Abordaje de las bases moleculares de los trastornos plaquetarios congénitos integrando el estudio del genoma, el transcriptoma y el uso de inteligencia artificial

PSEUDÓNIMO: Puccini

Trabajo presentado a la convocatoria 2021 del "Premio al proyecto de investigación en medicina clínica" de la Real Fundación Victoria Eugenia

Abordaje de las bases moleculares de los trastornos plaquetarios congénitos integrando el estudio del genoma, el transcriptoma y el uso de inteligencia artificial.

Resumen del proyecto

Los trastornos plaquetarios congénitos (TPCs) comprenden un grupo de enfermedades heterogéneas relacionadas con la alteración en la síntesis y/o función de las plaquetas. La complejidad en el diagnóstico de los TPCs se debe principalmente a que su etiología molecular es muy diversa y en gran parte desconocida. Además, los signos y las pruebas de laboratorio son generalmente insuficientes para obtener un diagnóstico definitivo.

La llegada de la *Next Generation Sequencing* (NGS) ha sido clave para profundizar en la base molecular de estos trastornos mediante el estudio de paneles de genes o, más recientemente, del exoma completo (*Whole Exome Sequencing*; WES). A pesar de las grandes ventajas que ofrece la aplicación de la NGS, todavía es elevado el porcentaje de pacientes para los que no es posible identificar los genes y/o las mutaciones que justifiquen el fenotipo de estudio. En este escenario, el objetivo del presente proyecto se centra en profundizar en el conocimiento de las bases moleculares responsables de los TPCs mediante una aproximación holística que combina el diagnóstico clínico, los estudios genómicos y transcriptómicos, y herramientas de análisis basadas en inteligencia artificial. Los pacientes reclutados hasta la fecha ya han sido analizados mediante WES y, en algunos casos, se han priorizado variantes candidatas en base a la correlación genotipo-fenotipo. Para demostrar la implicación de estas variantes en el fenotipo, se realizarán estudios familiares y se pueden plantear estudios funcionales que pueden incluir estudios especiales de plaquetas, estudios dirigidos sobre RNA u otros análisis bioquímicos a medida. En aquellos casos en los que no se identifique gen o variante candidata, se decidirá la necesidad de llevar a cabo estudios adicionales mediante secuenciación del RNA (RNA-seq) y se evaluarán los resultados, contrastándolos con los de la WES. Por otra parte, se emplearán herramientas de análisis inteligente de los datos generados a nivel individual (paciente a paciente) o en una evaluación global de la cohorte. La finalidad de estos estudios será identificar patrones genéticos comunes en diferentes individuos que faciliten la clasificación de la etiología

molecular subyacente y el diagnóstico de esta patología en el futuro. En conclusión, el presente proyecto pretende profundizar en las bases moleculares de los TPCs, dilucidar las potenciales consecuencias de variantes candidatas y, en definitiva, avanzar hacia una medicina personalizada que mejore la calidad asistencial ofrecida a estos pacientes.

Introducción

Los trastornos plaquetarios congénitos o TPCs son defectos de la hemostasia primaria que conducen a una mayor susceptibilidad al sangrado. Si bien la diátesis hemorrágica suele manifestarse con episodios de sangrado mucocutáneo como epistaxis o menorragia, estos trastornos pueden dar lugar a hemorragias graves que podrían llegar a poner en peligro la vida del paciente, sobre todo ante traumas severos o procedimientos quirúrgicos (1). La obtención de un diagnóstico definitivo para estos pacientes se ve comprometida por la complejidad de las pruebas de laboratorio que evalúan la función plaquetaria y la necesidad de estudios genéticos para identificar el defecto molecular. Además, la falta de acuerdo sobre su clasificación es un obstáculo adicional para su diagnóstico. Aunque muchos de los trastornos plaquetarios hereditarios tienen una base genética bien conocida, en muchos otros la etiología molecular última es difícil de identificar debido a la elevada heterogeneidad de fenotipo que presentan los pacientes y la limitación que existe en varios centros en cuanto a la realización de estudios funcionales completos.

Los TPCs representan globalmente el 8% de las diátesis hemorrágicas (2). Hasta hace relativamente poco tiempo, su caracterización genética se había limitado al diagnóstico mediante secuenciación tradicional de Sanger de las afecciones más prevalentes como el síndrome de Bernard-Soulier (BSS), la Trombastenia Glanzmann (GT) y la Enfermedad de Von Willebrand de tipo plaquetario (VWDP). Sin embargo, desde la introducción de la secuenciación masiva o *Next Generation Sequencing* (NGS) como herramienta diagnóstica, su aplicación en este campo ha permitido la asociación de estas patologías con un número creciente de genes. Equipos de investigación nacionales e internacionales (1)(3)(4) han desarrollado en los últimos años estrategias de diagnóstico de TPCs basados en la secuenciación de paneles de genes. Sin embargo, la descripción constante de nuevos genes relacionados con los déficits plaquetarios pone de manifiesto la limitación

de estos paneles. Por otro lado, la progresiva reducción de los costes de secuenciación conjuntamente con la simplificación de los procedimientos para la preparación de librerías hace de la secuenciación del exoma completo (WES) una aproximación cada vez más eficiente desde el punto de vista del coste-beneficio. Además, la posibilidad de adecuar el listado de genes a analizar en función del conocimiento que vayamos obteniendo sobre el tema, así como la oportunidad de reanalizar los datos cuando algún hallazgo científico del propio proyecto o de la literatura lo aconseje, explica que la WES sea de especial interés sobre todo en el campo de la investigación de estos trastornos.

Aun así, el estudio genético basado en la WES de los pacientes con trastornos plaquetarios congénitos, en numerosas ocasiones no es suficiente para esclarecer la base genética de la patología. Al analizar solo las variantes exónicas e intrónicas flanqueantes, esta técnica no permite detectar variantes de las zonas intrónicas profundas, así como grandes deleciones, inserciones y/o reordenamientos genómicos. Como consecuencia, el estudio genético centrado en el exoma de estos pacientes es una estrategia valiosa pero no suficiente para obtener un diagnóstico definitivo en todos ellos.

Por otra parte, el estudio del RNA nos ofrece un enfoque complementario de análisis tanto a nivel de estudio del *splicing* como de expresión génica o transcriptoma. En concreto, el análisis del RNA en plaquetas y leucocitos es una herramienta muy valiosa para desvelar el efecto definitivo de ciertas mutaciones (5).

Se ha observado que las plaquetas poseen un perfil único de expresión en comparación con otros tejidos o tipos celulares (6). Pese a no tener un núcleo definido, las plaquetas poseen una compleja red de mRNA y microRNA capaz de regular la trombopoyesis y la función plaquetaria. Además, se ha demostrado que el perfil transcriptómico de las plaquetas está relacionado con diferentes patologías, de manera que se han podido identificar patrones de expresión característicos para enfermedades inflamatorias, hematológicas, oncológicas, autoinmunes, metabólicas, nefrológicas y cardiovasculares (9). De hecho, el análisis transcriptómico de las plaquetas ya es una realidad en el estudio y diagnóstico de trombocitopenias autoinmunes (8). El análisis de la expresión del RNA de las plaquetas puede representar un potencial enfoque para delinear algunos de los defectos moleculares en pacientes con trastornos plaquetarios congénitos. Con el objetivo de esclarecer el diagnóstico genético de uno de estos pacientes, Sun *et al.* (7) en

el 2006 estudió sus perfiles de expresión plaquetaria. El paciente, con manifestaciones hemorrágicas y un perfil de deficiencia de agregación plaquetaria, presentaba una mutación en *RUNX1*. Con este análisis se pudo investigar la relación entre la mutación en *RUNX1* y la fosforilación alterada de la cadena ligera de la miosina (MLC), que se debe a su vez a la disminución de la expresión del polipéptido regulador de la cadena ligera de la miosina (*MYL9*).

Por otra parte, es relevante considerar la participación de los microRNAs en el control de la expresión génica, puesto que se estima que más del 60% de los genes humanos codificantes están regulados por microRNAs. Mediante estudios de expresión de estos pequeños fragmentos de RNA no codificante se ha podido demostrar su implicación en el proceso de megacariopoyesis y biogénesis de plaquetas, así como en la regulación de la función plaquetaria controlando la expresión de genes como *VAMP8* (10).

El presente proyecto se fundamenta en el estudio genético de pacientes con sospecha de TPCs mediante una estrategia holística basada en el estudio genómico (WES) y transcriptómico (mRNA y microRNA), utilizando herramientas de análisis inteligente de datos. Para la gestión del gran volumen de información obtenida mediante estos análisis se emplearán herramientas informáticas de elevado rendimiento que permitirán dotar de sentido biológico y extraer conclusiones de los resultados obtenidos al contrastarlos con la información clínica y biológica disponible.

Bibliografía

1. Gresele P, Falcinelli E, Bury L. Laboratory diagnosis of clinically relevant platelet function disorders. International journal of laboratory hematology. 2018;40 Suppl 1:34-45.
2. Sivapalaratnam S, Collins J, Gomez K. Diagnosis of inherited bleeding disorders in the genomic era. British journal of haematology. 2017;179(3):363-76.
3. Simeoni I, Stephens JC, Hu F, Deevi SV, Megy K, Bariana TK, et al. A high-throughput sequencing test for diagnosing inherited bleeding, thrombotic, and platelet disorders. Blood. 2016;127(23):2791-803.
4. Bastida JM, Lozano ML, Benito R, Janusz K, Palma-Barqueros V, Del Rey M, et al. Introducing high-throughput sequencing into mainstream genetic diagnosis practice in inherited platelet disorders. Haematologica. 2018;103(1):148-62.
5. Borras N, Orriols G, Batlle J, Perez-Rodriguez A, Fidalgo T, Martinho P, et al. Unraveling the effect of silent, intronic and missense mutations on VWF splicing: contribution of next generation sequencing in the study of mRNA. Haematologica. 2019;104(3):587-98.
6. Kissopoulou A, Jonasson J, Lindahl TL, Osman A. Next generation sequencing analysis of human platelet PolyA+ mRNAs and rRNA-depleted total RNA. PloS one. 2013;8(12):e81809.
7. Sun L, Gorospe JR, Hoffman EP, Rao AK. Decreased platelet expression of myosin regulatory light chain polypeptide (MYL9) and other genes with platelet dysfunction and CBFA2/RUNX1 mutation: insights from platelet expression profiling. Journal of thrombosis and haemostasis : JTH. 2007;5(1):146-54.
8. Hernandez-Sanchez JM, Bastida JM, Alonso-Lopez D, Benito R, Gonzalez-Porras JR, De Las Rivas J, et al. Transcriptomic analysis of patients with immune thrombocytopenia treated with eltrombopag. Platelets. 2020;31(8):993-1000.
9. Gutmann C, Joshi A, Mayr M. Platelet "-omics" in health and cardiovascular disease. Atherosclerosis. 2020;307:87-96.
10. Dangwal S, Thum T. MicroRNAs in platelet biogenesis and function. Thrombosis and haemostasis. 2012;108(4):599-604.

Racional

La aplicación de las nuevas herramientas moleculares de NGS, como la WES, nos ha permitido abordar de manera sistemática y unificada el estudio genético de los TPCs, anteriormente limitado a unos pocos genes mediante secuenciación tradicional de Sanger. Ello está impulsando el conocimiento de la etiología molecular de estas patologías abriendo posibilidades inabordables anteriormente. Sin embargo, el diagnóstico genético mediante WES se centra en mutaciones puntuales, limitando la detección de grandes deleciones, inserciones o reordenamientos, así como de mutaciones intrónicas o del promotor que pueden tener implicaciones en el procesamiento de la proteína.

Actualmente disponemos de una cohorte de 75 pacientes con trastornos plaquetarios bien caracterizada a nivel clínico y con el estudio de WES ya realizado. Asimismo, hemos identificado un amplio abanico de genes y mutaciones candidatas, la mayoría de las cuales no han sido descritas previamente, lo que conlleva diferentes grados de certidumbre: desde el diagnóstico de certeza hasta no tener ningún gen candidato. El estudio del efecto patogénico de las variantes mediante estudios funcionales y el análisis de transcriptoma mediante técnicas de RNA-seq, puede resultar de gran interés para dilucidar su contribución en el establecimiento de trastornos plaquetarios congénitos.

La hipótesis del proyecto se centra en la idea de que las nuevas herramientas moleculares que nos permiten analizar la funcionalidad de las mutaciones, el transcriptoma, los microRNAs, etc., unidas a la posibilidad actual de utilizar algoritmos de inteligencia artificial en estos análisis abren un nuevo escenario que contribuirá a profundizar en las bases moleculares de los TPCs, dilucidar la consecuencia de variantes candidatas y, en definitiva, dirigirse hacia una medicina personalizada que mejore la calidad asistencial ofrecida a estos pacientes.

Objetivos

El objetivo general del proyecto persigue profundizar en el conocimiento de las bases moleculares responsables de los TPCs mediante herramientas de secuenciación masiva y análisis genómico (Figura 1). Para ello, disponemos de una cohorte de pacientes que se irá ampliando progresivamente en los próximos años. Sobre esta cohorte se explorará la aplicación de las nuevas tecnologías de secuenciación al diagnóstico molecular e investigación de estos trastornos hemorrágicos, que en la mayoría de los casos tienen una base genética poco conocida o absolutamente inédita. Con esta orientación se combinarán estrategias de WES y secuenciación del transcriptoma con el fin de esclarecer las bases moleculares y diseñar e implementar un procedimiento holístico y versátil para el estudio genético integral de los trastornos plaquetarios. Este objetivo global se desglosa en los siguientes objetivos específicos:

1. Aplicación de WES al estudio molecular de los pacientes con diátesis hemorrágica y sospecha de trastorno plaquetario congénito que se vayan reclutando en la cohorte.

2. Identificación de variantes candidatas y validación de las mismas mediante secuenciación tradicional tanto en el paciente como en los familiares para hacer estudios de segregación. Clasificación y estratificación de los pacientes en base a las variantes encontradas y a la correlación genotipo-fenotipo, estudios *in silico,* estudios funcionales especiales adicionales (agregación, citometría, etc.).

3. Determinación y realización de los estudios genómicos ampliados necesarios para cada caso: análisis del transcriptoma; estudio de microRNAs, análisis funcional del procesamiento de genes concretos.

4. Integración y análisis basados en inteligencia artificial de los datos clínicos, de laboratorio, genotípicos y transcripcionales. Valoración final y diagnóstico de los pacientes. Elaboración de informes.

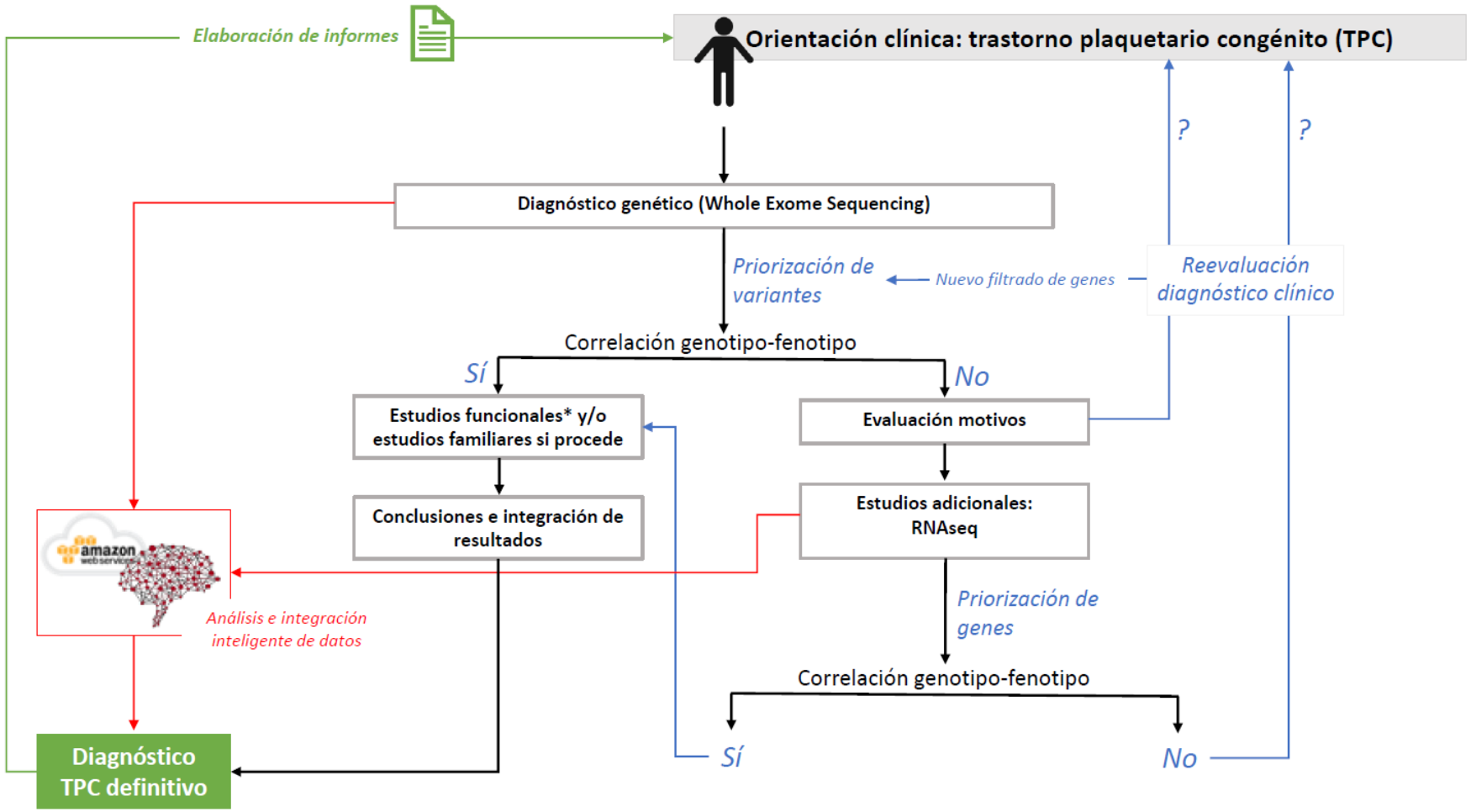

Figura 1. Representación esquemática del diseño del estudio.

Material y métodos

Diseño: estudio molecular epidemiológico, analítico, retrospectivo y prospectivo, a partir de muestras de sangre y/o ácidos nucleicos (DNA y/o RNA) de pacientes y familiares con diátesis hemorrágica de base genética poco conocida o heterogénea relacionada con TPCs. La selección de pacientes, el reclutamiento, el estudio de los diversos parámetros clínicos y la extracción de las muestras de sangre periférica se realiza en diferentes complejos hospitalarios del país.

Sujetos: En este proyecto se incluyen pacientes reclutados de manera retrospectiva y prospectiva con un diagnóstico o sospecha de trombocitopenia o trombocitopatía. Hasta la fecha se han reclutado un total de 75 pacientes con TPC, orientados clínicamente de manera muy heterogénea como: plaquetopatías, trombocitopenias, macrotrombocitopenias, trastornos plaquetarios, alteración en la liberación de gránulos o alteración en la agregación.

Muestras biológicas: Para todos los pacientes reclutados, así como para los familiares relacionados que se incluyan en el estudio, se dispondrá de una muestra de sangre total a partir de la cual se obtendrá el DNA genómico. La extracción de DNA se llevará a cabo mediante el kit QIAsymphony DNA Mini utilizando el sistema QIAsymphony SP (Qiagen). Además, se obtendrá RNA a partir de sangre total, plaquetas y leucocitos. La extracción y conservación de sangre total se realizará en tubos PAXgene, que permiten la estabilización del RNA intracelular. A continuación, la purificación del RNA, que incluye miRNAs, se llevará a cabo mediante el kit QIAsymphony PAXgene Blood RNA en el sistema QIAsymphony SP (Qiagen). Para la obtención de RNA total a partir de leucocitos y plaquetas, se centrifugará la muestra para separar los diferentes tipos celulares y se utilizará el kit QIAamp RNA Blood (Qiagen).

WES: La secuenciación del exoma completo de los pacientes reclutados prospectivamente se llevará a cabo mediante la misma metodología aplicada en los pacientes reclutados retrospectivamente de los que ya disponemos de resultados. En concreto, la WES se realizará mediante el kit Nextera Flex for Enrichment, actualmente

denominado Illumina DNA Prep with Enrichment, siguiendo las instrucciones del fabricante (Illumina). Las librerías se procesaran mediante tecnología de secuenciación *paired-end* en el sistema Illumina NextSeq 500 utilizando el kit NextSeq 500/550 High Output v2.5 (Illumina) de 150 ciclos (2x74), de acuerdo con el protocolo estándar de Illumina.

Análisis de variantes candidatas: Se llevará a cabo un análisis bioinformático para la obtención de variantes genéticas candidatas relacionadas con el fenotipo de estudio a partir de los resultados de la WES. En concreto, se utilizará la plataforma BaseSpace Sequence Hub y la aplicación BWA Enrichment (Illumina) para realizar el alineamiento de las lecturas obtenidas con la secuencia de referencia (GRCh37, hg19) y la identificación de las variantes genéticas.

La anotación, clasificación y filtrado de las variantes se llevará a cabo mediante los programas BaseSpace Variant Interpreter y VariantStudio Data Analysis (Illumina). Se tendrán en cuenta parámetros de calidad, frecuencia poblacional y la consecuencia de las variantes. Además, el uso de un panel de genes, a partir de información incluida en distintas bases de datos y bibliografía, permite filtrar los resultados obtenidos en función de la orientación clínica y el fenotipo presentado por los pacientes de este estudio.

Estudios *in silico*: El impacto de las mutaciones *missense* será evaluado por medio de algoritmos como PolyPhen-2, SIFT, Align y GVGD. Para la valoración de variantes que afectan potencialmente al *splicing* se usarán los programas SpliceSiteFinder, MaxEntScan, GeneSplicer, ESEfinder y los métodos RESCUE-ESE.

Validación: La validación técnica de las variantes genéticas candidatas potencialmente relacionadas con el fenotipo de estudio, se hará mediante el diseño de *primers* específicos, la optimización de los protocolos de amplificación y la secuenciación por *Sanger* de las regiones de interés. La puesta a punto de estos protocolos permitirá la realización de estudios familiares en aquellos casos que sea necesario o aconsejable.

Estudios funcionales en genes candidatos: Los estudios funcionales mediante mRNA podrán aplicarse a variantes genéticas potencialmente patogénicas en aquellos genes

que se expresen en sangre total, leucocitos y/o plaquetas. Se diseñarán *primers* específicos para identificar el efecto de las variantes que puedan relacionarse con el *splicing*. Las RT-PCRs resultantes serán secuenciadas por Sanger y/o NGS.

RNA-seq: El análisis del transcriptoma mediante la aproximación RNA-seq se llevará a cabo para muestras de leucocitos, plaquetas y/o sangre total. Para el análisis de las muestras de sangre total, se eliminará previamente el mRNA de globina no deseado mediante el kit GLOBINclear (Invitrogen). Para la preparación de las librerías se utilizará el kit Illumina Stranded mRNA Prep (Illumina). Las librerías se procesarán mediante tecnología de secuenciación *paired-end* en el sistema Illumina NextSeq 500 utilizando el kit NextSeq 500/550 High Output v2.5 (Illumina) de 150 ciclos, de acuerdo con el protocolo estándar de Illumina.

El estudio de la fracción de miRNAs se realizará con el kit QIAseq miRNA Library (Qiagen). Las librerías se procesaran en el sistema Illumina NextSeq 500 o Illumina MiSeq, de acuerdo con el protocolo estándar de Illumina.

Análisis Bioinformáticos y estadísticos: Se llevará a cabo una comparativa de distintos análisis bioinformáticos para evaluar los resultados derivados de los estudios de RNA-seq. Entre otros, se utilizará la plataforma BaseSpace Sequence Hub y la aplicación DRAGEN RNA Pipeline (Illumina), así como el programa CLC Genomics Workbench (Qiagen), para el análisis secundario y terciario de los datos obtenidos.

Inicialmente, se caracterizará el perfil de expresión específico de cada individuo. Además, se realizarán estudios de co-expresión entre RNA mensajero y microRNAs en cada paciente para identificar las posibles relaciones entre ambas moléculas. Por otro lado, se evaluará la existencia de patrones de expresión en función de la patología plaquetaria subyacente mediante estudios de expresión diferencial. Además, se tratará de identificar posibles asociaciones entre los mRNAs y microRNAs más informativos a nivel de grupo fenotípico. Debido a las limitaciones que presentan las herramientas analíticas clásicas utilizadas en este tipo de abordajes, también se aplicarán métodos de aprendizaje automatizado con el objetivo de identificar patrones de co-expresión entre individuos con fenotipos plaquetarios similares. Los distintos análisis descritos en los párrafos

anteriores se realizarán también mediante soluciones personalizadas basadas en los lenguajes R y Python, y distintas librerías específicas para este tipo de análisis.

Utilidad de los resultados

De acuerdo con los objetivos planteados, los resultados esperables de este proyecto tendrán una traslación inmediata a la práctica clínica. La identificación de nuevos genes y variantes implicadas en los TPCs permitirá profundizar en el desarrollo de paneles de diagnóstico genético rápido, fiable y económico. El presente proyecto pretende profundizar en la investigación de los fundamentos biológicos de los TPCs y en el desarrollo de nuevos algoritmos diagnósticos basados en la información genómica tratada mediante herramientas de *machine learning* e inteligencia artificial. Todo ello tiene como finalidad última mejorar el conocimiento del riesgo hemorrágico con el objetivo de facilitar una atención clínica adecuada e introducir innovación tecnológica en el campo asistencial, impulsando mejoras en la salud y calidad de vida de los pacientes. Consideramos que la naturaleza de este proyecto, en el que colaboran diferentes centros asistenciales, permite el posicionamiento de la investigación española en la vanguardia del conocimiento en este campo. Más allá de los aspectos epidemiológicos, el diagnóstico genético representa una herramienta básica para el estudio de los pacientes con implicaciones terapéuticas evidentes. Asimismo, los criterios para la inclusión de pacientes en ensayos clínicos convencionales o protocolos de terapia génica deben contar con una documentación completa de la mutación causante. En definitiva, se avanza en la dirección de la búsqueda de una medicina personalizada. La propuesta se alinea coherentemente con esta directriz actual de la biomedicina que permite rentabilizar las inversiones en investigación desde un punto de vista no solo económico sino también social.