of the Entire PKD1 Gene: Genetic and Diagnostic Implications. *The American Journal of Human Genetics* **68**, 46–63 (2001).

277. Read number Imact on Calculation Sensitivity. https://ionreporter.thermofisher.com/ionreporter/help/GUID-5B6D3D1C-EA01-440B-AA8B-18076455B078.html.

278. Integrative Genomics Viewer. http://www.broadinstitute.org/igv.

279. Integrative Genomic Viewer Ion Reporter. https://ionreporter.thermofisher.com/ionreporter/help/GUID-25CF9DAE-364F-4411-B668-99FD158F3874.html.

280. Demidov, G., Simakova, T., Vnuchkova, J. & Bragin, A. A statistical approach to detection of copy number variations in PCR-enriched targeted sequencing data. *BMC bioinformatics* **17**, 429 (2016).

281. PGDIS Position statement on chromosome mosaicism and preimplantation aneuploidy testing at the blastocyst stage. (2016).

282. Munné, S., Grifo, J. & Wells, D. Mosaicism: "survival of the fittest" versus "no embryo left behind". *Fertility and Sterility* **105**, 1146–1149 (2016).

283. ESHRE. Abstracts of the 32nd Annual Meeting of the European Society of Human Reproduction and Embryology. *Human Reproduction* **31**, i1–i513 (2016).

284. Well DAS, Taylor S, Kubikova N, Spath K, Turner K, Hickman C, F. E. Evidence that differences between embryology laboratories can influence the rate of mitotic errors, leading to increased chromosomal mosaicism, with significant implications for IVF success rates. in 31(supp_1.1):i25–6 (Human Reproduction, 2016).

285. Zhang, S. *et al.* Number of biopsied trophectoderm cells is likely to affect the implantation potential of blastocysts with poor trophectoderm quality. *Fertility and sterility* **105**, 1222-1227.e4 (2016).

286. Capalbo, A., Ubaldi, F. M., Rienzi, L., Scott, R. & Treff, N. Detecting mosaicism in trophectoderm biopsies: current challenges and future possibilities. *Human Reproduction* (2016) doi:10.1093/humrep/dew250.

287. Scott, R. T. & Galliano, D. The challenge of embryonic mosaicism in preimplantation genetic screening. *Fertility and sterility* **105**, 1150–1152 (2016).

288. Mertzanidou, A., Spits, C., Nguyen, H. T., Van de Velde, H. & Sermon, K. Evolution of aneuploidy up to Day 4 of human preimplantation development. *Human Reproduction* **28**, 1716–1724 (2013).

289. Mertzanidou, A. *et al.* Microarray analysis reveals abnormal chromosomal complements in over 70% of 14 normally developing human embryos. *Human reproduction (Oxford, England)* **28**, 256–64 (2013).

290. Daphnis, D. D. *et al.* Detailed FISH analysis of day 5 human embryos reveals the mechanisms leading to mosaic aneuploidy. *Human reproduction (Oxford, England)* **20**, 129–37 (2005).

291. Mantikou, E., Wong, K. M., Repping, S. & Mastenbroek, S. Molecular origin of mitotic aneuploidies in preimplantation embryos. *Biochimica et biophysica acta* **1822**, 1921–30 (2012).

292. Capalbo, A. *et al.* Correlation between standard blastocyst morphology, euploidy and implantation: an observational study in two centers involving 956 screened blastocysts. *Human reproduction (Oxford, England)* **29**, 1173–81 (2014).

293. Varela, M. A. & Amos, W. Heterogeneous distribution of SNPs in the human genome: Microsatellites as predictors of nucleotide diversity and divergence. *Genomics* **95**, 151–159 (2010).

294. Zhou, Y. & Zon, L. I. The Zon Laboratory Guide to Positional Cloning in Zebrafish. in 287–309 (2011). doi:10.1016/B978-0-12-374814-0.00016-1.

295. Sboner, A., Mu, X. J., Greenbaum, D., Auerbach, R. K. & Gerstein, M. B. The real cost of sequencing: higher than you think! *Genome biology* **12**, 125 (2011).

296. Dumont, B. L. & Payseur, B. A. Evolution of the genomic rate of recombination in mammals. *Evolution; international journal of organic evolution* **62**, 276–94 (2008).

297. Liu, T.-F. *et al.* Effective algorithms for tag SNP selection. *Journal of bioinformatics and computational biology* **3**, 1089–106 (2005).

298. Mahdevar, G., Zahiri, J., Sadeghi, M., Nowzari-Dalini, A. & Ahrabian, H. Tag SNP selection via a genetic algorithm. *Journal of Biomedical Informatics* **43**, 800–804 (2010).

299. Chen, W.-P., Hung, C.-L. & Lin, Y.-L. Efficient Haplotype Block Partitioning and Tag SNP Selection Algorithms under Various Constraints. *BioMed Research International* **2013**, 1–13 (2013).

300. Halldorsson, B. V. Optimal Haplotype Block-Free Selection of Tagging SNPs for Genome-Wide Association Studies. *Genome Research* **14**, 1633–1640 (2004).

301. Qin, Z. S., Gopalakrishnan, S. & Abecasis, G. R. An efficient comprehensive search algorithm for tagSNP selection using linkage disequilibrium criteria. *Bioinformatics* **22**, 220–225 (2006).

302. Sicotte, H., Rider, D. N., Poland, G. A., Dhiman, N. & Kocher, J.-P. A. SNPPicker: High quality tag SNP selection across multiple populations. *BMC Bioinformatics* **12**, 129 (2011).

303. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nature genetics* **31**, 241–7 (2002).

304. Zhang, K. *et al.* Haplotype block partitioning and tag SNP selection using genotype data and their applications to association studies. *Genome research* **14**, 908–16 (2004).

305. Schulze, T. G. *et al.* Defining haplotype blocks and tag single-nucleotide polymorphisms in the human genome. *Human Molecular Genetics* **13**, 335–342 (2004).

306. Lewontin, R. C. On measures of gametic disequilibrium. *Genetics* **120**, 849 LP – 852 (1988).

307. Stram, D. O. Tag SNP selection for association studies. *Genetic Epidemiology* **27**, 365–374 (2004).

308. Gopalakrishnan, S. & Qin, Z. S. TagSNP selection based on pairwise LD criteria and power analysis in association studies. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* 511–22 (2006).

309. Hofer, T., Ray, N., Wegmann, D. & Excoffier, L. Large Allele Frequency Differences between Human Continental Groups are more Likely to have Occurred by Drift During range Expansions than by Selection. *Annals of Human Genetics* **73**, 95–108 (2009).

310. Li, X., Self, S. G., Galipeau, P. C., Paulson, T. G. & Reid, B. J. Direct Inference of SNP Heterozygosity Rates and Resolution of LOH Detection. *PLoS Computational Biology* **3**, e244 (2007).

311. Rooney, K. L. & Domar, A. D. The relationship between stress and infertility. *Dialogues in clinical neuroscience* **20**, 41–47 (2018).

312. Costantini-Ferrando, M. F., Joseph-Sohan, M., Grill, E., Rauch, E. & Spandorfer, S. D. Does stress affect in vitro fertilization (IVF) outcome? *Fertility and Sterility* **106**, e61 (2016).

313. Li, X., Self, S. G., Galipeau, P. C., Paulson, T. G. & Reid, B. J. Direct inference of SNP heterozygosity rates and resolution of LOH detection. *PLoS computational biology* **3**, e244 (2007).

314. Wang, L. *et al.* Detection of Chromosomal Aneuploidy in Human Preimplantation Embryos by Next-Generation Sequencing1. *Biology of Reproduction* **90**, (2014).

ANEXOS

Media-2dv		Valor umbral	Índice exactitud	Sensibilidad	Error_Se	IC_Se+	IC_Se-	Especificidad	Error_Esp	IC_Esp+	IC_Esp-
Mosaicismo con trisomía	**100 - 90%**	3,589	1	1	0	1	1	1	0	1	1
	90 - 80%	3,348	0,6	0,266	0,223	0,490	0,042	0,933	0,126	1,059	0,807
	80 - 70%	2,977	0,633	0,266	0,223	0,490	0,042	1	0	1	1
	70 - 60%	2,699	0,6	0,2	0,202	0,402	-0,002	1	0	1	1
	60 - 50%	2,288	0,766	0,533	0,252	0,785	0,280	1	0	1	1
	50 - 40%	2,006	0,607	0,133	0,172	0,305	-0,038	1	0	1	1
	40 - 30%	1,804	0,533	0,066	0,126	0,192	-0,059	1	0	1	1
	30 - 20%	1,243	0,8	0,6	0,247	0,847	0,352	1	0	1	1
	20 - 10%	0,850	0,724	0,466	0,252	0,719	0,214	1	0	1	1
	10 - 0%	0,607	0,566	0,133	0,172	0,305	-0,038	1	0	1	1
	Promedio		**0,683**	**0,366**	**0,187**	**0,553**	**0,179**	**0,993**	**0,012**	**1,005**	**0,980**
Mosaicismo con monosomía	**0- 10%**	-0,855	0,444	0,095	0,125	0,220	-0,030	0,933	0,126	1,059	0,80
	10 - 20%	-0,967	0,804	0,619	0,207	0,826	0,411	1	0	1	1
	20 - 30%	-1,417	0,785	0,571	0,211	0,783	0,359	1	0	1	1
	30 - 40%	-1,976	0,642	0,333	0,201	0,534	0,131	0,952	0,091	1,043	0,861
	40 - 50%	-2,474	0,906	0,476	0,213	0,689	0,262	0,904	0,125	1,030	0,779
	50 - 60%	-2,968	0,785	0,619	0,207	0,826	0,411	0,952	0,091	1,043	0,861
	60 - 70%	-3,650	0,690	0,428	0,211	0,640	0,216	0,952380	0,091	1,043	0,861
	70 - 80%	-4,377	0,666	0,333	0,201	0,534	0,131	1	0	1	1
	80 - 90%	-4,935	0,880	0,761	0,182	0,944	0,579	1	0	1	1
	90 - 100%	-5,783	0,766	0,222	0,271	0,493	-0,049	1	0	1	1
	Promedio		**0,737**	**0,446**	**0,203**	**0,649**	**0,242**	**0,969**	**0,052**	**1,022**	**0,917**

Tabla 25: Resumen de datos obtenidos al clasificar las muestras empleando los puntos de corte relativos al set a) Media menos 2veces la desviación típica. Valor Umbral: Valor del punto de corte entre categorías según el método estudiado; Error_Se: Tasa de error de la sensibilidad; IC_Se+: Valor superior del intervalo de confianza al 95% para la sensibilidad; Error_Esp: tasa de error de la especificidad; IC_Se-: Valor inferior del intervalo de confianza al 95% para la sensibilidad; Error_Esp: Tasa de error de la especificidad; IC_Esp+: Valor superior del intervalo de confianza al 95% para la especificidad; Error_esp: tasa de error de la especificidad; IC_esp-: Valor inferior del intervalo de confianza al 95% para la especificidad.

3er Cuartil		Valor umbral	Índice exactitud	Sensibilidad	Error_Se	IC_Se+	IC_Se-	Especificidad	Error_Esp	IC_Esp+	IC_Esp-
Mosaicismo con trisomía	**100 - 90%**	3,315	0,619	0,933	0,126	1,059	0,807	0,733	0,223	0,957	0,509
	90 - 80%	3,05	0,7	0,8	0,202	1,002	0,597	0,642	0,250	0,893	0,391
	80 - 70%	2,704	0,866	0,733	0,223	0,957	0,509	0,733	0,223	0,957	0,509
	70 - 60%	2,428	0,733	0,666	0,238	0,905	0,428	0,733	0,223	0,957	0,509
	60 - 50%	1,902	0,9	0,533	0,252	0,785	0,280	0,666	0,238	0,905	0,428
	50 - 40%	1,684	0,6	1	0	1	1	0,8	0,202	1,002	0,597
	40 - 30%	1,435	0,7	0,733	0,223	0,957	0,509	0,733	0,223	0,957	0,509
	30 - 20%	0,949	0,733	0,933	0,126	1,059	0,807	0,8	0,202	1,002	0,597
	20 - 10%	0,559	0,724	0,6	0,247	0,847	0,352	0,8	0,202	1,002	0,597
	10 - 0%	0,122	0,833	0,333	0,377	0,710	-0,0438	0,733	0,223	0,957	0,509
	Promedio		**0,740**	**0,726**	**0,201**	**0,928**	**0,524**	**0,737**	**0,221**	**0,959**	**0,516**
Mosaicismo con monosomía	**0- 10%**	-0,370	0,722	0,714	0,193	0,907	0,521	0,733	0,223	0,957	0,509
	10 - 20%	-0,639	0,804	0,904	0,125	1,030	0,779	0,7	0,200	0,9008	0,499
	20 - 30%	-1,122	0,833	0,904	0,125	1,030	0,779	0,761	0,182	0,944	0,579
	30 - 40%	-1,609	0,880	0,904	0,125	1,030	0,779	0,857	0,149	1,006	0,707
	40 - 50%	-2,115	0,928	1	0	1	1	0,857	0,149	1,006	0,707
	50 - 60%	-2,646	0,833	0,904	0,125	1,030	0,779	0,761	0,182	0,944	0,579
	60 - 70%	-3,255	0,761	0,809	0,1679	0,977	0,641	0,714	0,1932	0,907	0,521
	70 - 80%	-3,898	0,785	0,857	0,149	1,006	0,707	0,714	0,193	0,907	0,521
	80 - 90%	-4,501	0,857	0,952	0,091	1,0434	0,861	0,761	0,182	0,944	0,579
	90 - 100%	-5,325	0,766	0,666	0,3077	0,9744	0,3586	0,809	0,167	0,977	0,641
	Promedio		**0,817**	**0,861**	**0,141**	**1,003**	**0,720**	**0,767**	**0,182**	**0,949**	**0,584**

Tabla 26: Resumen de datos obtenidos al clasificar las muestras empleando los puntos de corte relativos al set b) Tercer cuartil. Valor Umbral: Valor del punto de corte entre categorías según el método estudiado; Error_Se: Tasa de error de la sensibilidad; IC_Se+: Valor superior del intervalo de confianza al 95% para la sensibilidad; Error_Esp: tasa de error de la especificidad; IC_Se-: Valor inferior del intervalo de confianza al 95% para la sensibilidad; Error_Esp: Tasa de error de la especificidad; IC_Esp+: Valor superior del intervalo de confianza al 95% para la especificidad; Error_esp: tasa de error de la especificidad; IC_esp-: Valor inferior del intervalo de confianza al 95% para la especificidad