UNIVERSITAT DE BARCELONA

Análisis Genómico de Fenotipos de la Hemostasia Relacionados con la Enfermedad Tromboembólica Venosa

Laura Martín Fernández

Análisis Genómico de Fenotipos de la Hemostasia Relacionados con la Enfermedad Tromboembólica Venosa

Memoria presentada por:
Laura Martín Fernández

Para optar al grado de:
Doctor por la Universitat de Barcelona

Tesis realizada bajo la dirección del
Dr. José Manuel Soria Fernández
en la Unitat de Genòmica de Malalties Complexes del Institut de Recerca de l'Hospital de la Santa Creu i Sant Pau - Institut d'Investigació Biomèdica Sant Pau (IIB Sant Pau)

Tesis adscrita a la Facultad de Biología, Universitat de Barcelona (UB)
Programa de Doctorado en Genética
Tutor: **Dr. Daniel Raúl Grinberg Vaisman**

José Manuel Soria Fernández **Daniel Raúl Grinberg Vaisman** **Laura Martín Fernández**

Barcelona, 2017

Agradecimientos

Agradecimientos

Quiero expresar mi agradecimiento a todas las personas que han hecho posible la realización de esta Tesis Doctoral. Los que me conocéis sabéis que soy una persona reservada y que no suelo demostrar lo que siento por escrito. Aún así, espero que todos los que durante estos años me habéis ayudado os sintáis representados.

Principalmente, agradezco al Dr. José Manuel Soria y al Dr. Daniel Grinberg su especial dedicación y atención en la dirección y tutela de la presente Tesis. Gracias por la paciencia y por permitirme acudir a vosotros ante cualquier problema.

Quiero agradecer también la financiación parcial de este trabajo recibida por el Instituto de Salud Carlos III – Fondo de Investigación Sanitaria (PI 11/0184, PI 12/01494, PI 14/0582 y PI 15/00269), la Red de Investigación Cardiovascular (RD12/0042/0032 y RD12/0042/0053), la Agència de Gestió d'ajuts Universitaris i de Recerca (AGAUR SGR-1147, SGR-01068, SGR 1063 y SGR-1240) y el Ministerio de Economia y Competitividad a través de TEC2014-60337-R. Especialmente me gustaría dar las gracias al Instituto de Salud Carlos III por las Ayudas Predoctorales de Formación en Investigación en Salud (PFIS FI12/00322).

Gracias a todos mis compañeros de la Unidad de Genómica de Enfermedades Complejas por su apoyo y el aprendizaje recibido, y también a la Unidad de Trombosis y Hemostasia por su colaboración. En especial, al Dr. Juan Carlos Souto y a la Dra. Sonia López por su implicación en el progreso de mi carrera científica y por la ayuda

prestada en la elaboración de esta Tesis. Gracias por todas nuestras conversaciones, laborales e incluso personales, que tanto me han ayudado en ciertos momentos.
También gracias a todos mis compañeros del laboratorio de Coagulopatías Congénitas por permitirme compartir esta etapa con vosotros dentro y fuera del trabajo, especialmente al Dr. Francisco Vidal y a la Dra. Irene Corrales por sus consejos, orientación y por su atención a mis infinitas consultas. Gracias a todos vosotros y al laboratorio de Histocompatibilidad por hacerme sentir como una más de vuestro grupo.
Quiero agradecer asimismo al Dr. Pascual Marco la confianza depositada en mí y su participación en este estudio.

Gracias a mis compañeros y amigos doctorandos Alba, Anna, Emma, Jessica, Jordi, Mar, Miquel, Naiara, Nina y Víctor por vuestros consejos, comprensión y distracciones también necesarias. Vivir esta experiencia juntos ha sido la mejor parte.
Por supuesto, muchas gracias en general a todos mis amigos, a los de siempre y a los más recientes, a aquellos que por suerte puedo ver y abrazar cada día y a los que muy a mi pesar os tengo un poco más lejos pero muy presentes. No hubiera sido lo mismo sin vosotros. Gracias a mis cuquis compis de piso por la cantidad infinita de achuchones y a mi hermana postiza que tanto me conoce y necesito.

Mil gracias a mi familia, en concreto a mis padres y a mi hermana a los que tanto quiero, habéis sido mi principal sustento y alegría.

Y a ti, mi mejor amigo.

Muchas gracias.

Índice

Índice

Abreviaturas

Abreviaturas

3'ss: sitio aceptor de *splicing*

5'ss: sitio donador de *splicing*

ρ_a: correlación ambiental

ρ_f: correlación fenotípica

ρ_g: correlación genética

ACGS: asociación para la ciencia genética clínica

ADP: adenosín difosfato

ADPasa: adenosín difosfatasa

APC: proteína C activada

APCR: resistencia a la proteina C activada

aPTT: tiempo de tromboplastina parcial activada

AT: antitrombina

ATE: enfermedad tromboembólica arterial

bp: par de bases

C4BP: proteína de unión a C4b

Ca^{2+}: calcio

ddNTPs: didesoxinucleótidos

DNA: ácido desoxirribonucleico

DVT: trombosis venosa profunda

EMBL-EBI: instituto europeo de bioinformática del laboratorio europeo de biología molecular

emPCR: amplificación por reacción en cadena de la polimerasa en emulsión

ABREVIATURAS

EPCR: receptor endotelial de la proteina C

ETP: potencial endógeno de trombina

FI: factor I de la coagulación o fibrinógeno

FIa: factor I de la coagulación activado o fibrina

FII: factor II de la coagulación o protrombina

FIIa: factor II de la coagulación activado o trombina

FV: factor V de la coagulación

FVa: factor V de la coagulación activado

FVL: factor V Leiden

FVII: factor VII de la coagulación

FVIIa: factor VII de la coagulación activado

FVIII: factor VIII de la coagulación

FVIIIa: factor VIII de la coagulación activado

FIX: factor IX de la coagulación

FIXa: factor IX de la coagulación activado

FX: factor X de la coagulación

FXa: factor X de la coagulación activado

FXI: factor XI de la coagulación

FXIa: factor XI de la coagulación activado

FXII: factor XII de la coagulación

FXIIa: factor XII de la coagulación activado

FXIII: factor XIII de la coagulación

FXIIIa: factor XIII de la coagulación activado

FDPs: productos de degradación de fibrina

GAIT-1: *Genetic Analysis of Idiopathic Thrombophilia* 1

GAIT-2: *Genetic Analysis of Idiopathic Thrombophilia* 2

Gb: gigabase

GPIb: receptor plaquetario glicoproteína Ib

GPIIb/IIIa: receptor plaquetario glicoproteína IIb/IIIa

GWAS: estudio de asociación de genoma completo

h: hora

h^2: heredabilidad

HMWK: quininógeno de alto peso molecular

HRG: glicoproteína rica en histidina

IBD: idénticos por descendencia

ICA: análisis de componentes independientes

IPAQ: cuestionario internacional de actividad física

LD: desequilibrio de ligamiento

LOD: logaritmo en base 10 de la probabilidad de ligamiento

LR-PCR: amplificación larga por reacción en cadena de la polimerasa.

MAF: frecuencia del alelo minoritario

mRNA: ácido ribonucleico mensajero

NGS: secuenciación de nueva generación

NHGRI: instituto nacional de investigación del genoma humano

NS: resultado no estadísticamente significativo

OR: odds ratio

PAI-1: inhibidor del activador de plasminógeno

PC: proteína C

PCR: amplificación por reacción en cadena de la polimerasa

PE: embolia de pulmón

PL: fosfolípidos

SE: error estándar

PS: proteína S

PT: tiempo de protrombina

PZ: proteína Z

QTL: sitio cromosómico de un carácter cuantitativo

RNA: ácido ribonucleico

RR: riesgo relativo

SNP: polimorfismos de un solo nucleótido

TAFI: inhibidor de la fibrinólisis activable por trombina

TF: factor tisular

TFPI: inhibidor del factor tisular

TGT: test de generación de trombina

t-PA: activador tisular de plasminógeno

u-PA: activador tipo urocinasa de plasminógeno

USD: dólar americano

UTR: región no traducida

VKA: antagonista de la vitamina K

VTE: enfermedad tromboembólica venosa

vWF: factor de von Willebrand

WES: secuenciación de todo el exoma

WGS: secuenciación de todo el genoma

ZPI: inhibidor de la proteína Z

Figuras y Tablas

Figuras y Tablas

Introducción

Introducción

1. Sistema hemostático

El sistema hemostático es un mecanismo fisiológico complejo de defensa que mantiene la sangre circulando libremente en un estado fluido dentro del sistema vascular. A la vez, en caso de daño vascular, previene los sangrados mediante la formación de un coágulo sanguíneo o trombo, así como se provoca su posterior disolución antes de la obstrucción del sistema vascular con el fin de mantener un flujo normal. Los principales componentes del sistema hemostático son el sistema vascular, las plaquetas, el sistema de la coagulación y la fibrinólisis. Alteraciones del balance entre procesos procoagulantes y anticoagulantes pueden desencadenar fenómenos trombóticos o hemorrágicos (Austin 2013).

1.1 Hemostasia primaria

Éste es el mecanismo mediante el cual se lleva a cabo la rápida respuesta del sistema vascular y de las plaquetas en caso de daño vascular, con el objetivo de producir constricción vascular y el tapón plaquetario.

El sistema vascular está formado por vasos sanguíneos recubiertos en su interior por una fina capa de células endoteliales o endotelio. Las funciones del endotelio son diversas, como separar los componentes de la sangre del tejido procoagulante subendotelial y regular el tono vascular o su permeabilidad (Herrmann and Lerman 2001). En condiciones fisiológicas normales, el endotelio se mantiene en un estado anticoagulante. En concreto, la carga negativa de la superficie endotelial evita la adhesión de las plaquetas por repulsión y la secreción de substancias como el

monóxido de nitrógeno, la prostaciclina y la adenosín difosfatasa (ADPasa) protege contra la activación plaquetaria. Por otra parte, la inhibición de la coagulación se mantiene a través de la síntesis de trombomodulina, sulfato de heparina y el inhibidor del factor tisular (TFPI), así como se regula la fibrinólisis mediante la producción de activadores implicados en la disolución del trombo como el activador tisular de plasminógeno (t-PA). En el momento de daño vascular, se expone el tejido subendotelial y se altera el balance hemostático. En este caso, se promueve la vasoconstricción y la adhesión y activación de plaquetas y leucocitos. Además, se activa la coagulación mediante la exposición de factor tisular (TF) subendotelial y su expresión en la superficie de las células endoteliales, así como se limita la fibrinólisis mediante la secreción del inhibidor del activador de plasminógeno (PAI-1) (Austin 2013; Jobling and Eyre 2013).

Las plaquetas son los componentes celulares sanguíneos más pequeños. Éstas presentan una forma discoide, carecen de núcleo y están originadas en los megacariocitos (George 2000). Las plaquetas no se adhieren al endotelio en condiciones fisiológicas normales. En caso de daño vascular, su principal función consiste en la rápida formación del tapón plaquetario, que es una red compacta de plaquetas. Por lo tanto, el receptor plaquetario glicopoteína Ib (GPIb) se adhiere mediante el factor de von Willebrand (vWF) al colágeno subendotelial expuesto. El vWF es una glicoproteína multimérica presente en plasma, en los gránulos alfa de las plaquetas y en el subendotelio (Ruggeri 2001; Austin 2013). Esta interacción induce la activación de las plaquetas, las cuales sufren cambios morfológicos característicos como la transformación de forma discoide a esférica, la emisión de seudópodos, que aumentan la superficie de adhesión, y la exposición de fosfolípidos (PL) de membrana cargados negativamente y otros receptores, que aportan una superficie de interacción con factores de la coagulación (Heemskerk et al. 2002; Austin 2013). Además, se libera el contenido de sus gránulos alfa y densos. La secreción de moléculas como

tromboxano A2, calcio (Ca^{2+}), adenosín difosfato (ADP) y serotonina estimulan la activación de otras plaquetas y la vasoconstricción. Así mismo, el receptor glicoproteína IIb/IIIa (GPIIb/IIIa) sufre un cambio conformacional que permite la unión de moléculas como fibrinógeno, fibronectina y vWF. De esta manera, y una vez se ha formado la primera capa, se forman puentes de unión entre las plaquetas durante el proceso de agregación para formar el tapón plaquetario, que es temporal y se coordina con el sistema de la coagulación (hemostasia secundaria) para la formación del trombo (Dahlbäck 2005).

1.2 Hemostasia secundaria

Esta fase corresponde a la respuesta del sistema de la coagulación al daño vascular para la formación del trombo, que es una estructura formada por una red de fibrina que atrapa a distintos componentes sanguíneos entre los cuales se encuentran bloqueadas las plaquetas.

El sistema de la coagulación está formado por una serie de proteínas o factores de la coagulación que se identifican mediante números romanos del I al XIII. La síntesis de la mayoría de estos factores se lleva a cabo principalmente en el hígado (Peck-Radosavljevic 2007). En la coagulación los zimógenos inactivos se activan secuencialmente mediante una serie de reacciones enzimáticas con retroalimentación positiva y negativa que constituyen la cascada de la coagulación. Mediante el sistema de la coagulación se produce la transformación de protrombina o factor II (FII) a trombina o FII activado (FIIa), que conduce a la conversión de fibrinógeno o factor I (FI) en fibrina o FI activado (FIa) para la creación del trombo. La explicación de este sistema a partir del modelo de la cascada de la coagulación ha sido hasta la fecha de gran utilidad clínica, descrito en 1964 por MacFarlane (Macfarlane 1964) y por Davie y Ratnoff (Davie and Ratnoff 1964). Sin embargo, a partir del estudio de las deficiencias

de los factores de la coagulación se ha desarrollado un nuevo modelo basado en la participación conjunta de las células y proteínas de la coagulación que se emplea en la explicación *in vivo* del sistema.

1.2.1 Cascada de la coagulación

El modelo tradicional de la cascada de la coagulación se basa en distintas vías: la intrínseca, la extrínseca y la común. En este modelo, la coagulación se puede originar de manera independiente tanto a partir de la vía intrínseca como a partir de la vía extrínseca. La Figura 1 representa el modelo simplificado de la cascada de la coagulación.

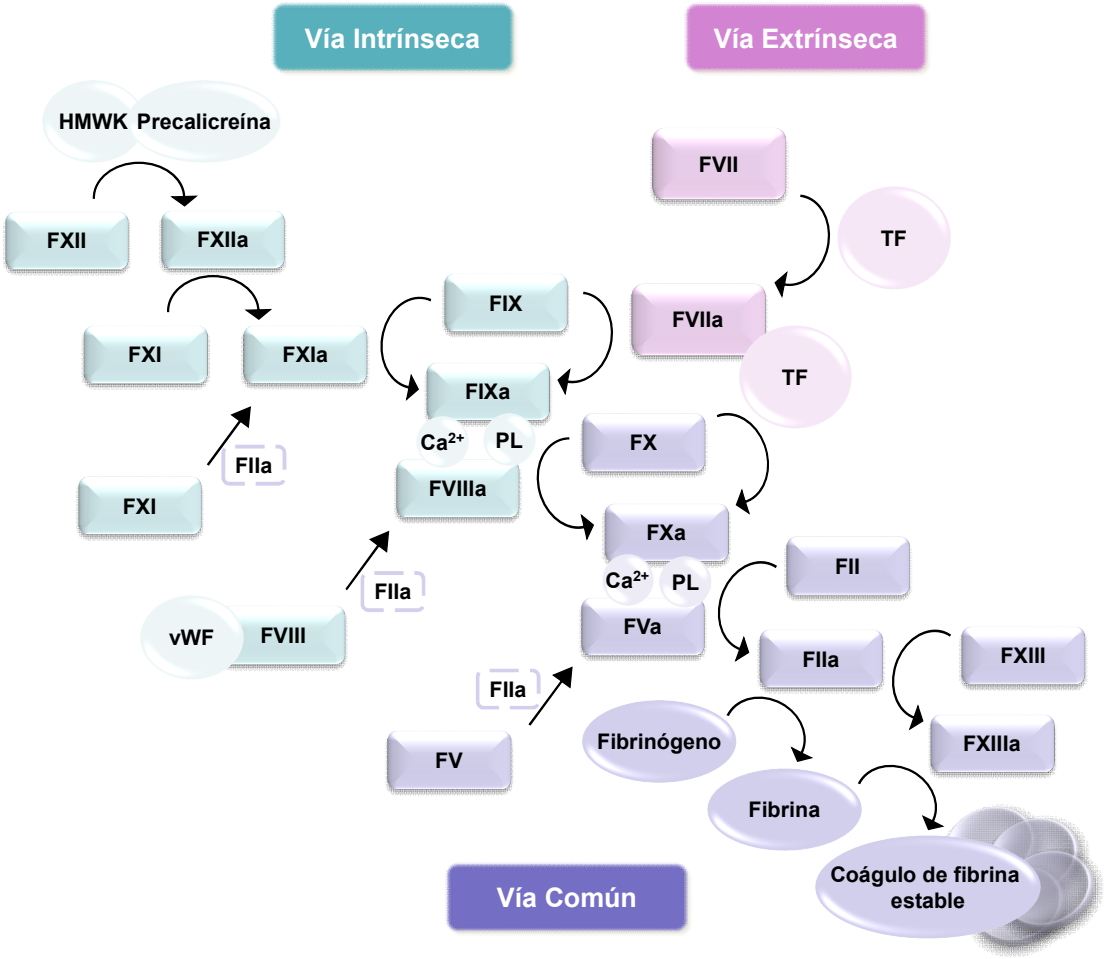

Figura 1. Cascada de la coagulación. Adaptado (Norris 2003).

1.2.1.1 Vía intrínseca

Está compuesta por el factor XII (FXII), la precalicreína, el quininógeno de elevado peso molecular (HMWK), el factor XI (FXI), el factor IX (FIX) y el factor VIII (FVIII). También se conoce como vía de activación por contacto.

Esta vía se inicia por el contacto de los componentes presentes en la sangre con superficies de carga negativa, como los PL de las plaquetas o el colágeno y, de esta manera, el FXII se autoactiva (FXIIa). Asimismo, la precalicreína, que se encuentra unida en plasma al HMWK, se sitúa en estas superficies mediante la asociación del HMWK con las cargas negativas. Esto permite que el FXIIa active a la precalicreína convirtiéndola en calicreína, y a su vez, que la calicreína actúe también sobre el FXII para generar FXIIa y amplificar así el proceso. Además, la calicreína produce el vasoactivo bradicinina a partir del HMWK. Por otra parte, el FXI circula en plasma formando un complejo junto al HMWK. Por lo tanto, también se localiza en estas superficies de carga negativa, lo que posibilita que el FXIIa genere FXI activo (FXIa) y que éste actúe sobre el FIX activándolo (FXIa) (Bouma and Meijers 1999; Gailani and Renné 2007). El FIXa forma junto al FVIII activado (FVIIIa), Ca^{2+} y PL el complejo tenasa, que permite la conversión del factor X (FX) a FX activado (FXa) (Norris 2003).

Respecto al FVIII, éste circula en plasma formando un complejo con el vWF que previene su degradación y se libera una vez es activado por la trombina. De forma alternativa, el FXI también puede ser activado por la trombina, lo que genera un proceso de retroalimentación positiva. Tanto el FIX como el FX son factores de la coagulación dependientes de vitamina K (Dahlbäck 2000).

Cabe destacar que la bradicinina generada está también implicada en el proceso de fibrinólisis mediante la estimulación de las células endoteliales para la generación de t-PA. Además, inhibe la activación mediada por trombina de las plaquetas (Sidelmann et al. 2000).

1.2.1.2 Vía extrínseca

El componente principal de la vía extrínseca es el factor VII (FVII), que es una proteína plasmática vitamina K dependiente. Esta vía se inicia por el contacto del TF expuesto, como consecuencia de la lisis celular en caso de daño vascular, tanto con el FVII como con el FVII activado (FVIIa), ya que una parte del FVII también puede circular en plasma en su forma activa. Este complejo enzimático potencia la activación del FIX y del FX (Mann et al. 2003; Crawley et al. 2011) que son elementos esenciales para continuar con el proceso de formación del trombo.

1.2.1.3 Vía común

Está formada por el FX, el factor V (FV), la protrombina, el fibrinógeno y el factor XIII (FXIII).

El FXa, catalizado por el complejo TF/FVIIa a partir de la vía extrínseca o por la acción del complejo tenasa en la vía intrínseca, forma el complejo protrombinasa junto con PL, Ca^{2+} y el FV activado (FVa) como cofactor (Norris 2003). En concreto, el FV puede ser activado por el FXa o por la trombina. A continuación, el complejo protrombinasa cataliza la transformación de protrombina (dependiente de vitamina K) en trombina, que desempeña un papel crucial en la formación del trombo y participa en otros procesos como la activación de las plaquetas, la retroalimentación positiva para la formación adicional de fibrina o la regulación anticoagulante de la cascada de la coagulación (Dahlbäck 2000).

La trombina proteoliza el fibrinógeno, que es una proteína soluble en plasma formada por tres pares de cadenas polipeptídicas A-alfa, B-beta y gamma unidas por enlaces disulfuro. En concreto, libera los fibrinopéptidos A y B, situados en el extremo N-terminal de las cadenas A-alfa y B-beta, respectivamente, para crear hebras insolubles de fibrina que polimerizan en la formación del trombo. Además, la trombina activa al FXIII (FXIIIa) que modifica la polimerización de la fibrina para formar una malla

entrelazada de moléculas de fibrina mediante uniones covalentes. Como resultado, el trombo adquiere mayor resistencia y elasticidad y, por lo tanto, estabilidad (Mosesson 2005; Smith 2009).

1.2.2 Modelo celular de la coagulación

El sistema de la coagulación se describe como una serie de etapas que tienen lugar en distintas superficies celulares: iniciación, amplificación y propagación. De esta manera, las células son elementos esenciales y decisivos en la coagulación para el control de la hemostasia y no únicamente superficies para la formación de los complejos procoagulantes. En particular, se considera que los procesos que corresponden a la vía intrínseca no inician la cascada de la coagulación, sino que aumentan la generación de trombina una vez iniciada la cascada a través de la vía extrínseca (Hoffman and Monroe 2001).

1.2.2.1 Iniciación

La etapa de iniciación se produce por el contacto del FVII con el TF. Cabe destacar que el TF se expresa en células situadas en el exterior del sistema vascular, por lo que se evita el inicio de la coagulación en condiciones normales. Además, también se ha detectado TF en el interior del sistema vascular y asociado principalmente a leucocitos, aunque se cree que de forma inactiva (Bach 2005; Chen and Hogg 2013). Una vez producido el daño vascular, el FVII se adhiere a las superficies dañadas que expresan TF para formar el complejo TF/FVIIa. El resultado final de esta etapa es la formación de pequeñas cantidades de trombina (Hoffman and Monroe 2001; Smith 2009).

1.2.2.2 Amplificación

Esta etapa consiste en la preparación para la producción de trombina a gran escala. En concreto, la trombina producida en la fase de iniciación activa las plaquetas y otros factores de la coagulación (FVIII, FXI y FV) (Smith 2009).

1.2.2.1 Propagación

Por último, la propagación ocurre en la superficie de las plaquetas activadas, que es donde se localizan el FVIIIa, FXIa y FVa de la fase de amplificación y el FIXa generado por el complejo TF/VIIa en la fase de iniciación. Como resultado, el FX presente en plasma se activa en la superficie de las plaquetas y se forman grandes cantidades de trombina para la creación de un trombo estable (Hoffman and Monroe 2001).

1.2.3 Regulación de la coagulación

Una vez generado el trombo existen diversos mecanismos que regulan la cascada de la coagulación limitando la continua formación de trombina y su localización exclusivamente al lugar de daño vascular.

Una vez se generan pequeñas cantidades de FXa en el inicio de la coagulación, el TFPI secretado por el endotelio forma rápidamente un complejo con el FXa y, posteriormente, también pueden unirse al complejo TF/FVIIa. De esta manera, se inhibe la vía extrínseca de la coagulación y se evita la generación descontrolada de trombina. A pesar de esto, se consigue producir pequeñas cantidades de trombina para la generación de fibrina (Bouma and Meijers 1999).

Por otra parte, la proteína plasmática antitrombina (AT) protege al sistema circulatorio de las enzimas activadas en la coagulación. Su actividad se intensifica por la unión a la heparina situada en la superficie de las células endoteliales. La AT inhibe a la trombina, FIXa, FXa, FXIa y al complejo TF-FVIIa, neutralizando de forma más eficaz a

las enzimas libres. Esto limita la localización del proceso de la coagulación a las zonas de lesión vascular (Norris 2003).

La proteína C (PC) es una proteína plasmática dependiente de vitamina K que, unida al receptor endotelial de PC (EPCR), se activa (APC) por interacción con la trombina unida previamente al receptor trombomodulina de las células endoteliales. La APC actúa en la superficie de las células endoteliales y de las plaquetas junto con el cofactor proteina S (PS) libre, otra proteína plasmática dependiente de vitamina K, para limitar la formación de nuevas moléculas de trombina mediante la conversión de FVIIIa y FVa a sus formas inactivas. La proteína S también puede estar presente en la circulación unida a la proteína de unión a C4b (C4BP) del sistema del complemento (Crawley et al. 2011).

Además, participan otros inhibidores como el inhibidor dependiente de la proteina Z (PZ) o ZPI, un anticoagulante plasmático que inhibe al FXa mediante el cofactor PZ, proteína dependiente de vitamina K (Crawley et al. 2011; Austin 2013).

1.3 Fibrinólisis

Es el mecanismo que permite la disolución del trombo estable (Figura 2).. En el lugar de daño vascular se recupera la estructura tisular normal una vez la fibrina ha realizado su función y el daño vascular se va cicatrizando.

El componente principal es la enzima plasmina, que se genera a partir de la activación del plasminógeno plasmático por el t-PA o el activador de plasminógeno tipo uroquinasa (u-PA). En esta fase, tanto el plasminógeno como el t-PA secretado por las células endoteliales pueden unirse a fibrina, y esto potencia y localiza la activación del plasminógeno. Por otra parte, el u-PA, que puede ser secretado por las células endoteliales, macrófagos, células epiteliales renales y algunas células tumorales, activa al plasminógeno de manera no dependiente de fibrina. Finalmente, el sustrato

de la plasmina es la fibrina del coágulo, que se fragmenta en productos solubles de degradación de fibrina (FDPs) como el dímero D (Sidelmann et al. 2000; Cesarman-Maus and Hajjar 2005).

Asimismo, la fibrinólisis debe producirse de manera controlada y los inhibidores de la fibrinólisis tienen un papel importante en su regulación. Éstos pueden inhibir a nivel del plasminógeno, como en el caso de PAI-1, que es el inhibidor más importante y de respuesta más rápida de la fibrinólisis. PAI-1 se secreta en las células endoteliales, macrófagos, monocitos, hepatocitos, adipocitos y plaquetas y actúa formando complejos con el t-PA y u-PA, evitando así la activación del plasminógeno. Además, los inhibidores de la fibrinólisis también pueden regular directamente la actividad de la plasmina mediante la formación de complejos con la alfa2-antiplasmina. Este inhibidor está presente en plasma y en los gránulos alfa de las plaquetas. Otro regulador es el inhibidor de la fibrinólisis activable por trombina (TAFI) presente en plasma y en las plaquetas. TAFI actúa tras la interacción de trombina con trombomodulina para reducir la unión de plasminógeno y t-PA a fibrina (Cesarman-Maus and Hajjar 2005).

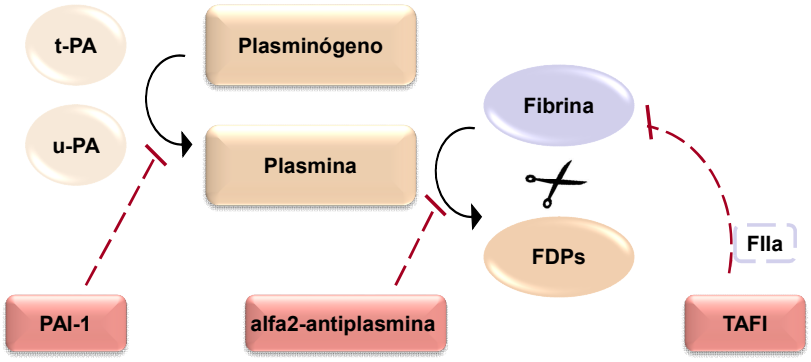

Figura 2. Sistema fibrinolítico. Adaptado (Sidelmann et al. 2000).

2. Enfermedad tromboembólica

La trombosis es la formación de un trombo obstructor en el sistema circulatorio que bloquea el flujo sanguíneo debido al desequilibrio del sistema hemostático, en el que predomina el estado protrombótico (Rosendaal 2005). La enfermedad tromboembólica es el conjunto de manifestaciones clínicas que abarcan tanto la trombosis arterial, o enfermedad tromboembólica arterial (ATE), como la trombosis venosa, o enfermedad tromboembólica venosa (VTE). La ATE y la VTE comparten el evento trombótico final, pero difieren en distintos aspectos como el lugar en el que se originan, el mecanismo activador y la composición del trombo.

2.1 Enfermedad tromboembólica arterial

Esta enfermedad se localiza en las arterias, o vasos sanguíneos por los que circula la sangre oxigenada del corazón hacia los distintos órganos. En concreto, se manifiesta en los vasos coronarios como infarto de miocardio, en la circulación cerebral causando accidente cerebrovascular isquémico, o como enfermedad arterial periférica (Ross 1999; Jackson 2011). La ATE da lugar a trombos ricos en plaquetas, conocidos como “trombos blancos”.

2.2 Enfermedad tromboembólica venosa

Se localiza en las venas, o vasos sanguíneos que llevan la sangre de vuelta al corazón. Por lo tanto, es el conjunto de alteraciones que comprenden la trombosis venosa profunda (DVT), que tiene lugar preferentemente en las venas de las piernas, y la embolia de pulmón (PE), una grave complicación que sucede al llegar al pulmón fragmentos desprendidos de un trombo que provocan el bloqueo del vaso. Si bien la DVT puede tener lugar en otras regiones como, por ejemplo, los senos venosos

cerebrales o las venas del brazo, del mesenterio o de la retina, esto sucede de manera menos frecuente. Las secuelas tras la VTE incluyen el síndrome postrombótico, que es una insuficiencia venosa crónica, y la hipertensión pulmonar tromboembólica crónica, que puede ocasionar insuficiencia cardiaca (Rosendaal 1999; Goldhaber and Bounameaux 2012). En la VTE se forman trombos ricos en fibrina y eritrocitos, con una cantidad variable de plaquetas, conocidos como “trombos rojos” (Figura 3) (Mackman 2012; Reitsma et al. 2012).

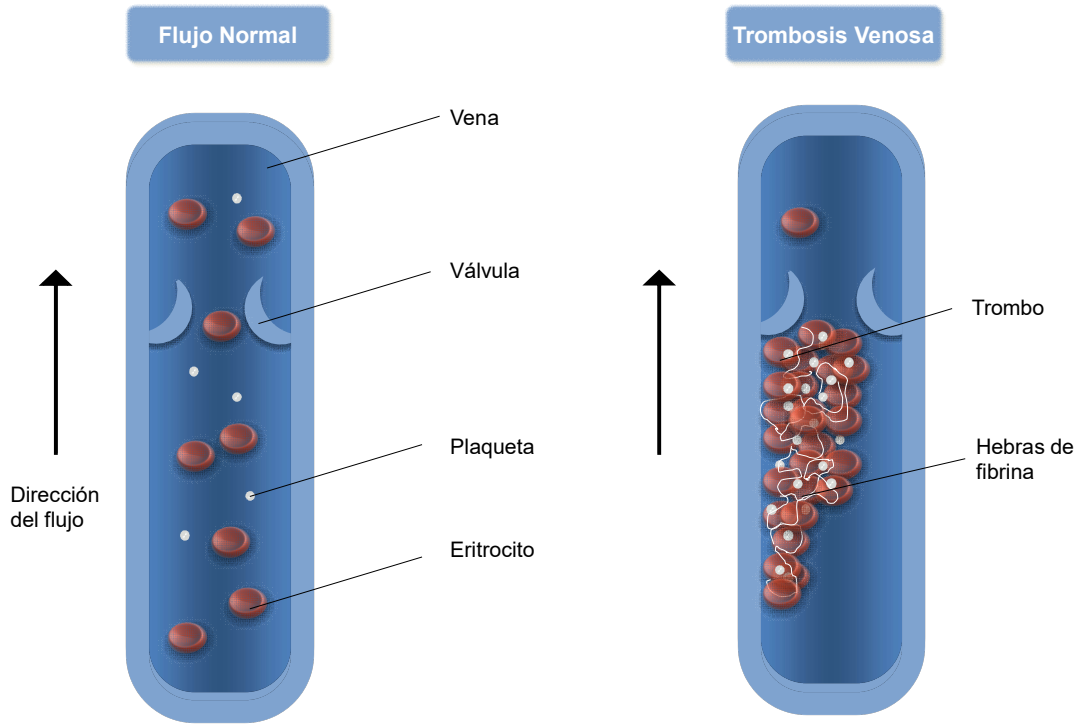

Figura 3. Comparativa entre el flujo normal en las venas y las alteraciones por la formación de un trombo obstructor patológico.

2.2.1 Evolución del concepto trombofília como enfermedad compleja

La trombofilia es la tendencia a desarrollar trombosis venosa. En la clínica se suele aplicar el término trombofilia solamente a un grupo de pacientes con una gran expresividad clínica: primera trombosis en edad joven (< 45 años), trombosis de repetición, historia familiar positiva, localización inusual de las trombosis y severidad desproporcionada con un estímulo reconocido. La trombofilia hereditaria hace referencia a la predisposición genéticamente determinada a desarrollar trombosis venosa. Ésta última se debe a la presencia de uno o varios factores genéticos que interaccionan con otros componentes que pueden ser genéticos o ambientales (Lane et al. 1996).

La primera causa de trombofilia hereditaria fue descrita por Egeberg en el año 1965 en una familia con trombofilia, la cual estaba asociada a la deficiencia de AT con un patrón de herencia autosómico dominante (Egeberg 1965). La variante genética identificada en esta familia se localiza en el gen estructural *SERPINC1* y ha sido caracterizada como AT Oslo (rs121909546; NM_000488.3: c.1306G>A). También se descubrieron otras causas de trombofilia hereditaria como la deficiencia de PC (Griffin et al. 1981) y la deficiencia de PS (Comp and Esmon 1984) en la década de los 80. Estas deficiencias no son comunes en la población general y se ha estimado una prevalencia de 1-3% en pacientes con VTE (Martinelli et al. 2014). Además, es importante destacar que aumentan entre 8-10 veces el riesgo de sufrir VTE en heterocigosis (Morange and Trégouët 2013). Por todo esto, se estableció que la trombofilia hereditaria podría tener una base genética monogénica (Lane et al. 1996; Rosendaal 1999).

En 1993 se dio a conocer la resistencia a la APC (APCR) (Dahlbäck et al. 1993) y se identificó la mutación responsable en el gen estructural del FV (*F5*) (Bertina et al. 1994), la mutación Factor V Leiden (FVL; rs6025; NM_000130.4: c.1601G>A).

También se identificó la mutación en el gen de la protrombina (*F2*) G20210A (rs1799963; NM_000506.3: c.*97G>A) (Poort et al. 1996) como factor de riesgo genético en VTE. A partir de este momento, se sugirió que la combinación de uno o más factores genéticos adicionales incrementarían el riesgo de trombosis (van Boven et al. 1996; Makris et al. 1997). Por lo tanto, se incorporó el modelo poligénico. Estas nuevas causas de trombofilia hereditaria cambiaron la concepción que se había establecido hasta el momento, puesto que se trata de alteraciones genéticas más comunes en la población general y con un impacto clínico menor. En concreto, la prevalencia de la mutación FVL es del 15-25% en pacientes con VTE y del 6% para la mutación G20210A. Por otro lado, el riesgo de VTE en portadores de la mutación de FVL en heterocigosis es 5 veces mayor y de 3-4 veces mayor en portadores de la mutación G20210A (Morange and Trégouët 2013). Teniendo en cuenta que estos trastornos presentan sintomatología variada debida a la interacción de factores de riesgo genéticos y factores de riesgo ambientales, actualmente se considera que la trombofilia hereditaria es una enfermedad multifactorial (Rosendaal 1999).

Por otra parte, también cabe destacar el cambio de concepción que ha sufrido el término enfermedad en los últimos años, cuyo estado se describe como un fenotipo dicotómico. Sin embargo, hoy en día se asume la existencia de un fenotipo continuo e inobservable que se corresponde con el riesgo o susceptibilidad (Figura 4) (Souto et al. 2000a; Souto 2002).

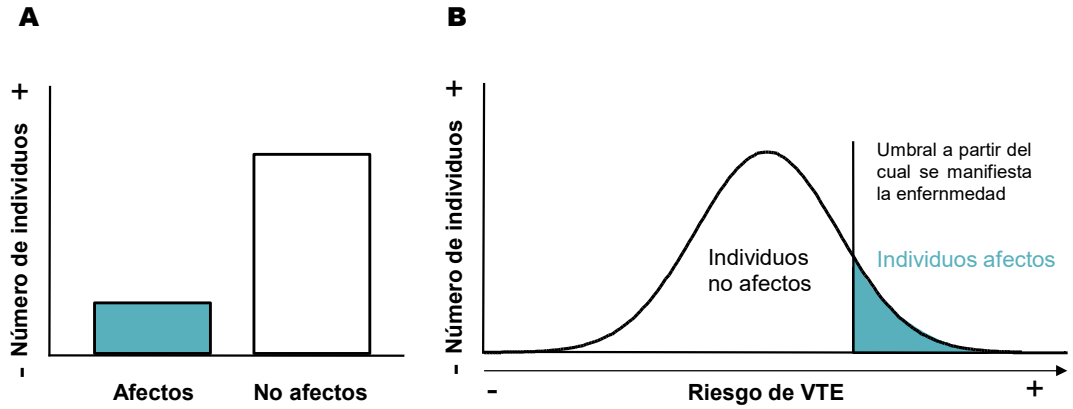

Figura 4. Distintos modelos de representación de la variable enfermedad. (A) Modelo dicotómico de la enfermedad. (B) Modelo de predisposición de distribución continua. Cada individuo tiene un grado de riesgo o susceptibilidad para padecer una enfermedad como la VTE. Si este grado supera un determinado umbral, el individuo es afecto.

De esta manera, la VTE se describe actualmente como una enfermedad compleja y multifactorial, donde no se sigue un patrón de herencia mendeliana debido a la interacción entre factores genéticos y factores ambientales (o adquiridos) que dan lugar a la susceptibilidad para padecer dicha enfermedad (Rosendaal 1999; Souto 2002). La proporción de la variabilidad de un fenotipo que se atribuye sólo a la acción de los genes se conoce como heredabilidad. Es importante destacar que la heredabilidad de la VTE se ha estimado en un 61%, lo que significa que los genes juegan el papel más importante en la variación del riesgo de padecer VTE (Souto et al. 2000a; Heit et al. 2004).

2.2.2 Mecanismos de la enfermedad tromboembólica venosa

Un trombo se forma a causa de una lesión de un vaso sanguíneo para evitar la pérdida de sangre. Cuando este proceso tiene lugar en otro contexto, nos encontramos con una trombosis patológica.

El patólogo Rudolf Virchow postuló en 1856 que los estados hipercoagulados o trombosis se producen por tres factores que constituyen la tríada de Virchow: cambios

en el flujo sanguíneo, en el vaso y/o en los componentes de la sangre (Virchow 1856). Actualmente se considera que estos mecanismos son el resultado de la interacción entre factores de riesgo genéticos y ambientales (Vilalta and Souto 2014).

Actualmente, se extiende la tríada de Virchow a nuevos mecanismos que originan un trombo patológico sin la presencia de traumatismo vascular. En los mecanismos de riesgo se incluye la lentitud del flujo sanguíneo o estasis, la hipoxemia, la activación del endotelio, la activación de la inmunidad innata (incluyendo monocitos y granulocitos) y adquirida, la activación de las plaquetas, la concentración y naturaleza de micropartículas y las variaciones cuantitativas y cualitativas de proteínas procoagulantes y anticoagulantes hacia un estado de hipercoagulabilidad (Mackman 2012; Reitsma et al. 2012).

2.2.3 Factores de riesgo genéticos

El 61% de la predisposición a VTE se atribuye a los factores genéticos, por lo que la heredabilidad de esta enfermedad es alta (Souto et al. 2000a; Heit et al. 2004). Teniendo esto en cuenta, y junto a los grandes avances producidos en los métodos de investigación de la genética humana, el estudio de los factores de riesgo genéticos ha sido de gran interés durante los últimos años.

Existen distintos mecanismos descritos en la literatura que afectan a la trombofilia, como las deficiencias en AT, PC y PS ya comentadas anteriormente, que provocan una pérdida de función anticoagulante. Otros mecanismos consisten en la ganancia de función procoagulante, siendo éste el caso de la mutación FVL y la mutación G20210A del gen de la protrombina. Cabe destacar que la mutación FVL es la alteración genética más común en los pacientes con trombosis. Debido a esta mutación, la inactivación del FV procoagulante resulta menos eficiente, lo que conlleva a un aumento del riesgo de trombosis. Por otra parte, la mutación G20210A de la

protrombina provoca un aumento de la concentración de esta proteína en plasma debido a una mayor eficacia en el procesamiento y una mayor estabilidad del ácido ribonucleico (RNA) mensajero (mRNA) (Poort et al. 1996; Gehring et al. 2001; Carter et al. 2002). Esto, a su vez, aumenta la generación de trombina y, de esta manera, la mutación G20210A se relaciona con un mayor riesgo de padecer trombosis. Además, la protrombina también está implicada en la inhibición de la actividad de la APC (Rosendaal 2005; Morange and Trégouët 2013).

El grupo sanguíneo ABO también se ha caracterizado como un factor genético de riesgo tromboembólico claramente establecido, siendo el más común en la población. En 1969 se describió que los grupos sanguíneos no-O tienen un riesgo superior de VTE que los individuos con grupo sanguíneo O (Jick et al. 1969). Teniendo en cuenta que los individuos con grupos sanguíneos no-O presentan concentraciones más altas de FVIII y vWF, éste parece ser el mecanismo mediante el cual se explica parte de su relación con la VTE (Martinelli et al. 2014). Durante los últimos años se ha demostrado que, en concreto, son los individuos con el alelo A1 los que presentan más riesgo de trombosis (Tirado et al. 2005). También cabe destacar la variante genética rs2066865 (NM_000509.5: c.*216C>T) situada en el gen *FGG*, que codifica para la cadena gamma del fibrinógeno. Esta variante genética ha sido asociada con una disminución de los niveles del tránscrito fibrinógeno gamma prima al disminuir la eficacia del procesamiento alternativo del mRNA (Uitte de Willige et al. 2005).

Aparte de estos factores de riesgo identificados antes de la aplicación de los estudios de asociación de genoma completo (GWAS) también se han registrado otros factores genéticos de riesgo trombótico durante los últimos años, algunos de los cuales se incluyen en la Tabla 1.

Tabla 1. Muestra de factores genéticos asociados con la VTE. Adaptado (Martinelli et al. 2014; Vilalta and Souto 2014; Morange et al. 2015).

Gen	Variante	RR	Fenotipo Asociado
SERPINC1	Más de 130 mutaciones privadas	10	Deficiencia AT
SERPINC1	rs121909548; AT Cambridge II	10	Deficiencia AT
PROC	Más de 160 mutaciones privadas	8	Deficiencia PC
PROS1	Más de 200 mutaciones privadas	8	Deficiencia PS
F5	rs6025; FVL	5	APCR
F5	rs118203906; FV Cambridge	5[a]	APCR
F5	rs118203905; FV Hong Kong	5[a]	APCR
F12	rs1801020	5[b]	Disminución FXII (alelo de riesgo: timina)
F2	rs1799963; Protrombina G20210A	3-4	Protrombina elevada
ABO	Grupos no-O	2-4	FVIII y vWF elevados
SERPINA10	rs2232698	3,30	Disminución ZPI
FGG	rs2066865	1,47	Disminución fibrinógeno gamma prima
F11	rs2036914	1,35	FXI elevado
F11	rs2289252	1,35	FXI elevado
SERPINC1	rs2227589	1,30	Disminución AT
TSPAN15	rs78707713	1,28	VTE (alelo de riesgo: timina)
STAB2	rs4981021	1,29	vWF elevados
TC2N	rs1884841	1,22	vWF elevados
PROCR	rs867186	1,22	EPCR elevado
F5	rs4524	1,21	VTE (alelo de riesgo: adenina)
HIVEP1	rs169713	1,20	VTE
SLC44A2	rs2288904	1,19	VTE
KNG1	rs710446	1,19	Disminución aPTT, FXI elevado
GP6	rs1613662	1,15	Activación plaquetaria elevada
VWF	rs1063856	1,15	vWF elevado (alelo de riesgo: guanina)
F13A1	rs5985	0,85	VTE (alelo de riesgo: guanina)
STXBP5	rs1039084	0,78	vWF elevados (alelo de riesgo: adenina)
THBD	rs1042579	0,72	APC elevada

FVL: Factor V Leiden; FV: factor V; RR: riesgo relativo; AT: antitrombina; PC: proteína C; PS: proteína S; APCR: resistencia a la proteína C activada; FXII: factor XII; FVIII: factor VIII; vWF: factor de von Willebrand; ZPI: inhibidor de la proteína Z; FXI: factor XI; EPCR: receptor endotelial soluble de la proteína C; aPTT: tiempo de tromboplastina parcial activada; APC: proteína C activada.[a] Se asume el mismo riesgo que FVL (Soria et al. 2014). [b] Riesgo asociado a homocigotos c.-4T.

2.2.4 Factores de riesgo ambientales

En pacientes con trombofília no se observa una afectación clínica contínua y esto sugiere la necesidad de factores desencadenantes. Por lo tanto, los factores ambientales y la interacción de los factores genéticos con éstos adquiere especial interés (Vilalta and Souto 2014). Los factores de riesgo ambientales descritos son diversos, entre los que se incluyen los reportados en la Tabla 2.

Tabla 2. Resumen de los factores de riesgo ambientales de la VTE. Adaptado (Anderson and Spencer 2003; Vilalta and Souto 2014).

Factor de riesgo	OR
Escayolamiento	> 10
Cirugía ortopédica	> 10
Cirugía general o traumatismo mayor	> 10
Lesión de la espina dorsal	> 10
Catéter venoso central	2 - 9
Quimioterapia/cáncer	2 - 9
Períodos prolongados de inmovilización	2 - 9
Insuficiencia cardíaca o respiratoria	2 - 9
Terapia de reemplazo hormonal	2 - 9
Anticonceptivos de uso oral	2 - 9
Parálisis post ictus	2 - 9
Embarazo y puerperio	2 - 9
Tromboembolismo venoso previo	2 - 9
Anticuerpos antifosfolípidos	2 - 9
Reposo en cama más de 3 días	< 2
Edad	< 2
Cirugía laparoscópica	< 2
Varices	< 2
Obesidad	< 2
Tabaquismo	< 2
Sexo masculino [a]	< 2
Inmovilización del viajero	< 2

OR: odds ratio. [a] Sólo aumenta el riesgo a recidiva.

2.2.5 Epidemiología

La VTE es una patología frecuente y grave con una incidencia de 1-3 por cada 1000 individuos al año, dos tercios de los cuales, aproximadamente, desarrollan DVT y un tercio sufren PE. De los pacientes que han sufrido un evento trombótico, el riesgo de recurrencia al acabar el tratamiento con anticoagulantes es del 6% a los 6 meses y se estima que el 40% de los casos de DVT presentará un síndrome postrombótico, mientras que la hipertensión pulmonar tromboembólica crónica ocurrirá en un 2-4% de los pacientes como complicación de la PE. A los 10 años, la recurrencia ocurre en un tercio de los pacientes, siendo más pronunciada en hombres que en mujeres (Rosendaal 1999; White 2003; Cushman et al. 2004; Colman et al. 2006; Cohen et al. 2007; Goldhaber and Bounameaux 2012). La VTE es actualmente la tercera causa de muerte cardiovascular, después del infarto de miocardio y el ictus, siendo la enfermedad cardiovascular la primera causa de muerte a nivel mundial. Durante el primer mes de diagnóstico la mortalidad es del 6% en los casos de DVT y del 12% en los casos de PE (White 2003; Goldhaber and Bounameaux 2012; Townsend et al. 2016). En concreto, en España se ha estimado una incidencia anual de 1,24 casos por cada 1000, con un coste sanitario de 60 millones de euros al año (Grupo Multidisciplinar para el Estudio de la Enfermedad Tromboembólica en España 2006). Además, la VTE sería la responsable del 12% de la mortalidad general anual (Cohen et al. 2007). Por lo tanto, teniendo en cuenta que es un problema de salud pública con un elevado coste sanitario, existe mucho interés en la mejora del diagnóstico, prevención y tratamiento.

2.2.6 Diagnóstico

La DVT puede presentar síntomas como dolor o sensibilidad, sensación de calor y edema en la extremidad afectada. La PE se caracteriza por disnea o sensación de

falta de aire, dolor de pecho, taquipnea o aumento de la frecuencia respiratoria y tos seca. Otros síntomas menos frecuentes incluyen fiebre, hemoptisis, cianosis, hipotensión o estado de choque (Wilbur and Shian 2012).

En la actualidad, el diagnóstico de VTE abarca el diagnóstico clínico, las pruebas analíticas de laboratorio como el dímero D y los estudios de imagen. En concreto, los criterios actuales recomiendan que en pacientes con bajo riesgo de VTE y dímero D negativo el diagnóstico sea de exclusión de VTE. Por otra parte, el diagnóstico se confirma mediante ecografías venosas en pacientes con riesgo medio o elevado de DVT y mediante la angiografía por tomografía computarizada en pacientes con riesgo de PE (Goldhaber and Bounameaux 2012; Wilbur and Shian 2012).

2.2.7 Prevención

La VTE es una enfermedad común que se puede prevenir con el uso adecuado de anticoagulantes profilácticos que inhiben la coagulación de la sangre. Esta profilaxis puede llevarse a cabo con anticoagulantes que inhiben la acción de las enzimas de la coagulación, como la heparina de bajo peso molecular o dosis bajas de heparina no fraccionada, anticoagulantes orales que inhiben la síntesis de los factores de la coagulación, como la warfarina, que es una antagonista de la vitamina K (VKA), y métodos físicos como la compresión intermitente (Anderson and Spencer 2003). Como alternativas a los VKA, teniendo en cuenta que presentan múltiples interacciones con otros fármacos, se han desarrollado otros anticoagulantes orales como inhibidores directos del FXa o un inhibidor directo de la trombina (Mateo 2013).

Es importante remarcar que el tratamiento profiláctico se debe administrar sólo a los pacientes que lo necesiten, puesto que existen posibles complicaciones como el riesgo de sangrado. Individuos de alto riesgo que se encuentren en situaciones protrombóticas es el caso de algunos pacientes hospitalizados, enfermos de cáncer o

embarazadas. Además, se debería considerar si son individuos de riesgo que pueden beneficiarse del diagnóstico genético de trombofilia tanto aquellos pacientes con una historia clínica de VTE, como individuos asintomáticos con familiares afectos de VTE, aunque no existe un consenso al respecto (Anderson and Spencer 2003; Morange and Trégouët 2013; Vilalta and Souto 2014).

Con el fin de tomar decisiones preventivas y terapéuticas, la modelización de toda la información genética y clínica de la que disponemos sobre la VTE en perfiles de riesgo individuales mejoraría la determinación precoz y permitiría estimar de forma cuantitativa y objetiva el riesgo de padecer VTE (Vilalta and Souto 2014). De esta manera se tiene en cuenta que la VTE es una enfermedad compleja, siendo la suma de factores genéticos y ambientales, así como las interacciones entre ellos, lo que proporciona la información necesaria para estratificar a los pacientes según el riesgo individual que presenten. Hasta la fecha, se han desarrollado distintos *scores* de riesgo o índices predictivos que integran la información genética para tratar de mejorar la capacidad de predicción de la enfermedad (de Haan et al. 2012; Soria et al. 2014), aunque sólo se ha logrado explicar hasta el 15% de la variación del riesgo de VTE a partir de la herramienta Thrombo inCode (Ferrer in Code, Barcelona, España), que permite el genotipado simultáneo de 12 variantes genéticas asociadas al riesgo de VTE: rs6025 (*F5*, Factor V Leiden), rs118203906 (*F5*, Factor V Cambridge), rs118203905 (*F5*, Factor V Hong Kong), rs1799963 (*F2*, G20210A), rs5985 (*F13A1*), rs121909548 (*SERPINC1*, AT Cambridge II), rs2232698 (*SERPINA10*), rs1801020 (*F12*), rs8176719 (*ABO*), rs7853989 (*ABO*), rs8176743 (*ABO*) y rs8176750 (*ABO*). (Soria et al. 2014). De hecho, los factores de riesgo genéticos conocidos permiten solamente identificar la causa genética principal a menos del 30% de los pacientes con VTE con ausencia de factores ambientales desencadenantes (Morange and Tregouet 2010). Debido a estas evidencias, el gran reto en la actualidad consiste en identificar

nuevos factores de riesgo trombótico y mejorar la integración de la información mediante algoritmos matemáticos.

2.2.8 Tratamiento

El tratamiento de la VTE consiste en el uso de fármacos anticoagulantes, con el fin de evitar la extensión del trombo y evitar otras complicaciones. Principalmente, se procede a la administración inicial de heparina, con un efecto anticoagulante inmediato, siendo la heparina de bajo peso molecular preferente frente a la heparina no fraccionada, teniendo en cuenta ventajas farmacocinéticas y coste-efectivas. El tratamiento con heparina se interrumpe al cabo de unos 5-7 días, después de solaparse con el inicio de los anticoagulantes orales como la warfarina, cuyos efectos no aparecen inmediatamente después de su administración. La duración de la segunda fase del tratamiento depende del riesgo de recurrencia de cada paciente, siendo habitualmente un mínimo de 3 a 6 meses (Paramo et al. 2007).

Respecto a los nuevos anticoagulantes orales, siguen presentando riesgo de sangrado, aunque éste podría ser menor que el riesgo de los fármacos VKA. Se están desarrollando nuevos estudios en esta dirección para la inhibición de factores de la coagulación que están implicados en el desarrollo patológico del trombo pero que no parecen ser tan esenciales para la generación de trombina, y por lo tanto su inhibición está menos asociada al sangrado. Éste es el caso del FXI y el FXII que se postulan como futuras dianas de terapias antitrombóticas para evitar, además, el sangrado (Müller et al. 2011).

3. Estudio genético de las enfermedades complejas

Descubrir la base genética de las enfermedades complejas es el primer paso para mejorar su diagnóstico, prevención y tratamiento (Manolio et al. 2009). Las enfermedades complejas suelen ser genéticamente muy heterogéneas y las causas pueden diferir entre individuos. La heterogeneidad genética puede deberse a la existencia de diversas variantes de riesgo en un mismo gen o por la implicación de diferentes genes en la susceptibilidad para padecer una misma enfermedad. Incluso, individuos con una misma variante genética pueden presentar diferencias fenotípicas debido a fenómenos de penetrancia incompleta, epigenética, o por la interacción con otros factores de riesgo genéticos o ambientales (Wang et al. 2015).

Para el análisis genético de las enfermedades complejas es de gran utilidad examinar algunos conceptos básicos como la heredabilidad del fenotipo de estudio y la correlación de fenotipos intermediarios con el resigo de sufrir la enfermedad de interés. Además, conocer los posibles diseños de la muestra y los fundamentos de diiferentes técnicas y métodos desarrollados permite valorar las ventajas e inconvenientes que nos ofrecen en función de los objetivos de estudio.

3.1 Heredabilidad

Es el parámetro que compara la importancia relativa de las influencias genéticas y ambientales en la variabilidad de un fenotipo. En otros términos, la variancia total de un fenotipo observado en una población, y en un momento determinado, se puede expresar como la suma de las variancias subyacentes no observadas de los componentes genéticos y ambientales. De esta manera, la heredabilidad, en sentido amplio, se define como un ratio de variancias, siendo la proporción de la variancia

fenotípica atribuible al efecto de los genes. La variancia genética, igualmente, se descompone en variancia aditiva (comprende los efectos aditivos de los genes sobre el fenotipo), variancia de dominancia (interacción entre alelos en un mismo sitio cromosómico) y variancia epistática (interacción entre alelos de distintos sitios cromosómicos). En sentido estricto, la heredabilidad (h^2) mide la proporción de la variancia fenotípica total que está determinada exclusivamente por la varianza genética aditiva. Concretamente, éste es el parámetro más utilizado y la causa principal de las similitudes entre parientes (Visscher et al. 2008).

La heredabilidad se estima en estudios familiares, teniendo en cuenta que el efecto genético aditivo influye en la variabilidad observada del fenotipo si entre familiares más próximos los valores del fenotipo tienden a ser más parecidos (Almasy and Blangero 2010).

3.2 Fenotipos intermediarios

Los parámetros de laboratorio que se asocian con el riesgo de padecer una enfermedad se consideran fenotipos intermediarios. Por ejemplo, se consideran fenotipos intermediarios la APCR y el FVIII, los cuales se han asociado previamente al riesgo de VTE (Vilalta and Souto 2014). Su uso ha sido de gran utilidad en el estudio de las enfermedades complejas para la identificación de determinantes genéticos aún desconocidos, teniendo en cuenta su potencial implicación en la fisiopatología de la enfermedad de estudio. Los fenotipos intermediarios también son complejos y, por lo tanto, son el resultado de la interacción entre factores genéticos y ambientales. Éstos presentan una serie de ventajas frente al uso de la propia enfermedad como fenotipo de estudio. Principalmente, están influidos por un menor número de factores de riesgo, por lo que en su análisis se evita la atenuación de las señales genéticas. Además, los fenotipos intermediarios se pueden medir de forma precisa en una escala cuantitativa

continua, tanto en individuos sintomáticos como asintomáticos. Esto ofrece, por lo tanto, más información sobre la predisposición a la enfermedad de estudio (Souto 2002; Blangero et al. 2003).

Con el fin de priorizar entre posibles fenotipos intermediarios, así como justificar su uso, se puede estimar la heredabilidad de éstos de forma previa a los análisis genéticos (Souto et al. 2000b). Además, se puede analizar la naturaleza de relación entre los fenotipos intermediarios y la enfermedad a partir del análisis bivariado de la partición de la covariancia fenotípica en el componente genético y el componente ambiental (Almasy and Blangero 2010). Una correlación genética estadísticamente significativa sugiere la existencia de determinantes genéticos implicados simultáneamente en la variabilidad de los fenotipos intermediarios y del fenotipo enfermedad, lo que se conoce como pleiotropía. Por lo tanto, esto indica que el estudio de los fenotipos intermediarios puede aportar información adicional sobre la base genética de la enfermedad (Souto et al. 2000a).

3.3 Tipo de muestra

La selección de la muestra es un factor crucial a tener en cuenta en la investigación genética, puesto que tiene una implicación directa en los resultados obtenidos. Los dos tipos de diseño utilizados son los basados en individuos no emparentados y en individuos emparentados.

3.3.1 Individuos no emparentados

El diseño más simple y más utilizado en la investigación de la base genética de las enfermedades complejas es el modelo caso-control. Esta estrategia se basa en el estudio de individuos afectos (casos) y no afectos (controles) que no están relacionados genéticamente entre sí. En este caso, obtener una muestra de un gran

número de individuos es asequible. Esto se debe a que el reclutamiento es relativamente sencillo, puesto que se evita la necesidad de incluir en el estudio individuos emparentados. Sin embargo, las frecuencias alélicas de los marcadores genéticos pueden diferir entre casos y controles por la diversidad en la distribución genotípica en la población, y sin estar en relación con el fenotipo de estudio. Esta estratificación poblacional puede conllevar la presencia de falsos positivos, por lo que se requiere de métodos estadísticos de control y corrección (Hirschhorn 2005; Evangelou et al. 2006).

3.3.2 Individuos emparentados

Los estudios basados en individuos emparentados pueden tener distintos diseños como los tríos formados por padres e hijos, pares de hermanos (como gemelos o mellizos), familiares seleccionados de las colas de distribución normal del fenotipo de estudio, familias extensas con múltiples individuos afectos o familias de poblaciones aisladas (Burmeister 1999; Ott et al. 2011). Por lo tanto, uno de los inconvenientes de este modelo es la dificultad a la hora de reclutar la muestra.

Los estudios familiares aportan ventajas frente a los estudios basados en individuos no relacionados en la detección de variantes con una frecuencia del alelo minoritario (MAF) muy baja (variantes raras o de muy baja frecuencia alélica) o baja (variantes de baja frecuencia alélica) así como para determinar su asociación con un fenotipo. Esto se debe a que las variantes que influyen en la variabilidad del fenotipo de estudio estarán presentes con una mayor frecuencia entre familiares afectos (Manolio et al. 2009). Además, los estudios familiares presentan ventajas como la posibilidad de realizar un estudio de los efectos de origen parental, que es esencial para comprender los procesos de epigenética, el estudio combinado de análisis de asociación y de ligamiento o la estimación de la heredabilidad. También ofrece una mayor protección

frente a la estratificación poblacional, teniendo en cuenta que entre los individuos de una misma familia se comparte un mismo origen genético (Blangero et al. 2003; Evangelou et al. 2006; Ott et al. 2011). Sin embargo, las familias pueden compartir otros factores ambientales que sean claves en el desarrollo de la enfermedad, por lo que el diseño de los estudios con familias debe tener en cuenta un adecuado control de estos efectos (Almasy and Blangero 2010).

3.4 Técnicas de genotipado

Con el objetivo de identificar genotipos implicados en la variabilidad del fenotipo de interés, las técnicas de genotipado del ácido desoxirribonucleico (DNA) se basan a menudo en el uso de marcadores genéticos, los cuales pueden informar sobre la variabilidad alélica de otro sitio cromosómico. En comparación, la técnica de secuenciación permite obtener directamente toda la variabilidad genética del individuo analizado (Schlötterer 2004).

3.4.1 Marcadores genéticos

Los microsatélites y los polimorfismos de un solo nucleótido (SNPs) han destacado como los más importantes durante los últimos 20 años, aunque existen otros tipos de marcadores genéticos (Schlötterer 2004).

3.4.1.1 Microsatélites

Estos marcadores multialélicos consisten en repeticiones consecutivas de 2 a 4 nucleótidos de la secuencia del DNA. De esta manera, se facilita que los individuos sean heterocigotos y resultan muy informativos para distinguir el alelo paterno del materno. Los microsatélites son abundantes y están distribuidos a lo largo del genoma. En concreto, su genotipado se obtiene a partir de la amplificación por reacción en

cadena de la polimerasa (PCR) y la separación y visualización de los fragmentos resultantes mediante secuenciadores automáticos de electroforesis capilar y detección de fluorescencia, entre otras técnicas (Burmeister 1999; Gray et al. 2000).

3.4.1.2 SNPs

Estas variaciones afectan a una sola base nitrogenada de una secuencia del genoma y se encuentran en más del 1% de la población. Sin embargo, a menudo se incluyen en esta definición inserciones, delecciones y otras variantes de distinta frecuencia alélica. Los SNPs son marcadores típicamente bialélicos y son el tipo de variante genética más abundante del genoma. En general, el genotipado de SNPs parte de la amplificación por PCR. La detección específica de los alelos se basa en técnicas de hibridación, ligación, escisión enzimática o extensión de cebadores y la identificación se basa en técnicas de espectrometría, fluorescencia o quimioluminiscencia. Cabe destacar los recientes avances en las técnicas de genotipado a gran escala mediante biochips, siendo Affymetrix (Santa Clara, CA, USA) e Illumina (San Diego, CA, USA) las dos principales plataformas. La tecnología GeneChip de Affymetrix se basa en la técnica de hibridación y detección por fluorescencia, mientras que la tecnología BeadArray de Illumina combina hibridación, extensión de cebadores y ligación, seguido de la detección del genotipo mediante fluorescencia (Kim and Misra 2007; Perkel 2008).

3.4.2 Secuenciación

La secuenciación del DNA es la técnica que permite la lectura del orden de los nucleótidos. Esto posibilita la identificación de variantes genéticas de todas las frecuencias alélicas, tanto si se han descrito previamente como si se trata de variantes genéticas nuevas.

La secuenciación tradicional se basa en la técnica de secuenciación de ácidos nucleicos diseñada por Sanger (Sanger et al. 1977) y por la cual recibió el Premio Nobel de Química en 1980. Durante los últimos años se han desarrollado las tecnologías de secuenciación masiva, también conocidas como secuenciación de nueva generación (NGS) que son óptimas para la secuenciación a gran escala de los genomas humanos (Metzker 2010).

3.4.2.1 Secuenciación tradicional

En general, se parte de millones de copias purificadas del DNA molde que contiene la secuencia de interés. La secuenciación consiste en ciclos de desnaturalización, hibridación y extensión, en los que se utiliza un cebador de secuencia conocida que marca el inicio de la secuencia de interés. En estos ciclos, la incorporación de didesoxinucleótidos (ddNTPs) marcados con fluorescencia provoca la terminación aleatoria de la extensión, por lo que el resultado final son fragmentos de distinto tamaño en los que la identidad del último nucleótido está marcada de forma diferencial. A continuación, los fragmentos de cadena simple se ordenan según su longitud mediante electroforesis. A medida que dichos fragmentos atraviesan el capilar, las moléculas fluorescentes se excitan mediante un láser, lo que permite la lectura ordenada de la identidad del nucleótido en cada posición de la secuencia. Las señales obtenidas se traducen en cromatogramas que facilitan la interpretación de la secuencia (Shendure and Ji 2008).

3.4.4.2 Secuenciación masiva

La NGS ha supuesto mejoras en el coste, velocidad y capacidad respecto a la secuenciación tradicional (van Nimwegen et al. 2016). En 2006 la fundación X-Prize (Santa Monica, CA, USA) ofreció 10 millones al primer equipo de investigadores que lograra secuenciar 100 genomas humanos en 10 días por menos de 10.000 dólares

(USD) por genoma. Sin embargo, la fundación canceló la competición al considerar que no incentivaba el avance natural que estaba experimentando esta tecnología. La NGS presenta una tendencia de disminución de los costes de secuenciación (Figura 5) que ha llegado a los 0,01 USD por megabase y a menos de 1.000 USD por genoma humano secuenciado con una cobertura de profundidad de 30x para el sistema Hiseq X Ten de Illumina (Wang et al. 2015).

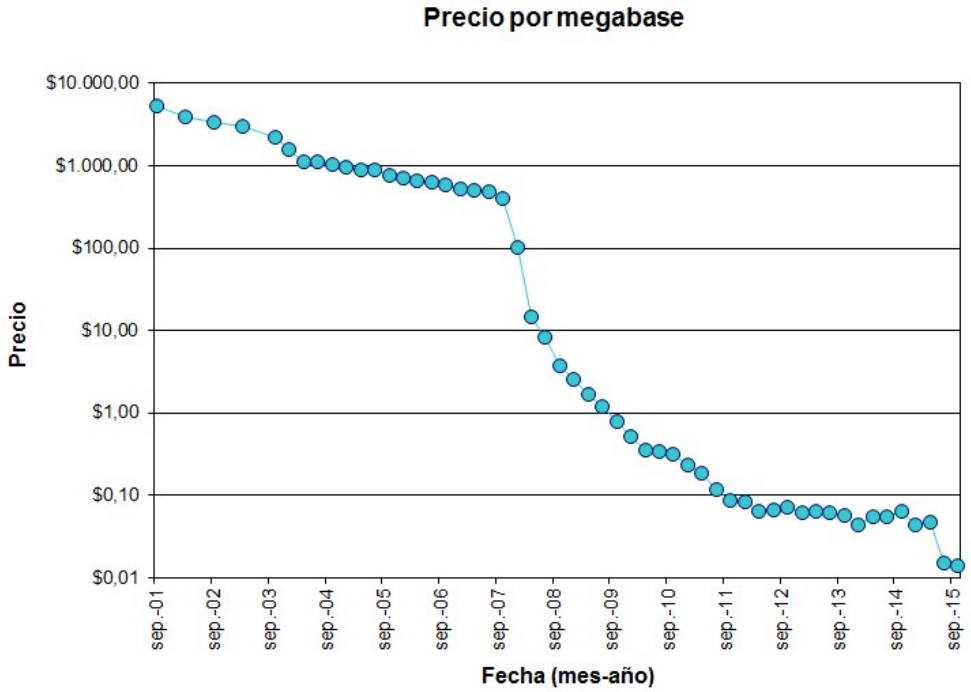

Figura 5. Gráfico de la evolución del precio de secuenciación de DNA por megabase. La caída drástica de los precios a partir de septiembre de 2007 representa la transición de la secuenciación tradicional a la NGS. Adaptado (NHGRI; www.genome.gov).

En general, se parte de la fragmentación al azar del DNA molde para crear las librerías, que son moléculas de DNA de interés junto con adaptadores de secuencia conocida que permiten la unión posterior a cebadores universal. La unión adicional de cada fragmento de la librería a un código de barras identificativo permite la secuenciación simultánea de muestras de distintos pacientes. Tras la inmovilización de cada uno de

los fragmentos separados entre sí, se pueden generan millones de reacciones de secuenciación de forma paralela y múltiples lecturas de cada base del genoma (Metzker 2010; Glenn 2011). Ésta es la principal diferencia respecto a la secuenciación tradicional. Además, puesto que las lecturas se obtienen a partir de cada uno de los fragmentos, la NGS permite el análisis de cada alelo por separado (Bentley et al. 2009). Una vez se obtiene las lecturas de secuencia, los resultados pueden alinearse contra un genoma de referencia para una identificación rápida de las variantes genéticas que difieren o bien se puede utilizar una alineación *de novo*. La gran producción de datos que conlleva la NGS ha propiciado el desarrollo de la tecnología bioinformática en cuanto a sistemas de almacenaje de datos, control de calidad, métodos de alineamiento o análisis y sistemas de gestión (Shendure and Ji 2008; Metzker 2010).

A lo largo de los últimos años se han desarrollado distintas plataformas basadas en combinaciones de estrategias para la preparación de las librerías, secuenciación y captura de imagen y análisis de datos (Metzker 2010). Sin embargo, las plataformas actualmente disponibles en el mercado son Illumina, ThermoFisher Ion Torrent, Pacific Biosciences y Oxford Nanopore (Levy and Myers 2016). Las características particulares de cada plataforma en función de la estrategia utilizada aporta distintas ventajas e inconvenientes que afectan a la capacidad de secuenciación, coste, ratio de error y longitud de lectura, entre otros (Tabla 3).

Tabla 3. Características de las principales plataformas de NGS y la tecnología de secuenciación tradicional. Adaptado (Park et al. 2015; Goodwin et al. 2016).

Plataforma	Amplificación	Secuenciación	Máxima longitud de lectura (bp)	Capacidad (Gb)	Tiempo por *run* (h)	Coste/Gb (USD)	Error dominante	Ratio de error
Illumina Miseq	Fase Sólida	Terminadores reversibles	300-600	13,2-15	21-56	109,2-996,0	Substituciones	0,2
Illumina Hiseq X Ten	Fase Sólida	Terminadores reversibles	150	800-900	72	7,1	Substituciones	0,2
Ion PGM	EmPCR	Semiconducción	200-400	1-2	2,3-7,3	450,0-800	Indel	1,0
Ion Proton	EmPCR	Semiconducción	200	10	2-4	11,4-81,6	Indel	1,0
PacBio RS II	-	Molécula Única	20.000	1	4	1.000	Indel	1-13,0
Oxford Nanopore MinION	-	Molécula Única	200.000	1.5	48	750	Indel	48

EmPCR: amplificación por reacción en cadena de la polimerasa en emulsión; bp: par de bases; h: hora; Gb: gigabase; USD: dólar americano.

La tecnología de Illumina/Solexa (Figura 6) se originó en el año 2006 (Fedurco et al. 2006; Turcatti et al. 2008). Esta plataforma realiza una amplificación de la librería en fase sólida, en la que cebadores en sentido directo o *forward* y en sentido inverso o *reverse* se encuentran anclados en una superficie sólida. Cada molécula de la librería hibrida con uno de los cebadores y se produce una amplificación en forma de puente con los cebadores que se encuentran en las zonas próximas. Como resultado, se crea de forma aislada para cada molécula un grupo de copias clonales o *cluster*. A continuación, la estrategia empleada en la secuenciación y captura de imagen es la utilización de terminadores reversibles. Para esto, un cebador se hibrida a la región de secuencia universal conocida con el fin de realizar pasos cíclicos de síntesis de la cadena complementaria. En esta secuenciación por síntesis los cuatro nucleótidos están marcados con fluorescencia diferencial y el grupo hidroxilo en la posición carbónica 3' está bloqueado. Por lo tanto, en cada paso sólo se permite la incorporación de un nucleótido y se procede a la captura de la imagen. A continuación, se elimina el bloqueo del grupo hidroxilo y el fluorocromo del nucleótido añadido para poder realizar el siguiente ciclo de síntesis (Shendure and Ji 2008).

La plataforma ThermoFisher Scientific/Life Technologies/Ion Torrent (Figura 7) se presentó en el año 2010 con una nueva tecnología de secuenciación (Rothberg et al. 2011). Esta plataforma parte de la amplificación por PCR en emulsión (emPCR) de la librería, en la cual la cadena simple de cada molécula de la librería se captura por separado en microesferas que se encuentran en emulsión y contienen todos los componentes para la reacción de PCR. A continuación, cada microesferas se deposita de manera separada en los chips de semiconducción, en los que se procede a la lectura de la secuenciación en base a la detección de cambios de pH por liberación de protones al incorporarse un nucleótido (Goodwin et al. 2016).

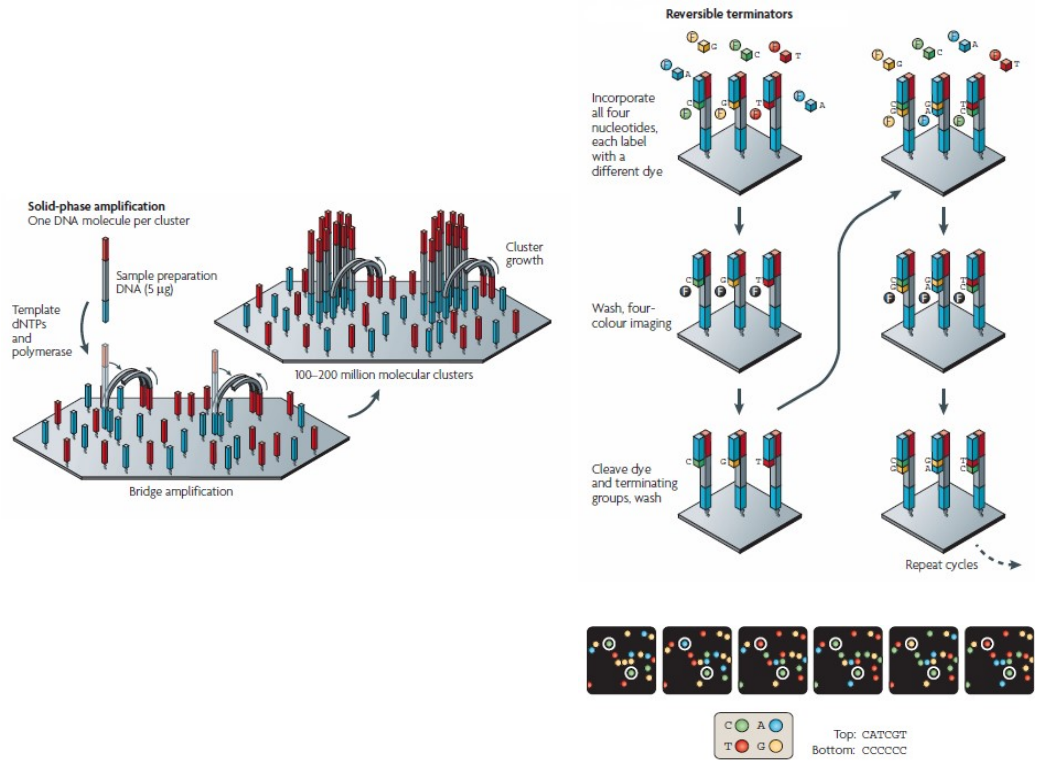

Figura 6. Estrategia de amplificación en fase sólida y secuenciación por terminadores reversibles. Adaptado (Metzker 2010).

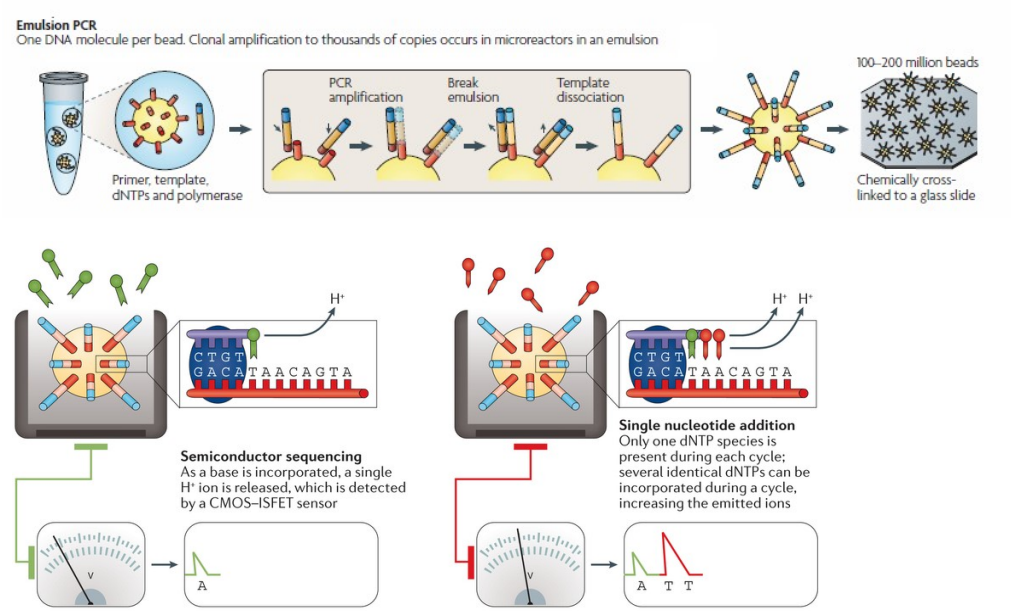

Figura 7. Estrategia de amplificación en emPCR y secuenciación por semiconducción. Adaptado (Metzker 2010; Goodwin et al. 2016).

Cabe destacar que los sistemas de captura de imagen de estas dos plataformas requieren de una amplificación clonal previa a la secuenciación de cada una de las moléculas de la librería. Por lo tanto, la señal observada para cada molécula de la librería es el consenso de las señales emitidas por cada uno de los clones del grupo de amplificación (Metzker 2010). Otras plataformas, en cambio, no necesitan una amplificación clonal de la librería sino que se basan en la secuenciación de una sola molécula, como es el caso de Pacific Biosciences (Menlo Park, CA, USA) que salió al mercado en el año 2010 (Eid et al. 2009) y Oxford Nanopore (Oxford, UK), cuya tecnología estuvo disponible en el 2014 (Jain et al. 2016). Ambas plataformas realizan una secuenciación en tiempo real, que consiste en capturas de imagen de la síntesis continua de la cadena complementaria de la molécula de interés (Figura 8).

Pacific Biosciences se basa en la actividad de la DNA polimerasa, que se encuentra inmovilizada en cada celda del chip. Tras la extensión de un nucleótido marcado con fluorescencia diferencial se captura la señal emitida. Oxford Nanopore se basa en el cambio de voltaje tras el paso de la secuencia por un poro (Goodwin et al. 2016).

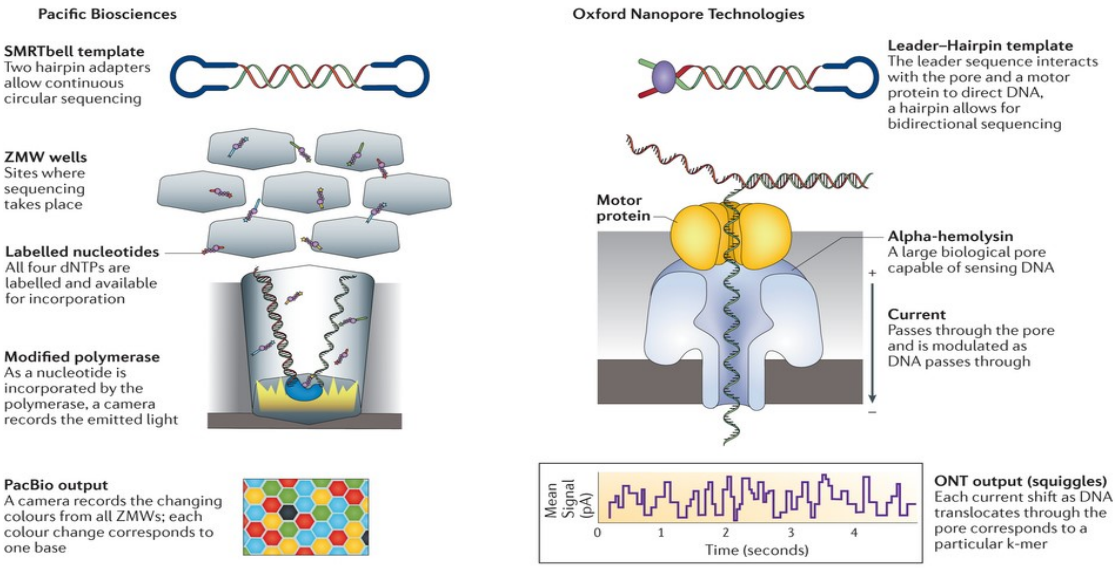

Figura 8. Estrategia de secuenciación a tiempo real de Pacific Biosciences y Oxford Nanopore. Adaptado (Goodwin et al. 2016).

3.5 Métodos de análisis genético de fenotipos complejos

Con el objetivo de identificar los genes y las variantes genéticas implicadas en las enfermedades complejas y en la variabilidad de fenotipos cuantitativos se han aplicado diversos métodos en el campo de la genética humana. En general, estas estrategias engloban los estudios de genes candidatos, los estudios del exoma y los estudios del genoma completo.

Los avances actuales en genómica se deben, entre otras cosas, al desarrollo bioinformático y a proyectos como Genoma Humano (Lander et al. 2001), HapMap (International HapMap Consortium 2005) o 1.000 Genomas (Abecasis et al. 2012) que han aportado un catálogo detallado de las variantes genéticas humanas. Además, han establecido los patrones de desequilibrio de ligamiento (LD), que es la relación entre variantes genéticas situadas a distancias pequeñas de decenas de miles de bases tal que los alelos se heredan de forma conjunta y no aleatoria en la población (Hirschhorn 2005).

3.5.1 Estudios de asociación de genes candidatos

Estos estudios consisten en el análisis de asociación de variantes genéticas o marcadores localizados en el gen candidato de interés. Por lo tanto, requieren del conocimiento previo de genes candidatos susceptibles de estar implicados en la variabilidad del fenotipo de estudio. En concreto, los análisis de asociación comparan si la distribución de la frecuencia alélica de una variante genética difiere estadísticamente y de forma significativa entre la presencia o ausencia de un fenotipo dicotómico, o bien entre los distintos valores de un fenotipo cuantitativo. Esto se puede llevar a cabo tanto en individuos no relacionados como en familias (Hirschhorn 2005).

Éste ha sido el primer tipo de estudio de asociación utilizado, antes de la aparición de los densos catálogos de variantes genéticas con patrones de LD. Aunque el estudio de

genes candidatos ha permitido la identificación de diversos genes y variantes genéticas que predisponen al riesgo de sufrir una enfermedad compleja, en ocasiones se ha fundamentado en bases biológicas incorrectas. De hecho, los que presentan más probabilidad de éxito se basan en resultados previos de análisis de asociación que marcan el gen o la región en la que se encuentra (Zeggini and Morris 2010).

3.5.2 Estudios de ligamiento genético

Los análisis de ligamiento consisten en identificar a lo largo del genoma y en familias la cosegregación o transmisión de padres a hijos de un alelo en un marcador genético que se hereda conjuntamente con el fenotipo de estudio. Dicha cosegregación se debe al ligamiento por proximidad cromosómica del marcador genético con el alelo causal. En contra, si el marcador se encuentra muy lejos del alelo causal podría separarse de éste durante la recombinación genética. Por lo tanto, se basa en alelos que se heredan conjuntamente durante algunas generaciones dentro de la familia de estudio mediante el uso de marcadores genéticos situados a lo largo de todo el genoma y separados por millones de bases (Burmeister 1999; Ott et al. 2011).

Para saber si el marcador y el alelo causal están ligados se utiliza generalmente la escala de puntuación LOD o logaritmo en base 10 de la *odds ratio* entre la probabilidad de que estén ligados y de que no haya ligamiento. Por lo tanto, una mayor puntuación LOD indica una mayor certeza estadística de ligamiento. A partir de una puntuación LOD=3 se acepta la evidencia, que corresponde a una relación de 1.000:1 a favor del ligamiento (Burmeister 1999). Es importante destacar que los análisis de ligamiento marcan una región cromosómica implicada en un fenotipo de interés sin concretar cuál es el alelo funcional. Esto se debe a que el ligamiento se observa por familiares que comparten el mismo alelo y que presentan una similitud

fenotípica mayor en comparación con los otros familiares, contemplando la coexistencia de diferentes variantes causales (Almasy and Blangero 2010).

Los métodos paramétricos comúnmente utilizados en el estudio de enfermedades monogénicas que se heredan siguiendo los clásicos patrones mendelianos necesitan conocer o presuponer el modelo de herencia para estimar la frecuencia de recombinación entre el marcador y la variante causal. Sin embargo, en el estudio de ligamiento genético de enfermedades complejas no se suele observar un tipo de herencia clara, por lo que se utilizan métodos no paramétricos. En éstos, no se asume ningún tipo de herencia concreta y se basan en la transmisión de alelos idénticos por descendencia (IBD) entre los individuos emparentados de estudio. Los alelos IBD hacen referencia a la compartición de copias idénticas de un mismo alelo ancestral, por lo que familiares más parecidos en fenotipo compartirán una proporción más alta de lo esperado de alelos IBD de un marcador que esté en ligamiento con el fenotipo de estudio. En concreto, cada uno de los sitios cromosómicos o *loci* que contienen genes que influyen en la variabilidad de los fenotipos cuantitativos se conoce como *quantitative trait locus* o QTL (Almasy and Blangero 2010; Ott et al. 2011).

Este tipo de estudios han tenido un gran éxito en la investigación de las enfermedades mendelianas, que son aquellas causadas por una sola variante genética en un solo gen, pero no tanto en el caso de las enfermedades complejas en las que están implicados diversos genes y variantes genéticas que aportan, a menudo, un riesgo pequeño o moderado (Risch and Merikangas 1996). Además, la penetrancia de la variante causal juega un papel importante en un análisis de cosegregación, siendo la penetrancia incompleta muy común en los fenotipos complejos (Marian 2012).

3.5.2.1 Mapeo fino

La estrategia para acotar las regiones marcadas por los análisis de ligamiento es el mapeo fino o *fine-mapping.* Este método supone una primera aproximación para la identificación de genes candidatos. En concreto, consiste en el genotipado de la región candidata de interés mediante marcadores adicionales con el fin de aumentar la información genética disponible (Carlson et al. 2004).

3.5.3 Estudios de asociación del genoma completo

Estos estudios consisten en el mismo concepto de análisis de asociación entre una variante genética y la variabilidad de un fenotipo de estudio que se ha utilizado en los estudios de asociación de genes candidatos. Sin embargo, con el desarrollo de chips para el genotipado de miles SNPs a gran escala se ha permitido la implantación de los GWAS sin necesidad de evidencias genéticas previas. Los SNPs se seleccionan como marcadores genéticos en base a patrones de LD. Es decir, en estos estudios se espera que uno de los marcadores utilizados sea causal o se encuentre suficientemente cerca de la variante causal, de manera que no haya afectación por los procesos de recombinación genética (Terwilliger and Weiss 1998; Marian 2012). Además, los SNPs pueden no presentar una funcionalidad evidente, al tratarse en muchas ocasiones de variantes genéticas localizadas en regiones no codificantes (Manolio et al. 2009). Los GWAS se han diseñado en base a la hipótesis de "enfermedades comunes - variantes comunes". Según esta hipótesis, las enfermedades complejas y comunes son el resultado de la suma del efecto de un gran número de variantes genéticas de efectos escasos y comunes (con una MAF alta). Esto puede deberse a que los alelos de riesgo no habrán sufrido una selección natural negativa intensa (Hirschhorn 2005; Marian 2012).

Estos estudios se han convertido en una poderosa herramienta con, hasta la fecha, más de 24.000 asociaciones entre SNPs y fenotipos relacionados con enfermedades humanas que han sido reportadas y recopiladas en el Catálogo de GWAS del Instituto Nacional de Investigación del Genoma Humano (NHGRI) y del Instituto Europeo de Bioinformática del laboratorio europeo de biología molecular (EMBL-EBI) (Welter et al. 2014).

3.5.3.1 Imputación de datos genotípicos

La imputación es la predicción del genotipo desconocido de variantes genéticas a partir de la comparación de las variantes genotipadas con un panel de referencia de haplotipos como HapMap o 1.000 Genomas. De esta manera, los datos genéticos de cada individuo aporta información sobre otras variantes genéticas en los mismos individuos, por lo que se aumenta significativamente la densidad génica de la muestra (Zeggini and Morris 2010). La imputación ha sido de gran utilidad en el aumento del poder estadístico de los estudios de GWAS (Kathiresan et al. 2008; Willer et al. 2008), así como en la aceleración de los estudios de mapeo fino (Liu et al. 2012) o para la facilitación de los meta-análisis de datos de asociación a nivel de todo el genoma (Neale et al. 2010). No obstante, el error de imputación aumenta a medida que la frecuencia alélica de la variante genética imputada disminuye (Tabla 4) (Li et al. 2011).

Tabla 4. Comparativa de la proporción de variantes imputadas de buena calidad según la frecuencia alélica del alelo minoritario. Adaptado (Li et al. 2011).

MAF de la variante imputada	Proporción de variantes imputadas de buena calidad [a]
≤0,1%	31%
0,1%<…≤0,5%	48%
0,5%<…≤1%	54%
1%<…≤5%	78%
>5%	97%

MAF: frecuencia del alelo menos común. [a] Plataforma Illumina 550k con una muestra de referencia de 3.713 individuos y r2>0,7.

3.5.4 Estudios de asociación de datos de secuenciación

Los análisis de asociación basados en datos obtenidos mediante NGS se postula como la solución a algunas de las limitaciones de los estudios de ligamiento y de GWAS (Wang et al. 2015). La NGS puede aplicarse a un grupo de genes candidatos, ofrecer una secuenciación de todo el exoma (WES) o de todo el genoma (WGS).

En conjunto, los estudios de genes candidatos y WES ofrecen muchas ventajas frente a las estrategias de WGS, como la posibilidad de utilizar muestras con mayor número de individuos debido a la optimización de los costes. Además, los análisis estadísticos se benefician de un mayor poder estadístico evitando las correcciones por comparación múltiple. Además, las variantes en regiones codificantes son más fáciles de interpretar (Wang et al. 2015). Sin embargo, existen evidencias en el estudio de las enfermedades complejas de variantes genéticas de riesgo localizadas en regiones no codificantes que influyen en la regulación de la transcripción con un impacto moderado en la expresión del gen (Hirschhorn 2005). Por este motivo, la inclusión de las regiones no codificantes en las estrategias de genes candidatos o WGS supone un cambio de orientación en la investigación de la genética humana, donde se ha restringido la búsqueda de variantes genéticas a las zonas codificantes y al promotor. A pesar de que la estrategia de WGS permite la identificación de más variantes genéticas, así como de algunos tipos de variantes estructurales, éste no ha sido un diseño tan utilizado. No obstante, ha sido elegida como estrategia alternativa en aquellos estudios sin resultados concluyentes mediante WES (Morange et al. 2015).

La gran cantidad de datos que aporta la NGS supone un reto a nivel bioinformático y estadístico. En concreto, los análisis de asociación de variantes raras y de baja frecuencia alélica representan el principal reto en la evaluación de los resultados de NGS. Estas variantes genéticas conllevan una disminución del poder estadístico frente al análisis de variantes comunes debido a su baja frecuencia alélica, a menos que el

efecto sea muy potente o la muestra contenga muchos individuos. A la vez, suelen necesitar muchas correcciones por comparaciones múltiples. En consecuencia, se han propuesto diversos métodos de asociación en los que no se evalúa cada variante por separado, sino que se tiene en cuenta el efecto acumulativo de distintas variantes agregadas por región o por gen. Por otra parte, los métodos estadísticos de metaanálisis permiten combinar la información de distintos estudios, aumentando el número de la muestra (Lee et al. 2015; Wang et al. 2015).

3.6 Caracterización funcional *in silico* de variantes genéticas

La implicación de variantes genéticas en la variabilidad del fenotipo de estudio se puede caracterizar a partir de la información disponible en las bases de datos. Las herramientas que permiten interrogar paralelamente a distintos programas bioinformáticos de simulación computacional o *in silico* resultan de especial interés, como es el caso de la interfaz Alamut Visual (Interactive Biosoftware, Ruan, Francia; www.interactive-biosoftware.com), que se emplea de forma habitual en el diagnóstico molecular. No obstante, los resultados predictivos no se deben interpretar como evidencias únicas y suficientes (Wallis et al. 2013).

La mayoría de programas de evaluación *in silico* se centran en las variantes situadas en regiones codificantes (Wang et al. 2015). Los programas predictivos para el análisis de variantes de cambio de sentido se diferencian en los algoritmos utilizados. Pueden estar basados en la conservación evolutiva de las secuencias de DNA, como es el caso de SIFT (Kumar et al. 2009) y Align-GVGD (Tavtigian et al. 2006), en la secuencia y estructura de las proteínas, modelo aplicado en PolyPhen-2 (Adzhubei et al. 2010), o en el aprendizaje automático, como en el caso de MutationTaster (Schwarz et al. 2010). Las recomendaciones actuales de la Asociación para la Ciencia

Genética Clínica (ACGS) se fundamentan en el uso de al menos tres programas predictivos, en los que deberían estar representados los distintos tipos de algoritmos desarrollados (Wallis et al. 2013). Otros programas bioinformáticos como Project HOPE (Centro de Informática Molecular y Biomolecular, Nimega, Países Bajos; www.cmbi.ru.nl/hope/) (Venselaar et al. 2010) predicen los cambios estructurales que puede sufrir la proteína.

Por el contrario, no se dispone de tanta información sobre la funcionalidad de variantes genéticas localizadas en posiciones no exónicas, a pesar de que constituyen aproximadamente el 98% del genoma humano (Wang et al. 2015). Para la evaluación del efecto de variantes genéticas en regiones de *splicing* cercanas al punto de corte y empalme entre exones e intrones en el sitio donador (5'ss) o en el sitio aceptor (3'ss) se han desarrollado distintas herramientas bioinformáticas. Éstas pueden basarse en distintos algoritmos como las matrices de peso posicionales (PWM) utilizados en los programas Splice Site Finder-like (integrado en la interfaz Alamut Visual v.2.6.1), MaxEntScan (Yeo and Burge 2004) o Human Splicing Finder (Desmet et al. 2009), los modelos de redes neuronales (NN) aplicados en NNSplice (Reese et al. 1997) y los modelos de máxima dependencia de descomposición (MDD) en el caso de GeneSplicer (Pertea et al. 2001). Actualmente, la ACGS recomienda el consenso de al menos dos de estas herramientas para la predicción de un efecto deletéreo (Wallis et al. 2013).

4. Proyecto *Genetic Analysis of Idiopathic Thrombophilia* (GAIT)

El primer estudio de la trombosis como enfermedad compleja y multifactorial a nivel de todo el genoma (estudios de ligamiento genético y GWAS) fue el Proyecto *Genetic Analysis of Idiopathic Thrombophilia* 1 (GAIT-1). Además, ha sido pionero en la incorporación de las técnicas de estadística genética para fenotipos cuantitativos y en el diseño basado en reclutamiento de familias grandes. Gran parte de las evidencias de las que partimos en este estudio se basan en los resultados obtenidos en este proyecto.

El Proyecto *Genetic Analysis of Idiopathic Thrombophilia* 1 (GAIT-1) comenzó en el Hospital de la Santa Creu i Sant Pau de Barcelona, España, en el año 1995. Su principal objetivo consiste en identificar los factores genéticos que influyen en la variación cuantitativa de los fenotipos relacionados con la hemostasia y en el riesgo de padecer una enfermedad tromboembólica.

Bajo este mismo objetivo comenzó en el año 2006 el Proyecto *Genetic Analysis of Idiopathic Thrombophilia* 2 (GAIT-2) en el Hospital de la Santa Creu i Sant Pau, el cual recoge un número considerablemente mayor de familias extensas con trombofilia, así como de datos clínicos y datos fenotípicos. Con este proyecto se pretende aumentar el poder estadístico para identificar la base genética de la trombosis, así como el estudio de la variabilidad genética de la expresión del RNA, la investigación de la base epigenética y la exploración de otros sistemas fisiológicos y metabólicos, como el estrés oxidativo, la estructura y función de las plaquetas o los mediadores de la inflamación. Además, permite la replicación de los resultados obtenidos en el Proyecto GAIT-1.

4.1 Proyecto GAIT-1

Se reclutaron 398 individuos distribuidos en 21 familias extensas españolas, de las cuales 12 fueron seleccionadas a partir de un individuo afecto de trombofilia idiopática (214 individuos) y las 9 restantes eran familias controles (184 individuos). Los requisitos para la inclusión en el proyecto han sido reportados previamente (Souto et al. 2000b; Souto et al. 2000a) y consisten en tener más de 10 miembros vivos en la familia y disponer, como mínimo, de 3 generaciones. El concepto trombofilia se definió como múltiples eventos trombóticos (siendo al menos uno de ellos espontáneo), un único evento trombótico espontáneo con un familiar de primer grado también afecto, o bien un primer evento trombótico antes de los 45 años. La trombosis se consideró idiopática a partir de la exclusión de todas las causas biológicas conocidas de trombosis durante el periodo de reclutamiento de las familias, entre las que se incluía deficiencia de AT, deficiencia de PS, deficiencia de PC, APCR, deficiencia de plasminógeno, deficiencia de cofactor II de la heparina, FVL, disfibrinogenemia, lupus anticoagulante y anticuerpos antifosfolípidos.

Todos los individuos fueron entrevistados por un médico que determinó su historial de salud y reproductivo, así como medidas antropométricas, como peso y altura, y los medicamentos habituales, por ejemplo, el uso de anticonceptivos orales. En concreto, se obtuvo información respecto a eventos trombóticos sufridos, tanto venosos como arteriales, la edad en la que estos eventos se produjeron y otros factores de riesgo cardiovascular como el tabaquismo, la diabetes y dislipemias. Además, se obtuvo información respecto al estilo de vida y la residencia principal, para poder determinar así la contribución de componentes ambientales compartidos entre los miembros de un mismo domicilio como, por ejemplo, la dieta. A cada uno de los individuos del proyecto se le midieron 66 fenotipos intermediarios cuantitativos relacionados con la hemostasia, el metabolismo del hierro, el metabolismo de los lípidos y el complemento.

Específicamente, el rango de edad de los individuos osciló entre 1 año y los 88 años, con una edad media de 37,7 años, y estaba formado por un porcentaje muy similar de hombres (45,98%) y mujeres (54,02%).

4.1.1 Aportación científica del Proyecto GAIT-1

Algunos de los hallazgos más relevantes del Proyecto GAIT-1 incluyen la heredabilidad o cuantificación del componente genético del riesgo de VTE y de los 27 fenotipos intermediarios cuantitativos relacionados con la hemostasia (Tabla 5). Igualmente, el Proyecto GAIT-1 también ha descrito las correlaciones fenotípicas, genotípicas y ambientales entre estos fenotipos y la VTE (Tabla 6) (Souto et al. 2000b; Souto et al. 2000a). Teniendo en cuenta estos resultados, cabe destacar la existencia de un grupo de especial interés por su alta heredabilidad y su implicación en el riesgo de enfermedad tromboembólica. Concretamente, este grupo incluye el FXII, FXI, FIX y FVIII, que definen la vía intrínseca de la coagulación. Las importantes evidencias descritas confirman la relevancia de incluir estos fenotipos en el estudio de la base genética implicada en la variabilidad del riesgo de VTE.

Tabla 5. Heredabilidades de la VTE y de los fenotipos cuantitativos del Proyecto GAIT-1. Adaptado (Souto et al. 2000b; Souto et al. 2000a).

Fenotipo	Heredabilidad ± SE
VTE	0,61 ± 0,16
aPTT	0,83 ± 0,07
APCR	0,71 ± 0,08
FXII	0,67 ± 0,09
FVII	0,52 ± 0,09
HRG	0,52 ± 0,09
TFPI	0,52 ± 0,09
PT	0,50 ± 0,09
PC	0,50 ± 0,09
Protrombina	0,49 ± 0,0
AT	0,49 ± 0,09
PS total	0,46 ± 0,09
PS funcional	0,45 ± 0,10
FXI	0,45 ± 0,10
FV	0,44 ± 0,09
Cofactor II de la heparina	0,44 ± 0,09
FX	0,43 ± 0,13
FVIII	0,40 ± 0,09
FIX	0,39 ± 0,09
Fibrinógeno	0,34 ± 0,10
vWF	0,32 ± 0,11
PAI-1	0,30 ± 0,08
t-PA	0,27 ± 0,07
Homocisteína	0,24 ± 0,08
Plasminógeno	0,24 ± 0,10
PS libre	0,22 ± 0,10
TF	0,17 ± 0,08
Dímero D	0,11 ± 0,09

SE: error estándar; VTE: enfermedad tromboembólica venosa; aPTT: tiempo de tromboplastina parcial activada; APCR: resistencia a la proteína C activada; FXII: factor XII; FVII: factor VII; HRG: glicoproteína rica en histidina; TFPI: inhibidor del factor tisular; PT: tiempo de protrombina; PC: proteína C; AT: antitrombina; PS: proteína S; FXI: factor XI; FV: factor V; FX: factor X; FVIII: factor VIII; FIX: factor IX; vWF: factor de von Willebrand; PAI-1: inhibidor del activador de plasminógeno; t-PA: activador tisular de plasminógeno; TF: factor tisular.

Tabla 6. Correlaciones fenotípicas, genéticas y ambientales entre los fenotipos cuantitativos en el Proyecto GAIT-1 y la VTE. Adaptado (Souto et al. 2000a).

Fenotipo [a]	ρ_f	p-valor (ρ_f)	ρ_g	p-valor (ρ_g)	ρ_a	p-valor (ρ_a)
APCR	-0,23	$3{,}00\times10^{-04}$	-0,65	$1{,}00\times10^{-06}$	0,67	$6{,}00\times10^{-04}$
FVII	0,03	NS	-0,35	$5{,}64\times10^{-02}$	0,57	$9{,}10\times10^{-03}$
FVIII	0,29	$2{,}00\times10^{-04}$	0,69	$5{,}00\times10^{-04}$	-0,13	NS
FIX	0,15	$7{,}87\times10^{-02}$	0,60	$1{,}31\times10^{-02}$	-0,20	NS
FXI	0,21	$1{,}80\times10^{-02}$	0,56	$2{,}45\times10^{-02}$	0,07	NS
FXII	0,17	$3{,}39\times10^{-02}$	0,35	$5{,}00\times10^{-02}$	-0,15	NS
Homocisteína	0,23	$1{,}80\times10^{-03}$	0,65	$1{,}50\times10^{-03}$	-0,03	NS
t-PA	0,18	$2{,}00\times10^{-04}$	0,75	$7{,}00\times10^{-03}$	-0,10	NS
vWF	0,26	$1{,}00\times10^{-03}$	0,73	$5{,}00\times10^{-04}$	-0,18	NS

APCR: resistencia a la proteína C activada; FVII: factor VII; FVIII: factor VIII; FIX: factor IX; FXI: factor XI; FXII: factor XII; t-PA: activador tisular de plasminógeno; vWF: factor de von Willebrand; ρ_f: correlación fenotípica; ρ_g: correlación genética; ρ_a: correlación ambiental; NS: resultado no estadísticamente significativo (p-valor > 0,1). [a] Fenotipos relacionados con la hemostasia con al menos una correlación estadísticamente significativa (p-valor <0,05).

4.2 Proyecto GAIT-2

El Proyecto GAIT-2 está formado por un nuevo grupo de 935 individuos agrupados en 35 familias extensas españolas. En este proyecto, los criterios de inclusión y reclutamiento han sido los mismos descritos en el Proyecto GAIT-1. En concreto, los individuos incluidos han sido entrevistados por un médico que determina el historial de salud y reproductivo, así como medidas antropométricas, medicamentos habituales, como el uso de anticonceptivos orales o terapia de reemplazo hormonal, y el consumo de alcohol. Por otra parte, también se ha determinado la actividad física mediante el formato corto del cuestionario internacional de actividad física (IPAQ). Además, se ha registrado información detallada respecto a los antecedentes trombóticos (venosos y arteriales), la edad y la presencia de factores de riesgo relacionados como el tabaquismo. Además, se ha recogido más información sobre enfermedades como la diabetes, dislipemias, asma, rinitis alérgica, dermatitis atópica, enfermedad autoinmune y cáncer. Finalmente, se ha registrado la residencia principal de cada uno

de los individuos, para determinar el efecto de componentes ambientales compartidos entre los individuos de un mismo domicilio, como la dieta. En este proyecto, se han estudiado 472 fenotipos cuantitativos, el rango de edad de los individuos varía de los 2,6 años a los 101 años, con una edad media de 39,5 años, y el porcentaje de hombres (49,73%) y mujeres (50,27%) es muy similar. La comparativa de las principales características del Proyecto GAIT-1 y GAIT-2 se muestran en la Tabla 7.

Tabla 7. Comparativa entre el Proyecto GAIT-1 y el Proyecto GAIT-2.

Características	Proyecto GAIT-1	Proyecto GAIT-2
Inicio del reclutamiento	1995	2006
Fin del reclutamiento	1997	2010
Individuos totales	398	935
Edad	1-88	2,6-101
Familias	21	35
Parejas de parentesco	2744	8649
Niños (<18 años)	69	197
Individuos con trombosis	53	120
Individuos con VTE	40	86
Individuos con ATE	17	47
Individuos con VTE y ATE	4	13
Total fenotipos estudiados	66	472

Objetivos

Objetivos

El objetivo global de esta Tesis Doctoral consiste en identificar variantes genéticas que determinan la variabilidad de fenotipos de la hemostasia relacionados con el riesgo de padecer eventos trombóticos. Para ello queremos aplicar una estrategia global, integradora e innovadora de diversos métodos de análisis genético y herramientas bioinformáticas disponibles para el estudio de las enfermedades complejas en genética humana. Los resultados esperados pueden contribuir a una mejora en el campo diagnóstico, preventivo y terapéutico de la VTE. Los objetivos concretos son los siguientes:

- Identificar, caracterizar y validar fenotipos intermediarios potencialmente relacionados con la VTE como herramientas para la detección de genes candidatos.

- Identificar genes que participan en la variabilidad de fenotipos de la hemostasia como potenciales factores de riesgo de VTE empleando métodos de análisis de genoma completo. Para ello se proponen dos abordajes:

    - El estudio de la variabilidad de fenotipos individuales de la cascada de la coagulación.

- El estudio de la variabilidad de fenotipos que aporten información global e integradora de la cascada de la coagulación y de la fibrinólisis.

- Incorporar la metodología de NGS para determinar la estructura alélica de los genes candidatos identificados y las variantes genéticas responsables de la variabilidad de los niveles de los fenotipos de la cascada de la coagulación.

- Explorar la idoneidad de la NGS para el estudio de genes relacionados con la VTE y su aplicación al diagnóstico clínico.

Resultados

Informe del director

La memoria de la Tesis Doctoral “Análisis Genómico de Fenotipos de la Hemostasia Relacionados con la Enfermedad Tromboembólica Venosa” se presenta como un compendio de cinco artículos científicos. Tres de estos arículos han sido publicados en revista internacionales. Otro está en estos momentos sometido en una revista internacional. Por último, uno de los artículos está actualmente en proceso de preparación. La participación de la doctoranda se especifica a continuación:

Artículo 1

Título: *Genetic Determinants of Thrombin Generation and Their Relation to Venous Thrombosis: Results from the GAIT-2 Project.*

Autores: Laura Martin-Fernandez, Andrey Ziyatdinov, Marina Carrasco, Juan Antonio Millon, Angel Martinez-Perez, Noelia Vilalta, Helena Brunel, Montserrat Font, Anders Hamsten, Juan Carlos Souto y José Manuel Soria.

Referencia: *PloS one* 2016; 11(1):e0146922. doi: 10.1371/journal.pone.0146922. PMID: 26784699. Factor de Impacto (2015): **3,057**.

Aportación de la doctoranda en el artículo: La doctoranda, Laura Martín Fernández, ha participado activamente en el diseño del estudio y ha sido la resposable de la elaboración del manuscrito final, integrando todos los resultados y generando la discusión de los mismos. Ha participado principalmente en la interpretación de los datos estadísticos y ha sido la encargada de la discusión de los resultados.

Artículo 2

Título: *Genetics Determinants for Factor VIII Levels: Genome-Wide Linkage and Association Analyses from the GAIT Project.*

Autores: Sonia Lopez, **Laura Martin-Fernandez**, Andrey Ziyatdinov, Angel Martinez-Perez, Ares Rocañín, Giovana Gavidia-Bovadilla, Juan Carlos Souto, y José Manuel Soria.

Referencia: Artículo en preparación, pendiente de enviar a *Journal of Thrombosis and Haemostasis*. Factor de Impacto (2015): **5,565.**

Aportación de la doctoranda en el artículo: Como director de la Tesis Doctoral de Laura Martín Fernández hago constar que la doctoranda ha participado activamente en el diseño y en la preparación del manuscrito final. La doctoranda ha sido la encargada de todo el trabajo experimental realizado mediante la metodología de NGS y ha participado en la selección de los individuos que se han incluido en el estudio concreto de secuenciación del gen candidato. La doctoranda ha diseñado los *primers* y se ha encargado de optimizar las condiciones de las PCRs largas (LR-PCR) para asegurar la amplificación completa del gen candidato seleccionado. Además, ha preparado la normalización mediante cuantificación de las LR-PCRs, así como ha creado las librerías. La doctoranda ha realiazado la optimización de la técnica de laboratorio de secuenciación y ha realizado el análisis bioinformático de los datos obtenidos mediante NGS y su anotación. Por último, ha participado en el diseño de los análisis de asociación entre las variantes genéticas identificadas mediante NGS y los niveles de FVIII.

Artículo 3

Título: *The Central Role of KNG1 Gene as a Genetic Determinant of Coagulation Pathway-Related Traits: Exploring Metaphenotypes.*

Autores: Helena Brunel, Raimon Massanet, Angel Martinez-Perez, Andrey Ziyatdinov**, Laura Martin-Fernandez**, Juan Carlos Souto, Alexandre Perera y José Manuel Soria.

Referencia: *PloS one* 2016. 11(12): e0167187. doi: 10.1371/journal.pone.0167187. PMID: 28005926. Factor de Impacto (2015): **3,057**.

Aportación de la doctoranda en el artículo: La doctoranda, Laura Martín Fernández, ha participado en el diseño de la metodología y en la elaboración del manuscrito final. Además, ha participado en la interpretación biológica de los resultados. Parte de la información incluida en esta publicación (especialmente la parte del desarrollo estadístico) forma parte de la Tesis Doctoral de Helena Brunel (http://hdl.handle.net/10803/134362). La Tesis de la Dra. Brunel, vinculada al Departamento de Ingeniería de Sistemas, Automática e Informática Industrial (Universidad Politécnica de Cataluña), se realizó en formato clásico sin compendio de publicaciones. Concretamente, este artículo científico estaba sometido cuando se defendió la Tesis de la Dra. Brunel.

Artículo 4

Título: *Next Generation Sequencing to Dissect the Genetic Architecture of KNG1 and F11 Loci using Factor XI Levels as an Intermediate Phenotype of Thrombosis.*

Autores: Laura Martin-Fernandez, Giovana Gavidia-Bovadilla, Irene Corrales, Helena Brunel, Lorena Ramírez, Sonia López, Juan Carlos Souto, Francisco Vidal y José Manuel Soria.

Referencia: *PloS one* 2016. Artículo sometido. Factor de Impacto (2015): **3,057**.

Aportación de la doctoranda en el artículo: La doctoranda, Laura Martín Fernández, ha participado activamente en el diseño, análisis de resultados, discusión de conclusiones y preparación del manuscrito final. La doctoranda se ha encargado de toda la realización experimental incluida, de la preparación del manuscrito final y de la discusión de las conclusiones. En concreto, ha participado en la fase de selección de individuos, ha diseñado los *primers* y ha realizado tanto la optimización de las condiciones de amplificación (tanto PCRs cortas como LR-PCRs) así como la propia amplificación de los individuos seleccionados. Además, se ha encargado de la normalización de dichas PCRs y de la preparación de las librerías. La doctoranda ha llevado a cabo la secuenciación mediante NGS y el análisis bioinformático posterior de los resultados obtenidos. También ha realizado la anotación de las variantes genéticas identificadas y se ha ocupado de la validación, replicación y realización de análisis *in silico* de potenciales mutaciones patogénicas. Por último, ha participado en el diseño del análisis de los estudios de asociación entre las variantes genéticas obtenidas mediante NGS y los niveles de FXI.

Artículo 5

Título: *The Unravelling of the Genetic Architecture of Plasminogen Deficiency and its Relation to Thrombotic Disease*.

Autores: **Laura Martin-Fernandez**, Pascual Marco, Irene Corrales, Raquel Pérez, Lorena Ramírez, Sonia López, Francisco Vidal y José Manuel Soria.

Referencia: *Scientific Reports* 2016; 6:39255. doi: 10.1038/srep39255. PMID: 27976734. Factor de Impacto (2015): **5,228**.

Aportación de la doctoranda en el artículo: La doctoranda, Laura Martín Fernández, ha participado en el diseño del estudio y ha sido la encargada de la obtención, análisis e interpretación de los datos genéticos obtenidos, así como de la elaboración del

manuscrito final. Ha participado en la obtención del DNA de los individuos incluidos en el estudio. Además, ha llevado a cabo el diseño de los *primers*, la puesta a punto y creación de las LR-PCRs, la preparación de las librerías y la secuenciación mediante NGS. También ha realizado el análisis bioinformático de las secuencias obtenidas, así como la anotación de las variantes genéticas y la identificación, validación y evaluación *in silico* de las potenciales mutaciones patogénicas.

Barcelona, 2017

José Manuel Soria Fernández

Artículo 1

Título

Genetic Determinants of Thrombin Generation and Their Relation to Venous Thrombosis: Results from the GAIT-2 Project.

Autores

Laura Martin-Fernandez, Andrey Ziyatdinov, Marina Carrasco, Juan Antonio Millon, Angel Martinez-Perez, Noelia Vilalta, Helena Brunel, Montserrat Font, Anders Hamsten, Juan Carlos Souto y José Manuel Soria.

Referencia

PloS one 2016; 11(1):e0146922. doi: 10.1371/journal.pone.0146922. PMID: 26784699.

Resumen

La VTE es una enfermedad común y multifactorial. A pesar de que la base genética parece ser la más importante, sólo se ha conseguido identificar, hasta la fecha, una pequeña parte de ésta. Por lo tanto, el uso de fenotipos intermediarios como puede ser el caso del test de generación de trombina (TGT) es de gran interés en los análisis para la identificación de nuevos factores de riesgo genéticos que contribuyan al riesgo de padecer VTE. Con el objetivo de caracterizar la relación del TGT con la VTE y, además, poder explorar las bases genéticas de este test clínico, se ha medido el tiempo de latencia, la altura del pico de trombina y el potencial endógeno de trombina (ETP), como fenotipos cuantitativos del TGT, en los 935 individuos agrupados en 35 familias extensas que forman el Proyecto GAIT-2. Los resultados muestran que los factores genéticos son los responsables del 67% de la variación del riesgo de padecer

VTE. Se han estimado también las heredabilidades del tiempo de latencia (49%), de la altura del pico de trombina (54%) y del ETP (52%) y se ha demostrado una correlación genética positiva tanto de la altura del pico de trombina como del ETP con el riesgo de VTE. Además, se ha realizado un GWAS y análisis de imputación que muestran que el mayor determinante genético de la generación de trombina es el gen *F2*, aunque también se observan otras señales sugestivas. Por lo tanto, concluimos que los factores genéticos juegan un papel importante no sólo en la enfermedad, sino también en los fenotipos cuantitativos del TGT, y reportamos, por primera vez, evidencias de los efectos de pleiotropía entre estos fenotipos y el riesgo de VTE. En especial, este trabajo ha permitido identificar al gen *F2* como el principal responsable genético de la variabilidad de la generación de trombina.

PLOS ONE

RESEARCH ARTICLE

Genetic Determinants of Thrombin Generation and Their Relation to Venous Thrombosis: Results from the GAIT-2 Project

Laura Martin-Fernandez[1], Andrey Ziyatdinov[1], Marina Carrasco[2], Juan Antonio Millon[2], Angel Martinez-Perez[1], Noelia Vilalta[2], Helena Brunel[1], Montserrat Font[2], Anders Hamsten[3], Juan Carlos Souto[2]*, José Manuel Soria[1]

1 Unit of Genomics of Complex Diseases, Biomedical Research Institute Sant Pau (IIB-Sant Pau), Barcelona, Spain, **2** Unit of Hemostasis and Thrombosis, Department of Hematology, IIB-Sant Pau, Hospital de la Santa Creu i Sant Pau, Barcelona, Spain, **3** Cardiovascular Genetics and Genomics Group, Atherosclerosis Research Unit, Department of Medicine, Karolinska Institutet, Karolinska University Hospital Solna, Stockholm, Sweden

* jsouto@santpau.cat

OPEN ACCESS

Citation: Martin-Fernandez L, Ziyatdinov A, Carrasco M, Millon JA, Martinez-Perez A, Vilalta N, et al. (2016) Genetic Determinants of Thrombin Generation and Their Relation to Venous Thrombosis: Results from the GAIT-2 Project. PLoS ONE 11(1): e0146922. doi:10.1371/journal.pone.0146922

Editor: Esteban Gándara, Ottawa Hospital Research Institute, CANADA

Received: November 20, 2015

Accepted: December 23, 2015

Published: January 19, 2016

Data Availability Statement: All relevant data are within the paper.

Funding: This study was supported by funds from the Instituto de Salud Carlos III Fondo de Investigación Sanitaria PI 11/0184, Red Investigación Cardiovascular RD12/0042/0032 and AGAUR 2009 SGR 1147 from Generalitat de Catalunya. Laura Martin-Fernandez was supported by Ayudas Predoctorales de Formación en Investigación en Salud (PFIS) FI12/00322. Juan Antonio Millon was supported by Fundació Josep Carreras contra la leucemia and Institut Josep Carreras (IJC). The

Abstract

Background

Venous thromboembolism (VTE) is a common disease where known genetic risk factors explain only a small portion of the genetic variance. Then, the analysis of intermediate phenotypes, such as thrombin generation assay, can be used to identify novel genetic risk factors that contribute to VTE.

Objectives

To investigate the genetic basis of distinct quantitative phenotypes of thrombin generation and its relationship to the risk of VTE.

Patients/Methods

Lag time, thrombin peak and endogenous thrombin potential (ETP) were measured in the families of the Genetic Analysis of Idiopathic Thrombophilia 2 (GAIT-2) Project. This sample consisted of 935 individuals in 35 extended families selected through a proband with idiopathic thrombophilia. We performed also genome wide association studies (GWAS) with thrombin generation phenotypes.

Results

The results showed that 67% of the variation in the risk of VTE is attributable to genetic factors. The heritabilities of lag time, thrombin peak and ETP were 49%, 54% and 52%, respectively. More importantly, we demonstrated also the existence of positive genetic correlations between thrombin peak or ETP and the risk of VTE. Moreover, the major genetic determinant of thrombin generation was the *F2* gene. However, other suggestive signals were observed.

funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Conclusions

The thrombin generation phenotypes are strongly genetically determined. The thrombin peak and ETP are significantly genetically correlated with the risk of VTE. In addition, *F2* was identified as a major determinant of thrombin generation. We reported suggestive signals that might increase our knowledge to explain the variability of this important phenotype. Validation and functional studies are required to confirm GWAS results.

Introduction

Venous thromboembolism (VTE) is a common disease involving genetic and environmental risk factors and their interactions [1]. VTE, which consist of deep vein thrombosis and pulmonary embolism, has an annual incidence of approximately 1 in 1000 individuals in developed countries [2]. Previous studies have estimated that the risk of VTE has a heritability of approximately 60% [3,4]. This means that the proportion of the variance that is attributable to genetic effects is approximately 60%. Specifically, hemostasis phenotypes have been used as intermediate phenotypes to identify novel genetic risk factors that contribute to thrombotic disease [3,5,6].

One of these intermediate phenotypes could be a measure of the thrombin generation assay, which measures the capacity to generate thrombin, a key enzyme in the coagulation cascade [7]. This phenotype is a global coagulation assay developed by Hemker *et al.* [8]. There are distinct quantitative phenotypes such as lag time, thrombin peak and the endogenous thrombin potential (ETP) representative of the dynamics of thrombin generation. Lag time corresponds to the time from the beginning of the test to the time when thrombin formation begins. Thrombin peak is defined as the highest thrombin concentration detectable. The ETP is determined from the area under the thrombin generation curve that measures the enzymatic capacity of thrombin generated during its lifetime [8,9].

Several studies have reported an association between these quantitative thrombin generation phenotypes and the risk of cardiovascular diseases using different test conditions. Specifically, an association has been described between thrombin generation and the risk of first VTE [10–13]. However, the results of the risk of recurrence are equivocal [11,14–16]. Thrombin generation phenotypes have been associated also with the risk of ischemic stroke [17,18] and acute myocardial infarction [19]. Also, the heritability of thrombin generation has been estimated previously at 32% [12]. These results indicate that the study of the genetic basis of thrombin generation phenotypes should identify novel genetic factors involved in the risk of VTE.

In addition to the Genetic Analysis of Idiopathic Thrombophilia 1 (GAIT-1) Project [3,20], we initiated another project—the Genetic Analysis of Idiopathic Thrombophilia 2 (GAIT-2) Project, using a new set of families. Several intermediate phenotypes were analyzed in the families of the GAIT-2 Project, including thrombin generation phenotypes. One aim of the present study was to determine the heritability of the risk of VTE as compared to the heritability values found in the GAIT-1 Project, and to estimate the genetic basis of 3 phenotypes related to thrombin generation. To evaluate their function as intermediate phenotypes with the risk of VTE, another aim was to estimate the phenotypic, genetic and environmental correlations of these 3 phenotypes with the risk of VTE. Finally, we wanted to perform genome wide association studies (GWAS) to identify susceptibility loci for thrombin generation phenotypes.

Material and Methods

Subjects

The Spanish families in our study were recruited in the GAIT-2 Project at the Hospital de la Santa Creu i Sant Pau of Barcelona, Spain.

The recruitment and the criteria used for inclusion were the same as that in GAIT-1 and have been described in detail previously [20]. Briefly, to be included in this study, a family was required to have at least 10 living individuals in 3 or more generations. Families were selected through a proband with idiopathic thrombophilia, which was defined as recurrent thrombotic events (at least one of which was spontaneous), a single spontaneous thrombotic episode plus a first-degree relative also affected, or onset of thrombosis before age 45. Thrombosis in these probands was considered idiopathic following the same criteria as in the GAIT-1 Project (excluding biological causes as antithrombin deficiency, Protein S and C deficiencies, activated protein C resistance, plasminogen deficiency, heparin cofactor II deficiency, Factor V Leiden, dysfibrogenemia, lupus anticoagulant and antiphospholipid antibodies). The subjects were interviewed by a physician to determine their health and reproductive history, current medications, alcohol consumption, use of sex hormones (oral contraceptives or hormonal replacement therapy) and their smoking history. Physical activity was also determined using the short form of the International Physical Activity Questionnaire (IPAQ) [21].

Our subjects were questioned also about previous episodes of venous and arterial thrombosis, the age at which these events occurred, and the presence of potentially correlated disorders such as diabetes, lipid disease, asthma, allergic rhinitis, atopic dermatitis, autoimmune disease and cancer. The residence of each subject was determined to assess the contribution of shared environmental influences (such as diet) common to members of the same household. The study was performed according to the Declaration of Helsinki. All procedures were reviewed and approved by the Institutional Review Board of the Hospital de la Santa Creu i Sant Pau, Barcelona, Spain. Adult subjects gave written informed consent for themselves and for their minor children.

This sample consisted of 935 individuals in 35 extended families. All of the pedigrees contained at least 3 generations, and 14 families had more than 3 generations. The individuals ranged in age from 2.6 to 101 years old (SD = 21.4), with a median age of 39.5 and approximately

Table 1. Relative pairs.

n pairs	Kinship coefficient x 2	Relation
935	1	Self
1001	0.5	Parent-offspring
597	0.5	Siblings
467	0.25	Grandparent-grandchild
1248	0.25	Avuncular
7	0.25	Half siblings
3	0.25	Double 1st cousins
1807	0.125	3rd degree
1697	0.0625	4th degree
1296	0.03125	5th degree
349	0.015625	6th degree
177	0.0078125	7th degree
7116	0	Unrelated

doi:10.1371/journal.pone.0146922.t001

an equal number of males (465) and females (470). The depth and complexity of the pedigrees are illustrated in Table 1 with the number of pairs of relatives contained therein.

The sample had 120 subjects with thromboembolism, including venous and arterial thrombosis. Specifically, they were 86 subjects with venous thrombosis, 47 with arterial thrombosis and 13 with both venous and arterial thrombosis. Of the 935 subjects, we obtained the thrombin generation phenotypes of 919 subjects.

Blood Collection

Blood was collected by venipuncture following a 12-hour fast. Samples were collected in 1/10 volume containing 0.129 mol/L sodium citrate. None of the participants was using oral anticoagulants or heparins at the time of blood collection. Platelet-poor plasma (PPP) was obtained by centrifugation at 2000 *g* for 20 minutes at room temperature (22±2°C) and stored at −80°C before performing the thrombin generation assay. DNA was extracted from whole blood samples using a standard salting out procedure [22].

Thrombin Generation Assay

Thrombin generation was evaluated in PPP according to the method described by Hemker *et al.* [23] by means of Fluoroskan Ascent (Thermo Labsystems, Helsinki, Finland), an automated fluorometer with a 390/460 nm filter set. Measurements were conducted on 80 μL of PPP and 20 μL of PPP reagent was added (Thrombinoscope BV, Maastricht, The Netherlands) consisting of a final concentration of 5 pM tissue factor and 4 μM phospholipids. The thrombin calibrator and the fluorogenic substrate with $CaCl_2$ FluCa-Kit from Thrombinoscope BV (Maastricht, The Netherlands) were used. Thrombin generation curves were calculated using the Thrombinoscope™ software (Synapse BV, Maastricht, The Netherlands) and the parameters analyzed were lag time (min), thrombin peak (nM), and ETP (nM*min). The medians of crude lag time, crude thrombin peak and crude ETP were 2.67 min (first quartile 2.33 min and third quartile 3.08 min), 315.80 nM (first quartile 224.80 nM and third quartile 391.50 nM) and 1,594 nM*min (first quartile 1,260 nM*min and third quartile 1,980 nM*min). The intra-assay and inter-assay coefficients of variation for thrombin generation parameters were below 6% and below 8%, respectively.

Genotyping and Imputation

We genotyped the samples with a combination of HumanOmniExpressExome-8v1.2 (324 individuals) and HumanCoreExome-12v1.1 (610 individuals). We applied inclusion filters in the datasets based on call rate (>98%), HWE (p-value $>1.00x10^{-06}$) and MAF (>1%). We merged the data and obtained 395,556 SNPs in all of the samples. Then, we estimated haplotypes using SHAPEIT v2 [24] and imputed genotypes to the 1000 genomes phase 1 panel using IMPUTE2 [25]. We obtained 37,985,264 SNPs.

Statistical Analyses

Prior to the data analyses, statistics logarithmic transformations were performed to normalize distributions in traits.

The statistical methods used in our study have been described elsewhere [3,26]. The analysis of heritability (h^2, the relative proportion of phenotypic variance of the trait attributable to the additive effects of genes) was performed using the variance component method [27]. The total phenotypic variance was partitioned into three components: (1) an additive genetic variance that is caused by the sum of the average effects of all of the genes that influence the trait; (2) a shared environmental variance that is caused by the environmental factors that are common to members

of a household (c^2); and (3) a residual random environmental variance that is specific to each individual. The random environmental variance also includes non-additive genetic effects such as interactions between alleles within loci (dominance effects), interactions between alleles at different loci (epistatic effects), and effects caused by gene-environment interactions. Therefore, this model generally underestimates the role of genetics in the determination of the trait.

The covariances among individuals within a family that are due to additive genetic effects were estimated as a function of their expected genetic kinship relationships. Covariances among individuals that are due to shared environments were modeled by using the information whether individuals live in the same household. The power of this variance component approach of partition genetic and environmental effects stems from the high information content in the case of the extended pedigrees where families cut across multiple household [28]. Because the pedigrees were ascertained through a thrombophilic proband, all analyses included an ascertainment correction to allow unbiased estimation of parameters relevant to the general population [29,30].

Trait-specific covariates in the variance component models were evaluated from the following list of candidate covariates: age, sex, physical activity, oral contraceptives and smoking. The regression coefficient for the continuous covariate age represents the effect associated with a 1-year deviation from the mean age. The covariate age^2 captured the non-linear relationship between the trait and age. The regression coefficients for the discrete covariates (female sex, physical activity, oral contraceptives and smoking) represent the effect of the covariate versus its absence. P-values of less than 0.05 were considered statistically significant.

The disease status of VTE is recorded as a binary trait encoding affected or unaffected status. In the statistical model, it is transformed to a continuous trait referred to as liability or susceptibility (risk) to the disease by means of probit link function. An interpretation of the effect of a continuous covariate like age in our model with probit link function follows a rule: the negative sign of the coefficient, for example for age covariate, means positive effect in the model (increase liability to the disease).

The correlations between a given pair of traits were analyzed by multivariate variance component models, which are an extension of the univariate model [27]. Similarly to the univariate model for estimation of the heritability of a single trait, the bivariate model partitioned the phenotypic covariances between traits into genetic and environmental components. The derived parameter of genetic correlation coefficient quantifies the pleiotropic genetic effects (i.e., one gene may have effects on several traits). This partition is potentially valuable, since hidden relationships between traits can be revealed [31]. By studying these traits in extended families, we can estimate robustly both the genetic (ρ_g), and the environmental (ρ_e) correlations between traits. The phenotypic correlation (ρ_p) can be derived from these two constituent correlations and the heritabilities of the traits as follows:

$$\rho_p = \sqrt{(h^2{}_1 h^2{}_2)}\,\rho_g + \sqrt{(1 - h^2{}_1)}\,\sqrt{(1 - h^2{}_2)}\,\rho_e.$$

where $h^2{}_1$ and $h^2{}_2$ are the heritabilities for trait one and trait two.

For the association analysis with the imputed genotypes it was applied two excluding filters based on imputation score (info <0.3) and MAF (<1%). The final number of SNPs was 9,303,497. Analyses were performed using the measured genotype method by testing for genotype-specific differences in the means of traits while allowing for the nonindependence among family members. Genome wide significance was defined by p-values $<5\text{x}10^{-8}$, and suggestive significance was defined by p-values $<1\text{x}10^{-5}$. The association analysis was adjusted for *F2* G20210A mutation as known genetic determinant of thrombin generation, if needed.

Table 2. Significant covariates affecting the risk of VTE and thrombin generation.

Trait	Mean	SE (Mean)	Covariate	β	SE (β)	p-value	Variance due to covariates
VTE	NA	NA	AGE	-0.023*	0.004	$6.28x10^{-12}$	NA
LT	1.02	0.01	AGE	0.001	0.0003	$1.39x10^{-03}$	0.04
			SEX	-0.05	0.01	$7.77x10^{-04}$	
			OC	-0.11	0.04	$2.04x10^{-03}$	
TP	32.09	0.06	AGE	0.01	0.001	$1.67x10^{-06}$	0.06
			AGE^2	-0.0002	0.00005	$3.72x10^{-05}$	
			SEX	0.13	0.05	$1.46x10^{-02}$	
			OC	0.72	0.13	$8.16x10^{-08}$	
ETP	48.53	0.06	AGE	0.01	0.001	$1.70x10^{-18}$	0.10
			AGE^2	-0.0003	0.00005	$7.53x10^{-10}$	
			OC	0.87	0.14	$9.81x10^{-10}$	

Mean and SE (Mean) indicate mean of the phenotype value and standard error; β, SE (β) and p-value: regression coefficient, standard error and p-value of regression coefficient; VTE: venous thromboembolism; NA: not applicable; OC: oral contraceptives; LT: Lag Time; TP: Thrombin Peak; and ETP: endogenous thrombin potential.

* The effect of AGE covariate to the risk of VTE is positive, as the negative sign of the coefficient means positive effect in this model, where the VTE response (binary) variable was transformed by probit link function.

doi:10.1371/journal.pone.0146922.t002

All statistical analyses were performed employing the computer package Sequential Oligogenic Linkage Analysis Routines (SOLAR, version 8, official) [28]. SOLAR employs the maximum likelihood approach for variance component models with the standard likelihood ratio tests (LRT) to evaluate the statistical significance of the model parameters [32].

Results

Effect of covariates and heritabilities

The effects of significant covariates of the risk of VTE and thrombin generation phenotypes are shown in Table 2. The covariate age was significantly and positively related to the risk of VTE as well as to the lag time, thrombin peak and ETP. Furthermore, the results of the analysis of the covariate sex showed that women, in comparison with men, have significantly shorter lag times and greater thrombin peak values. Finally, use of oral contraceptives were related to a decreased lag time and associated with an increased thrombin peak and ETP. In contrast, smoking and physical activity were not related to the risk of VTE nor to thrombin generation phenotypes. Interestingly, significant covariates explained only from 4% to 10% of the variability of these traits. Residual kurtosis was within normal range.

The analysis of the contribution of genetics to the variability of each trait (trait heritability = h^2) was performed after the correction by its significant covariates. It is notable that genetic factors accounted for 67% of the variation in the risk of VTE. Furthermore, the heritabilities of lag time, thrombin peak and ETP were estimated as 49%, 54% and 52%, respectively. The estimates of additive genetic effects and the household effect on the variability of the traits are shown in Table 3.

Phenotypic, genetic and environmental correlations of thrombin generation phenotypes with the risk of VTE

The lag time was not phenotypically correlated with VTE. In contrast, the thrombin peak and ETP showed significant and positive phenotypic correlations with VTE (Table 4).

Table 3. Heritabilities and household effect.

Trait	h^2	SE (h^2)	p-value (h^2)	c^2	SE (c^2)	p-value (c^2)
VTE	0.67	0.17	$1.60x10^{-06}$	-*	-	-
LT	0.49	0.07	$3.32x10^{-15}$	0.21	0.05	$2.80x10^{-06}$
TP	0.54	0.07	$3.14x10^{-16}$	0.27	0.05	$9.66x10^{-10}$
ETP	0.52	0.06	$5.71x10^{-18}$	0.23	0.05	$2.27x10^{-08}$

h^2, SE (h^2) and p-value (h^2) indicate heritability, standard error and p-value of the heritability; c^2, SE (c^2) and p-value (c^2): household effect, standard error and p-value of the household effect; VTE: venous thromboembolism; LT: Lag Time; TP: Thrombin Peak; and ETP: endogenous thrombin potential. The covariates used in each model were reported in Table 2.

*The household effect (c^2) was removed from the model of VTE trait, as its estimation was 0.

doi:10.1371/journal.pone.0146922.t003

Table 4. Phenotypic, genetic and environmental correlations of thrombin generation phenotypes with the risk of VTE.

Trait 1	Trait 2	ρ_p	p-value (ρ_p)	ρ_g	SE (ρ_g)	p-value (ρ_g)	ρ_e	SE (ρ_e)	p-value (ρ_e)
VTE	LT	0.05	$3.96x10^{-01}$	0.21	0.22	$3.53x10^{-01}$	-0.04	0.13	$7.50x10^{-01}$
VTE	TP	0.16	$1.10x10^{-02}$	0.47	0.25	$3.21x10^{-02}$	-0.05	0.15	$7.25x10^{-01}$
VTE	ETP	0.20	$1.11x10^{-03}$	0.50	0.28	$3.31x10^{-02}$	0.04	0.14	$7.96x10^{-01}$

ρ_p and p-value (ρ_p) indicate phenotypic correlation and p-value of phenotypic correlation; ρ_g, SE (ρ_g) and p-value (ρ_g): genetic correlation, standard error and p-value of genetic correlation; ρ_e, SE (ρ_e) and p-value (ρ_e): environmental correlation, standard error and p-value of environmental correlation; VTE: venous thromboembolism; LT: Lag Time; TP: Thrombin Peak; and ETP: endogenous thrombin potential.

doi:10.1371/journal.pone.0146922.t004

The lag time did not show significant genetic or environmental correlations. In contrast, and most relevant, the thrombin peak and the ETP showed significant genetic correlations with the risk of VTE, but we did not find environmental correlations (Table 4).

Genetic correlations among thrombin generation phenotypes

The lag time showed significant and negative genetic correlations with the thrombin peak and ETP, but they were not correlated environmentally. Otherwise, the significant genetic and environmental correlations between the thrombin peak and ETP were positive (Table 5).

GWAS

No inflation of the test statistic was observed ($\lambda = 1.05$, $\lambda = 1.03$, $\lambda = 1$ and $\lambda = 0.99$ for the risk of VTE, lag time, thrombin peak and ETP, respectively). No variants showed genome wide significance for association with the risk of VTE (data not shown). Manhattan plots of the results

Table 5. Genetic and environmental correlations among thrombin generation phenotypes.

Trait 1	Trait 2	ρ_p	p-value (ρ_p)	ρ_g	SE (ρ_g)	p-value (ρ_g)	ρ_e	SE (ρ_e)	p-value (ρ_e)
TP	LT	-0.40	$3.39x10^{-28}$	-0.50	0.06	$2.21x10^{-11}$	-0.19	0.10	$6.74x10^{-02}$
ETP	LT	-0.14	$2.10x10^{-04}$	-0.23	0.08	$5.38x10^{-03}$	0.02	0.09	$8.07x10^{-01}$
ETP	TP	0.85	$3.23x10^{-48}$	0.87	0.02	$1.52x10^{-37}$	0.82	0.03	$2.14x10^{-12}$

ρ_p and p-value (ρ_p) indicate phenotypic correlation and p-value of phenotypic correlation; ρ_g, SE (ρ_g) and p-value (ρ_g): genetic correlation, standard error and p-value of genetic correlation; ρ_e, SE (ρ_e) and p-value (ρ_e): environmental correlation, standard error and p-value of environmental correlation; LT: Lag Time; TP: Thrombin Peak; and ETP: endogenous thrombin potential.

doi:10.1371/journal.pone.0146922.t005

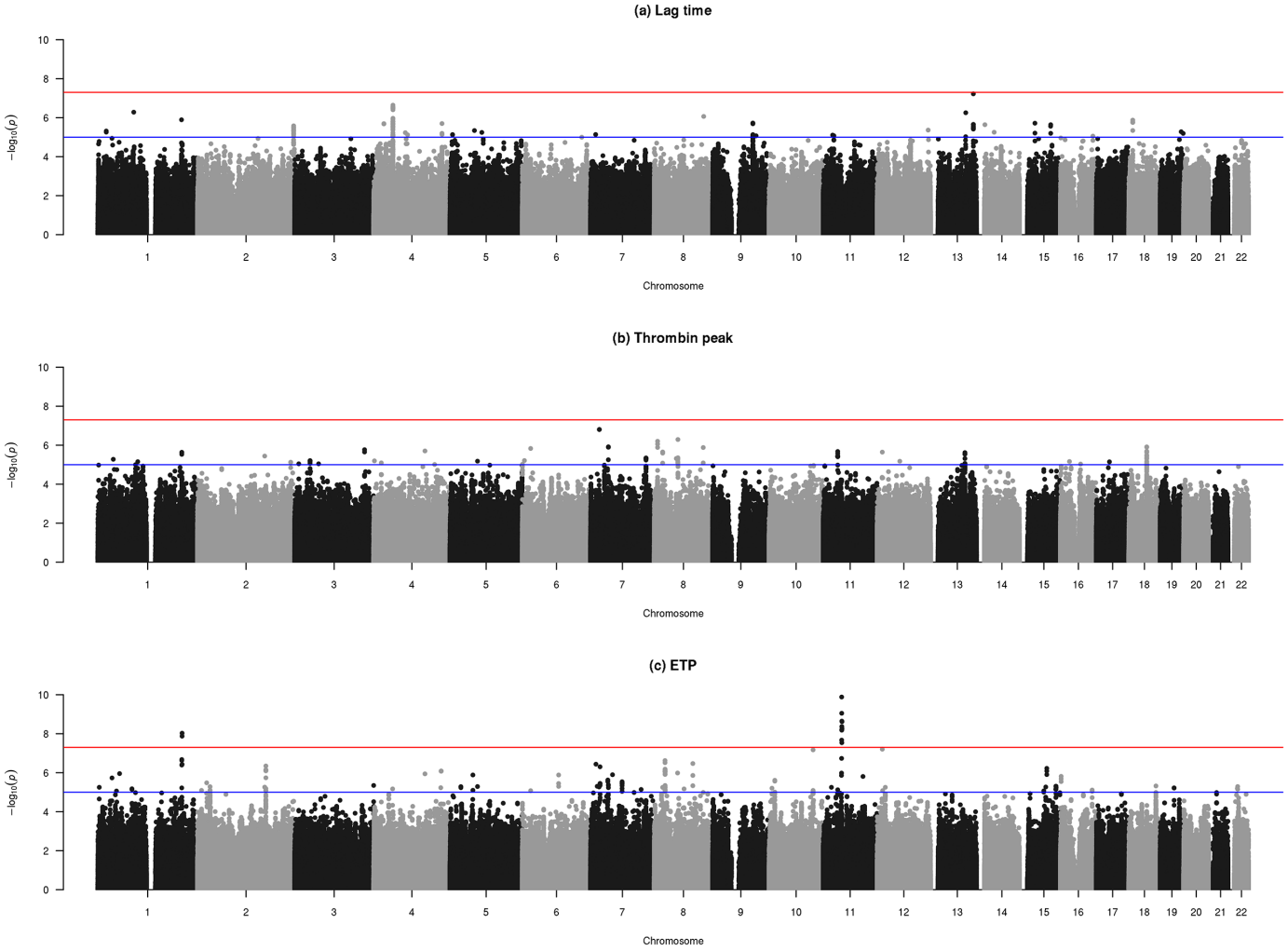

Fig 1. Manhattan plots of the genome wide association studies on the 3 thrombin generation phenotypes. Lag time (a), thrombin peak (b), and ETP (c). Dots correspond to SNPs organized by chromosomal order and position and the *y* axis shows the statistical significance expressed as $-\log_{10}$ of the p-values. The horizontal lines correspond to genome wide significant threshold taken at 5×10^{-8} and genome wide suggestive significance threshold at 1×10^{-5}.

doi:10.1371/journal.pone.0146922.g001

from genome wide associations on the thrombin generation traits are shown in Fig 1. Briefly, no variant attained genome wide significance for association with lag time or thrombin peak. In contrast, 2 peaks on Chromosomes 1 and 11 reached genome wide significance level (p-value = $5\text{x}10^{-08}$) in association with the ETP trait. The significant SNPs rs61828128 (MAF = 2.18% and p-value = $1.33\text{x}10^{-08}$) and rs61828133 (MAF = 2.18% and p-value = $9.36\text{x}10^{-09}$) on Chromosome 1 are located in the *HHAT* gene. Interestingly, the SNP rs1799963 (MAF = 2.82% and p-value = $2.25\text{x}10^{-09}$) in *F2* gene is included in the association peak at chromosome 11 region. After adjustment for G20210A (rs1799963) mutation in the *F2* gene, the signals did not remain significant. However, our results suggested associations involved in both thrombin peak and ETP traits in or near *IRF6* (rs75594643; MAF = 4.31%, p-value = $6.13\text{x}10^{-06}$ for thrombin peak, p-value = $2.94\text{x}10^{-06}$ for ETP, and rs1474608; MAF = 3.80%, p-value = $4.66\text{x}10^{-06}$ for thrombin peak, p-value = $2.46\text{x}10^{-06}$ for ETP), *OCLN* (rs76696742; MAF = 1.13%, p-value = $5.19\text{x}10^{-06}$ for thrombin peak, p-value = $6.59\text{x}10^{-06}$ for ETP), *CDKAL1* (rs16884308; MAF = 1.16%, p-value = $2.57\text{x}10^{-06}$ for thrombin peak, p-

value = 1.47x10^{-06} for ETP), *CAMK2B* (rs180694332; MAF = 1.72%, p-value = 6.55x10^{-06} for thrombin peak, p-value = 5.58x10^{-06} for ETP), and *NUDCD3* (rs144256107; MAF = 1.99%, p-value = 9.80x10^{-06} for thrombin peak, p-value = 1.26x10^{-06} for ETP) genes. We defined as "near" a distance between 1 and 1,500 bp upstream.

Discussion

To investigate the risk of diseases, the study of complex intermediate phenotypes provide a useful means to identify genetic risk factors. The intermediate phenotypes are closer to the action of genes than the presence or absence of a complex disease. In addition, the susceptibility to a complex disease is primarily a process that represents an unobservable continuous liability. This means that this variable can not be measured directly in an individual, and consequently the use of intermediate phenotypes is statistically more powerful [33]. Our study demonstrates that the thrombin generation assay is a useful tool for the study of the risk of VTE.

Using a variance component method and maximum likelihood estimations age was found to be the covariate with statistically significant effect on the risk of VTE. This is not surprising since the covariate age has been associated previously with VTE [1]. Except for smoking and physical activity, all covariates, including age, sex and oral contraceptives, were significantly related to the parameters of thrombin generation assay. Specifically, age, sex and oral contraceptives influenced the lag time. The thrombin peak was significantly related with age, age^2, sex and oral contraceptives and the ETP was influenced by age, age^2 and oral contraceptives. These results are consistent with previous reports under different experimental conditions. Specifically, it has been reported that thrombin generation increases with age and is higher in women than in men [34]. In addition, oral contraceptives have been related also with an increased thrombin generation [11,35].

Our study was based on the recruitment of families, which provided direct data of familial transmission and allowed an estimation of genetic factors affecting the variation in the risk of VTE within a Spanish population [3]. It is notable that, in our population, genetic factors accounted for 67% of the variation in the risk of VTE. This high heritability is similar to what we reported previously in the GAIT-1 Project study [3]. The heritabilities of thrombin generation phenotypes in this new set of families ranged from 49% to 54%. These findings are consistent with previous published data [12]. These high heritabilities indicate that genes play a major role in the determination of the variability of these parameters.

One of our most important results is that they add to the previous evidence of the relationship between the risk of VTE and thrombin generation. For the first time to our knowledge, we determined genetic correlations between the susceptibility to VTE and thrombin peak or ETP. Not surprisingly, we did not find phenotypic, genetic nor environmental correlations between the risk of VTE and lag time. Our results agree with previous publications which reported no association between the lag time and the risk of first or recurrent VTE [11]. In contrast, thrombin peak and ETP showed positive phenotypic correlations with the risk of VTE which are similar to previous evidence. Interestingly, our study is the first that shows strong evidence that the positive associations between the risk of VTE and thrombin peak or ETP is caused by pleiotropic factors. These results provide evidence that there are genes acting jointly on both the risk of VTE and thrombin peak or ETP. Therefore, we believe that it is prudent to use the thrombin generation assay as an intermediate phenotype to identify novel genes that affect the risk of VTE.

Taking together, our results support the hypothesis that the thrombin generation test is a very useful test to investigate the risk of VTE. Specifically, thrombin generation phenotypes have been associated significantly with genetic variants in haemostatic genes such as *F5*, *F2*, *FGA*, *F10*, *F12* and *TFPI* [7,36]. It has been reported recently that there is a new association

between thrombin generation variability and the *ORM1* locus [37]. It is important to note that known genetic risk factors have been estimated to account for <30% of VTE cases [38]. Consequently, this global coagulation assay might identify novel genetic variants that contribute to the heritability of the risk of VTE.

In agreement with a previous study [37] we observed a high significant association between thrombin generation and prothrombin G20210A mutation. This genetic variation is a well-known genetic risk factor for VTE [1]. In addition, we observed suggestive signals (presented in at least two out of the three phenotypes studied) that might increase our knowledge to explain the variability of thrombin generation. However, we do not have biological evidences to relate these observed signals with the phenotypes due to the lack of information about functionality in the data bases. Thus, further studies are needed to elucidate the implication of these new genes with thrombin generation and the risk of VTE.

The thrombin generation assay is sensitive to preanalytic conditions and there is no standardized protocol [39,40]. In detail, we used 5 pM tissue factor as trigger in PPP. At this concentration, thrombin generation assay is sensitive to factors VIII and IX. Lower concentrations (1 pM) of tissue factor could lead to an increase of the variability of the assay. It is difficult to compare various studies considering that this assay can be performed under widely different laboratory conditions [41]. Despite this, our results are consistent with those of previous studies, but add important new data.

In summary, the high heritabilities that we found indicate that genetic factors play a significant role in the risk of VTE and in the thrombin generation test in this population. In addition, the significant genetic correlations suggest that there are pleiotropic genetic effects between the risk of VTE and thrombin peak or ETP. Finally, our study indicates that the use of the thrombin generation assay should help to detect genes responsible for thrombophilia. In fact, from results of the GWAS, *F2* has been shown as the main contributor to thrombin generation.

Acknowledgments

We are deeply grateful to the families who participated in this study. Also, we would like to thank Professor Bill Stone for reviewing the manuscript. Genotyping was performed by the SNP&SEQ Technology Platform in Uppsala. The platform is part of Science for Life Laboratory at Uppsala University and supported as a national infrastructure by the Swedish Research Council.

Author Contributions

Conceived and designed the experiments: JCS JMS. Performed the experiments: MC JAM NV MF. Analyzed the data: LMF AZ AMP HB. Wrote the paper: LMF AZ JMS. Final approval of the version to be published: LMF AZ JCS JMS AH. Search of funding: JCS JMS AH.

References

1. Rosendaal FR. Venous thrombosis: the role of genes, environment, and behavior. Hematology Am Soc Hematol Educ Program. 2005; 2005: 1–12.
2. Rosendaal FR. Venous thrombosis: a multicausal disease. Lancet. 1999; 353: 1167–1773. PMID: 10209995
3. Souto JC, Almasy L, Borrell M, Blanco-Vaca F, Mateo J, Soria JM, et al. Genetic susceptibility to thrombosis and its relationship to physiological risk factors: the GAIT study. Genetic Analysis of Idiopathic Thrombophilia. Am J Hum Genet. 2000; 67: 1452–1459. PMID: 11038326
4. Heit JA, Phelps MA, Ward SA, Slusser JP, Petterson TM, De Andrade M. Familial segregation of venous thromboembolism. J Thromb Haemost. 2004; 2: 731–736. PMID: 15099278

5. Soria JM, Almasy L, Souto JC, Bacq D, Buil A, Faure A, et al. A quantitative-trait locus in the human factor XII gene influences both plasma factor XII levels and susceptibility to thrombotic disease. Am J Hum Genet. 2002; 70: 567–574. PMID: 11805911
6. Buil A, Trégouët D-A, Souto JC, Saut N, Germain M, Rotival M, et al. C4BPB/C4BPA is a new susceptibility locus for venous thrombosis with unknown protein S-independent mechanism: results from genome-wide association and gene expression analyses followed by case-control studies. Blood. 2010; 115: 4644–4650. doi: 10.1182/blood-2010-01-263038 PMID: 20212171
7. Segers O, van Oerle R van, ten Cate H ten, Rosing J, Castoldi E. Thrombin generation as an intermediate phenotype for venous thrombosis. Thromb Haemost. 2010; 103: 114–122. doi: 10.1160/TH09-06-0356 PMID: 20062924
8. Hemker HC, Wielders S, Kessels H, Béguin S. Continuous registration of thrombin generation in plasma, its use for the determination of the thrombin potential. Thromb Haemost. 1993; 70: 617–624. PMID: 7509511
9. Hemker HC, Béguin S. Phenotyping the clotting system. Thromb Haemost. 2000; 84: 747–751. PMID: 11127849
10. Dargaud Y, Trzeciak MC, Bordet JC, Ninet J, Negrier C. Use of calibrated automated thrombinography +/- thrombomodulin to recognise the prothrombotic phenotype. Thromb Haemost. 2006; 96: 562–567. PMID: 17080211
11. Van Hylckama Vlieg A, Christiansen SC, Luddington R, Cannegieter SC, Rosendaal FR, Baglin TP. Elevated endogenous thrombin potential is associated with an increased risk of a first deep venous thrombosis but not with the risk of recurrence. Br J Haematol. 2007; 138: 769–774. PMID: 17760809
12. Wichers IM, Tanck MWT, Meijers JCM, Lisman T, Reitsma PH, Rosendaal FR, et al. Assessment of coagulation and fibrinolysis in families with unexplained thrombophilia. Thromb Haemost. 2009; 101: 465–470. PMID: 19277406
13. Lutsey PL, Folsom AR, Heckbert SR, Cushman M. Peak thrombin generation and subsequent venous thromboembolism: the Longitudinal Investigation of Thromboembolism Etiology (LITE) study. J Thromb Haemost. 2009; 7: 1639–1648. doi: 10.1111/j.1538-7836.2009.03561.x PMID: 19656279
14. Hron G, Kollars M, Binder BR, Eichinger S, Kyrle PA. Identification of patients at low risk for recurrent venous thromboembolism by measuring thrombin generation. JAMA. 2006; 296: 397–402. PMID: 16868297
15. Besser M, Baglin C, Luddington R, van Hylckama Vlieg A, Baglin T. High rate of unprovoked recurrent venous thrombosis is associated with high thrombin-generating potential in a prospective cohort study. J Thromb Haemost. 2008; 6: 1720–1725. doi: 10.1111/j.1538-7836.2008.03117.x PMID: 18680535
16. Tripodi A, Legnani C, Chantarangkul V, Cosmi B, Palareti G, Mannucci PM. High thrombin generation measured in the presence of thrombomodulin is associated with an increased risk of recurrent venous thromboembolism. J Thromb Haemost. 2008; 6: 1327–1333. doi: 10.1111/j.1538-7836.2008.03018.x PMID: 18485081
17. Faber CG, Lodder J, Kessels F, Troost J. Thrombin generation in platelet-rich plasma as a tool for the detection of hypercoagulability in young stroke patients. Pathophysiol Haemost Thromb. 2003; 33: 52–58. PMID: 12853713
18. Carcaillon L, Alhenc-Gelas M, Bejot Y, Spaft C, Ducimetière P, Ritchie K, et al. Increased thrombin generation is associated with acute ischemic stroke but not with coronary heart disease in the elderly: the Three-City cohort study. Arterioscler Thromb Vasc Biol. 2011; 31: 1445–1451. doi: 10.1161/ATVBAHA.111.223453 PMID: 21454811
19. Orbe J, Zudaire M, Serrano R, Coma-Canella I, Martínez de Sizarrondo S, Rodríguez JA, et al. Increased thrombin generation after acute versus chronic coronary disease as assessed by the thrombin generation test. Thromb Haemost. 2008; 99: 382–387. doi: 10.1160/TH07-07-0443 PMID: 18278189
20. Souto JC, Almasy L, Borrell M, Garí M, Martínez E, Mateo J, et al. Genetic determinants of hemostasis phenotypes in Spanish families. Circulation. 2000; 101: 1546–1551. PMID: 10747348
21. International Physical Activity Questionnaire (IPAQ) [Internet]. [cited 30 Mar 2014]. Available: www.ipaq.ki.se
22. Miller SA, Dykes DD, Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res. 1988; 16: 1215. PMID: 3344216
23. Hemker HC, Giesen P, Al Dieri R, Regnault V, de Smedt E, Wagenvoord R, et al. Calibrated automated thrombin generation measurement in clotting plasma. Pathophysiol Haemost Thromb. 2003; 33: 4–15. PMID: 12853707

24. Delaneau O, Zagury J-F, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. Nat Methods. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2013; 10: 5–6.

25. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009; 5: e1000529. doi: 10.1371/journal.pgen.1000529 PMID: 19543373

26. Souto JC, Almasy L, Blangero J, Stone W, Borrell M, Urrutia T, et al. Genetic regulation of plasma levels of vitamin K-dependent proteins involved in hemostasis: results from the GAIT Project. Genetic Analysis of Idiopathic Thrombophilia. Thromb Haemost. 2001; 85: 88–92. PMID: 11204594

27. Lynch M, Walsh B. Genetics and analysis of quantitative traits. Sunderland, Mass: Sinauer Associates Inc; 1998.

28. Almasy L, Blangero J. Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet. 1998; 62: 1198–1211. PMID: 9545414

29. Hopper JL, Mathews JD. Extensions to multivariate normal models for pedigree analysis. Ann Hum Genet. 1982; 46: 373–383. PMID: 6961886

30. Boehnke M, Lange K. Ascertainment and goodness of fit of variance component models for pedigree data. Prog Clin Biol Res. 1984; 147: 173–192. PMID: 6547532

31. Comuzzie AG, Blangero J, Mahaney MC, Haffner SM, Mitchell BD, Stern MP, et al. Genetic and environmental correlations among hormone levels and measures of body fat accumulation and topography. J Clin Endocrinol Metab. 1996; 81: 597–600. PMID: 8636274

32. Self S, Liang K. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J Am Stat Assoc. 1987; 82: 605–610.

33. Souto JC. Search for new thrombosis-related genes through intermediate phenotypes. Genetic and household effects. Pathophysiol Haemost Thromb. 2002; 32: 338–340. PMID: 13679669

34. Dielis AWJH, Castoldi E, Spronk HMH, van Oerle R, Hamulyák K, Ten Cate H, et al. Coagulation factors and the protein C system as determinants of thrombin generation in a normal population. J Thromb Haemost. 2008; 6: 125–131. PMID: 17988231

35. Tchaikovski SN, van Vliet HAAM, Thomassen MCLGD, Bertina RM, Rosendaal FR, Sandset P-M, et al. Effect of oral contraceptives on thrombin generation measured via calibrated automated thrombography. Thromb Haemost. 2007; 98: 1350–1356. PMID: 18064335

36. Kyrle PA, Mannhalter C, Béguin S, Stümpflen A, Hirschl M, Weltermann A, et al. Clinical studies and thrombin generation in patients homozygous or heterozygous for the G20210A mutation in the prothrombin gene. Arterioscler Thromb Vasc Biol. 1998; 18: 1287–1291. PMID: 9714136

37. Rocanin-Arjo A, Cohen W, Carcaillon L, Frère C, Saut N, Letenneur L, et al. A meta-analysis of genome-wide association studies identifies ORM1 as a novel gene controlling thrombin generation potential. Blood. 2014; 123: 777–785. doi: 10.1182/blood-2013-10-529628 PMID: 24357727

38. Morange P-E, Tregouet D-A. Deciphering the molecular basis of venous thromboembolism: where are we and where should we go? Br J Haematol. 2010; 148: 495–506. doi: 10.1111/j.1365-2141.2009.07975.x PMID: 19912223

39. Loeffen R, Kleinegris M-CF, Loubele STBG, Pluijmen PHM, Fens D, van Oerle R, et al. Preanalytic variables of thrombin generation: towards a standard procedure and validation of the method. J Thromb Haemost. 2012; 10: 2544–2554. doi: 10.1111/jth.12012 PMID: 23020632

40. Baglin T. The measurement and application of thrombin generation. Br J Haematol. 2005; 130: 653–661. PMID: 16115120

41. Van Veen JJ, Gatt A, Makris M. Thrombin generation testing in routine clinical practice: are we there yet? Br J Haematol. 2008; 142: 889–903. doi: 10.1111/j.1365-2141.2008.07267.x PMID: 18564356

Artículo 2

Título

Genetics Determinants for Factor VIII Levels: Genome-Wide Linkage and Association Analyses from the GAIT Project.

Autores

Sonia Lopez, Laura Martin-Fernandez, Andrey Ziyatdinov, Angel Martinez-Perez, Ares Rocañín, Giovana Gavidia-Bovadilla, Juan Carlos Souto, y José Manuel Soria.

Referencia

Artículo en preparación, pendiente de enviar a *Journal of Thrombosis and Haemostasis*.

Resumen

Los niveles plasmáticos del FVIII presentan una alta heredabilidad y están asociados al riesgo de VTE. Considerando que los factores genéticos responsables de niveles elevados de FVIII son en gran parte desconocidos, nos hemos planteado como objetivo la realización de un análisis de ligamiento de genoma completo para poder determinar las regiones genómicas responsables de la variabilidad de los niveles del FVIII en las familias del Proyecto GAIT-1. Para esto, se han medido los niveles plasmáticos del FVIII funcional en los 398 individuos de este proyecto, los cuales están agrupados en 21 familias extensas españolas. Como marcadores genéticos se han utilizado 485 microsatélites y se ha empleado el método de los componentes de la variancia en el análisis de ligamiento. Los resultados han mostrado evidencias de ligamiento con los niveles de FVIII en 2 QTLs, uno de los cuales se encuentra en el

cromosoma 2 (puntuación LOD=3,41, p-valor=1,00x10^{-03}) y el otro está localizado en el cromosoma 3 (puntuación LOD=3,90, p-valor=1,00x10^{-03}). De esta manera, los genes *CIB4* (cromosoma 2), *ARHGEF3* y *ADAMTS9* (cromosoma 3) se postulan como principales genes candidatos. A continuación, y mediante un mapeo fino de 2593 SNPs en estas regiones candidatas, se ha determinado que la variante genética rs1276123 en *CIB4* muestra una asociación estadísticamente significativa con los niveles de FVIII. Cabe destacar que esta señal ha sido replicada en el Proyecto GAIT-2, con 935 individuos de 35 nuevas familias extensas. Además, mediante NGS de *CIB4* en individuos no relacionados del Proyecto GAIT-2 se han identificado 7 variantes genéticas comunes adicionales y 2 grupos de variantes genéticas raras y de baja frecuencia alélica asociados a los niveles de FVIII. En este trabajo se ha contribuido al descubrimiento de nuevos factores genéticos implicados en la alta heredabilidad del FVIII que, además, pueden ayudar a esclarecer los mecanismos biológicos ocultos en la variabilidad de los niveles del FVIII y del riesgo de VTE.

Material suplementario

Tablas suplementarias de la publicación (ver Anexo I).

Genetic Determinants for Factor VIII levels: Genome-wide Linkage and Association Analyses from the GAIT Project

Sonia Lopez[1]*, Laura Martin-Fernandez[1], Andrey Ziyatdinov[1], Angel Martinez-Perez[1], Ares Rocañín[1], Giovana Gavidia-Bovadilla [1], Noelia Vilalta, Juan Carlos Souto[2], Jose Manuel Soria[1].

[1]Unit of Genomic of Complex Diseases. Research Institute of Hospital de la Santa Creu i Sant Pau, 08025 Barcelona, Spain

[2]Haemostasis and Thrombosis Unit, Department of Hematology, Hospital de la Santa Creu i Sant Pau, 08025 Barcelona, Spain

***Corresponding author:**

E-mail: slopezm@santpau.cat (SL)

Abstract

Plasma factor VIII (FVIII) levels have been reported to be highly heritable and strongly associated with thrombosis risk. The genetic factors that predispose to elevated FVIII are largely unknown. We performed a genome-wide linkage analysis followed by fine-mapping to delineate the genomic regions that influence the variance in FVIII levels in families from the Genetic Analysis of Idiopathic Thrombophilia (GAIT) Project. Functional clotting FVIII (FVIII:C) was measured in plasma of 398 individuals belonging to 21 Spanish families from GAIT 1. A total of 485 DNA microsatellite markers were typed at a mean distance of 7.1 cM. A variance component linkage method was used to evaluate linkage and to detect quantitative trait loci (QTLs). Two QTLs showed strong evidence of linkage with FVIII levels. One of them was located on Chromosome 2 (LOD = 3.41) and the other one located on Chromosome 3 (LOD = 3.90). The gene *CIB4* on Chromosome 2 and the genes *ARHGEF3* and *ADAMTS9* on Chromosome 3 were positional candidate genes that could influence FVIII levels. Subsequent fine-mapping with 2593 single nucleotide polymorphisms (SNPs) from the Golden Gate Illumina platform was conducted in the genomic regions containing linkage signals to find out genetic variants associated with the variance in FVIII levels. Fine-mapping in these linkage regions revealed that the rs1276123 on Chromosome 2 was significantly associated with FVIII levels. Replication of this association in an independent sample was performed by using 935 individuals from 35 new extended Spanish families belonging to GAIT 2. Interestingly, this intron-variant is located in the *CIB4* gene, suggesting a functional effect of this polymorphism on FVIII levels. However, no association was found on Chromosome 3, suggesting that a rare variant may contribute to the linkage signal. Also, next generation sequencing (NGS) in unrelated individuals from the GAIT 2 Project was applied to identify additional genetic variants in *CIB4* involved in the variability of FVIII levels. A total of 7 common genetic variants and 2 low frequency variants sets in *CIB4* were associated with FVIII levels. Our results shed light on the genetic factors that contribute to the heritability of FVIII and could help to elucidate the biological mechanisms underlying the variance in FVIII levels and thrombosis risk.

Introduction

Factor VIII (FVIII) is a key protein of the intrinsic pathway of the coagulation cascade. It functions as a cofactor for activated FIX (FIXa), in the presence of calcium and phospholipids, to increase the rate of activation of FX in a dose-dependent manner. Elevated plasma levels of FVIII has been well described as an independent and common risk factor for venous thrombosis (VT) [1,2], recurrent thrombosis [3], coronary heart disease [4,5], ischemic stroke [6] and ischaemic heart disease [7]. FVIII is considered a strong contributor to thrombosis disease as much as deficiencies of coagulation inhibitor proteins (proteins C and S) and activated protein C resistance (APCR). However, the biological mechanism that cause elevated FVIII levels is not yet understood.

Elevated FVIII levels have shown to persist over time and not to be the result of an acute phase reaction [8,9]. Several reports, being most of them twin-based studies, have shown that genetic factors contributed from 57% to 61% to the variation in FVIII antigen levels [10–12]. In agreement with these estimates, our group have previously reported an additive genetic heritability of 40% for FVIII:C levels in families from the Genetic Analysis of Idiopathic Thrombophilia (GAIT) Project [13]. These heritabilities indicate that genetic factors are the most important determinants of this quantitative trait and encourage the search for genes. In addition, a strong genetic correlation between FVIII levels and thrombosis risk has been reported in GAIT by our group [14]. The identification of major genes involved in quantitative variation of FVIII in plasma is important given the high contribution of this protein to thrombosis risk. However, genetic studies performed up to now have not been able to identify them. For example, Kamphuisen et al. suggested in 1998 a control of FVIII variation by X-linked alleles, possible the *F8* gene itself [15]. However, a lack of sequence variations associated with high FVIII levels in the promoter and 3' terminus, as well as in introns 13 and 22 of *F8* was reported in subjects with high plasma FVIII levels and VT [16,17]. Furthermore, a study including families with a proband with VT and carrying the FV Leiden mutation (FVL) failed to detect a linkage signal in *F8* for FVIII levels [18]. Other studies have identified a few polymorphisms and haplotypes in *F8* that have shown association with FVIII levels in different populations, although with little effect on

this clotting protein [19–22]. More recently, copy number variations of *F8* has been shown to influence FVIII activity and the risk of VT [23]. In addition, other genes different of the structural *F8* gene have been associated with the quantitative variation of FVIII. We know that the *ABO* locus is a major determinant that influences both plasma concentration of Von Willebrand factor (vWF) and FVIII, being all of them strongly associated with the risk of thrombosis [2,11,14]. The *ABO* genes influence 30% of the genetic variance of vWF [11] and accounts for up to 20% of the variance in FVIII levels [24]. In particular, levels of vWF and FVIII in plasma are higher in subjects of non-O blood group than in those of group O, thus conferring two-fold increased thrombosis risk [2]. Most of the effects of blood groups on FVIII levels are mediated by vWF, since it is the carrier of FVIII in plasma and protects this procoagulant factor from proteolysis. Thus, FVIII levels depend on vWF availability in plasma. Nevertheless, FVIII has been described as an independent risk factor for deep vein thrombosis (DVT) in a multivariate analysis with individuals from the Leiden Thrombophilia Study (LETS) [2]. Adjusted thrombosis risk for FVIII is dose-dependent and it has been estimated to be 4.8 higher in subjects with FVIII:C ≥150 IU/dL than the risk for subjects with FVIII:C <100 IU/dL. Importantly, high plasma levels of FVIII not only have been associated with the risk of a first event of thrombosis, but also predisposes for the recurrence of the disease [3]. For each 10 IU/dL increment of FVIII, the risk for a single episode of VT increases by 10%, whereas for recurrent disease increases by 24% [1]. In agreement with the association of ABO blood group with both vWF and FVIII levels, our group have found significant linkage between the *ABO* locus and vWF levels and a suggestive linkage signal with FVIII levels in GAIT [25]. In addition to ABO blood group and vWF, an acquired APCR phenotype has also been related to high levels of FVIII and increased thrombosis risk in the absence of the FVL [26]. Therefore, high FVIII levels could enhance susceptibility to thrombosis either by increasing the rate of thrombin formation or by an effect on APCR, due to a decreased efficiency in FVIII inactivation. However, the risk of thrombosis due to elevated FVIII levels also has been found to be independent on APCR [27]. All these data suggests that high levels of FVIII may be controlled by additional genetic factors other than *ABO*, *VWF* and *F8*. Accordingly, we have reported a QTL on Chromosome 18 with a pleiotropic effect on APCR and FVIII levels, as well as on susceptibility to thrombosis in GAIT by using bivariate

linkage analyses [28]. This means that common genes may influence both phenotypes and contribute to thrombosis risk. Also, a suggestive locus on Chromosome 8 and two imprinted QTLs on Chromosomes 5 and 11 for FVIII levels were detected in thrombophilic families with elevated FVIII levels [29]. However, these loci deserve further exploratory interventions to assign them a functional role. In general, a small biological effect with a low variation in the susceptibility to thrombosis has been observed from these studies.

Regardless of the effort to identify other genes with a major effect on FVIII concentration, until date little is known about the mechanism underlying the variability of this clotting protein. All the above data support that other still-unknown genetic factors might influence the quantitative variation of FVIII in plasma, so new research on genes implicated in the control of this hemostasis-related quantitative trait is necessary. In addition, there is a lack of studies including extended pedigrees, which are more suitable to investigate the genetic architecture of a complex trait.

The major aim of the present study was to identify genes that influence plasma FVIII levels. For this purpose, we conducted a genome-wide variance component linkage analysis using data from the GAIT Project. We fine-mapped these linkage regions to identify common variants that could influence the phenotype and we performed also next generation sequencing (NGS) of the top candidate gene to explore the genetic spectrum related to FVIII levels. To our knowledge, this is the first family-based genome-wide scan combining linkage and association analyses to look for genomic regions that influence plasma levels of FVIII.

Material and Methods

Study Subjects

The present study included 398 individuals from 21 extended Spanish families (≥ 10 living individuals in ≥ 3 generations) belonging to the GAIT Project (GAIT 1) [13]. Among the total pedigreed families, 12 of them were selected through a proband with idiopathic thrombophilia and 9 families were obtained randomly from the general population. Idiopathic thrombophilia was defined as recurrent thrombosis (≥ 1 spontaneous), a single spontaneous thrombotic event with a first-degree relative also affected, or early-onset thrombosis (< 45 years), with all known

biological causes of thrombophilia excluded at inclusion (1995-1997). Accordingly, families with deficiencies of coagulation proteins or cofactors including protein S, protein C, antithrombin, plasminogen, factor V Leiden, heparin cofactor II, APCR, dysfibrogenemia, lupus anticoagulant, and antiphospholipid antibodies were not accepted for the study. Detailed composition of the families and the collection of lifestyle, medical, and family history data have been previously described [13].

In addition, we included in this study a new set of extended families belonging to the GAIT Project (GAIT 2) to replicate and validate our major results in an independent sample. This sample consisted of 935 individuals from 35 Spanish families recruited since 2006 to 2010 [30]. All of the individuals were selected through a proband with idiopathic thrombophilia according to the same inclusion criteria used for the GAIT 1 study [13].

A total of 105 individuals not related by blood from the GAIT 2 Project were selected for NGS analysis. Of them, we chose as a discovery sample 22 individuals to have low plasma FVIII levels (47-94 IU/dL) and 19 individuals to have high plasma FVIII levels (229-450 IU/dL) compared to the normal distribution. This subset of 41 individuals contained an approximately equal number of males (N=22) and females (N=19). Thus, variants identified by the NGS in the discovery sample were denoted as “variants of interest” and the genotypes of these “variants of interest” of individuals from the whole sample of 105 individuals were used for association analyses to increase its statistical power.

Ethics Statement

All procedures were approved by the Institutional Review Board of the Hospital de la Santa Creu i Sant Pau (Barcelona). Adult subjects gave informed consent for themselves and for their minor children.

Laboratory Studies and Phenotyping

Thrombophilic participants were not using oral anticoagulants at the time of sampling. To minimalize the influence of the acute phase in FVIII levels, blood was obtained at least 6 months after the thrombotic event in affected individuals.

Blood was collected by venipuncture in 1/10 volume of 0.129 mol/l sodium citrate from fasting subjects of the GAIT Project. After centrifugation at 2000 g during 20 minutes, platelet-poor plasma (PPP) was obtained and used for the biochemical assays that require fresh plasma sample. The remaining plasma was stored at −80ºC until use. The remaining fraction of blood cells was washed (1:1) with sodium chloride 0.9% and centrifuged 10 minutes at 1580 g to obtain the buffy coat, which was stored at -20 ºC until DNA extraction.

FVIII procoagulant activity (FVIII:C) was measured in fresh PPP samples of all of the GAIT participants by using the automated coagulometer STA-R Evolution® Expert Series (Diagnostica Stago, Asnières sur Seine, France) and the kit STA®-Deficient VIII (Diagnostica Stago, Asnières sur Seine, France) as previously described [25]. FVIII:C levels were expressed in international units per decilitre (IU/dL).

Total DNA from the buffy coats of the GAIT samples was extracted by a standard salting out procedure [31] and it was used for subsequent genotyping of all of the subjects.

Genotyping with Microsatellites and Single Nucleotide Polymorphisms (SNPs)

All of the subjects of GAIT 1 were genotyped for a genome-wide scan including 485 DNA microsatellite markers, distributed through the autosomal genome at a density of 7.1 cM and 0.79 of average heterozygosity. Microsatellites consisted primarily of the ABI Prism Linkage Mapping Set MD-10 (Applied Biosystems, Foster City, CA). Linkage mapping was undertaken with the PE LMS II fluorescent marker set (ABI Prism, Foster City, CA) with multiplex polymerase chain reaction (PCR) [32]. The PCR products were analysed on the PE 310, PE 377, and PE 3700 automated sequencers, and genotyped using the Genotyper software. Information on microsatellite markers can be found in the public-accessible genomic database (http://www.gdb.org). Marker maps for multipoint analyses were obtained from the Marshfield Medical Research Organization (http://research.marshfieldclinic.org/genetics/).

Subsequent genotyping in GAIT 1 with a set of 2593 SNPs from the Golden Gate Genotyping Assay (Illumina, San Diego, CA, USA) was conducted to fine-map the genomic regions delineated by the linkage signals to find out common variants associated with the variance in

plasma FVIII levels. Then, allelic frequencies for each polymorphism were estimated from the GAIT sample. The SNPs with a genotype call rate of <0.95, a minor allele frequency of <0.025 or failing the Hardy-Weinberg equilibrium (HWE) test ($p < 1e-5$) were excluded from the analysis. HWE was ascertained by a standard χ^2 with 1 degree of freedom and it was tested using parental data only. Thus, 2579 SNPs were finally available for genotyping. Genotyping in GAIT 2 of the SNPs that showed significant association with FVIII levels in GAIT 1 was carried out by using individual TaqMan® SNP Genotyping Assays (Applied Biosystems, Life Technologies, Foster City, CA).

Consistency of the genetic data with Mendelian inheritance, for both the genome scan and the fine-mapping genotypes, was checked using the program INFER (PEDSYS, San Antonio, TX, USA) [33]. When the inconsistency could not be resolved, discrepant individuals were either corrected or excluded from the analyses.

Linkage and Association Analyses

A standard multipoint variance component linkage method was used to assess linkage between autosomal markers and plasma FVIII levels in GAIT 1 using the Sequential Oligogenic Linkage Analysis Routines (SOLAR) v. 6.6.2 software package [34]. The covariates age, gender, ABO blood group, vWF, APCR, smoking behaviour, oral contraceptives use and the household effect were included in the variance components framework to identify those one with a significant effect on FVIII levels. Only the covariates with a significant effect were considered in the analyses.

Given the great influence that the genes coding the ABO blood group proteins exert on plasma FVIII levels, all of the subjects were classified according to the ABO blood group genotypes. In particular, the O allele of the ABO blood group was taken as reference, since it has been associated with low levels of FVIII and low risk of thrombosis. Thus, we considered an additive genetic model of allelic effect in which all of the subjects were grouped in three categories according to the presence of the O allele in the ABO blood group (no O allele, one O allele or two O alleles). As most of the families were ascertained through a thrombophilic proband, all analyses included an ascertainment correction to allow unbiased estimation of parameters

relevant to the general population. To achieve this, the likelihood for each family ascertained through a thrombophilic proband was conditioned on the likelihood of their respective proband [35].

Quantitative trait association between the SNPs and FVIII plasma levels was performed with SOLAR using a measured genotype association analysis assuming an additive model of allelic effect [36]. The *P*-values for each SNP association test were corrected for multiple comparisons by the generally more stringent Bonferroni adjustment, which establishes a significance threshold corresponding to a family-wise error rate of 0.05. SNP associations that were ranked within the 10 most significant hits, but did not survive the Bonferroni correction were considered to be suggestive.

NGS

PCR Primer Design and PCR Amplification

A total of 8 long range (LR) PCRs were used to amplify 63,264 bp of *CIB4* (GRCh37/hg19 chr2:26,804,073-26,864,211). Primers were designed within the intronic region or outside the gene and primer sequences, primer positions and LR PCR amplicons size are described in S1 Table. The overlapped LR PCR amplicons covered exons, introns, 5'-UTR, 3'UTR and, approximately, 1,500 bp of the 5' flanking region (promoter region). The PCR amplicons designed were tested for target specificity by Sanger sequencing [37] of both strands.

LR PCR amplifications were performed with the SequalPrep Long PCR Kit with dNTPs (Invitrogen, Thermo Fisher Scientific Inc., MA, USA). The LR PCR mix solution contained ~50 ng of DNA, SequalPrep 1X reaction buffer, 0.4 µl of dimethylsulfoxide (DMSO), SequalPrep 1X enhancer B, 0.75 µM of forward and reverse primers and 1.8 units of SequalPrep Long Polymerase in a total volume of 20 µl. After initial denaturation at 94ºC for 2 minutes (min), 10 cycles of 94ºC for 10 seconds (sec), 64ºC for 30 sec (in all PCR except for CIB4 LR6: 60ºC), and 68ºC for 18 min were performed, followed by 22 cycles of 94ºC for 10 sec, 64ºC for 30 sec (in all PCR except for CIB4 LR6: 60ºC), and 68ºC for 18 min (+ 20 sec/cycle). In addition, an elongation step at 72ºC for 5 min and a final cooling at 4°C were performed.

LR PCR amplicons of the 105 individuals were analysed on 0.7% agarose gel electrophoresis and visualized by SYBR safe (Invitrogen, Thermo Fisher Scientific Inc.) staining. Moreover, LR

PCR amplicons were quantified by using the Qubit technology (Invitrogen, Thermo Fisher Scientific Inc.) to prepare a normalized pool of the 8 LR PCR amplicons for each individual by combining equimolar amounts. The Qubit technology was used also to adjust the 105 PCR pools at 0.2 ng/µl for library preparation.

Library Preparation, Sequencing and Data Analysis

The sequencing libraries were prepared from PCR pools using the Nextera XT DNA Sample Preparation kit (Illumina, San Diego, CA, USA) with double indexing, according to the manufacturer's protocol. Paired-end sequencing was used to improve the mapping quality [38]. We obtained 105 paired-end libraries that were pooled and ran simultaneously on an Illumina Miseq sequencing system (Illumina) using the Miseq V2 300 cycle run kit (2x150 bp) (Illumina). Indexed sequences were de-multiplexed and analyzed individually. Paired sequence files in fastq files format were used for the analysis with CLC Genomic Workbench version 6.5 software (CLC bio - Qiagen, Aarhus, Denmark). Raw data was trimmed with length (minimum, 25 bp; maximum, 500 bp), ambiguous nucleotide (maximum, 2) and quality score (0.05) filters. This software permits the alignment of the trimmed reads against the human genome sequence (hg19) and concurrent *in silico* analysis. The read mapping was performed with specific parameter setting (mismatch count, 2; indel count, 3; length fraction, 0.7; similarity fraction, 0.9). In addition, adjusted parameters were also used for quality-based variant detection (minimum coverage, 30x; minimum variant frequency, 25%). Results in variant call format (VCF) file format were used as input for the Illumina VariantStudio Data Analysis version 2.1 Software (Illumina) to annotate genetic variants.

Association Analysis for Common and Low Frequency Variants

The genotypes from all of the 105 individuals of the "variants of interest" from the discovery sample of 41 individuals were used for the candidate gene association study with plasma FVIII levels. First, MAF (minor allele frequency) in our population of 105 individuals were calculated using PLINK package version 1.07 [39]. For this association analysis, the common variants were defined as genetic variants with MAF ≥10%. Thus, low frequency variants were defined as genetic variants with MAF <10%.

Then, linkage disequilibrium (LD) based variant pruning was applied for each variant group (common and low frequency) using PLINK package to find informative variants. Thus, we used the variance inflation factor (VIF) method to check for multi-collinearity of variants and recursively removes variants within a sliding window. The VIF is calculated as $1/(1-R^2)$ where R^2 is the multiple correlation coefficient for a variant being regressed on all other variants simultaneously at each step. In this way, this method considered the correlation between variants but also between linear combinations of them. It was established a variant windows size of 30, a number of variants to shift the windows at each step equal to 3 and the VIF threshold equal to 2 (i.e. implies R^2 of 0.5), which implied that variants greater than this VIF value were removed. Association with plasma FVIII levels was performed by using two different approaches: (a) a single linear association for common variants using PLINK package, and, (b) a collapsing method based on sliding window for low frequency variants using the SNP-set (Sequence) Kernel Association Test (SKAT version 1.0.9) available in R [40]. In the collapsing method, the set of variants were obtained by shifting one low frequency variant to the right within a sliding 2-kb window. SKAT aggregated individual score test statistics of each variant and computed variant-set level p-values, while adjusting for covariates. This allowed different variants to have different directions and magnitudes, including no effects [41]. Both single linear association and collapsing method analyses were adjusted by age, gender and ABO type.

Results

Phenotypes

The phenotypic data of all of the subjects included in our study are shown in Table 1. Similar phenotypic values were obtained in the two groups of samples (GAIT 1 and GAIT 2). In brief, the total number of subjects was similarly distributed in males and females. Also, the mean age and the use of oral contraceptives were comparable in both groups of samples, while the percentage of smokers decreased in GAIT 2.

We measured FVIII:C levels in fresh PPP samples from all of the GAIT participants and found similar mean values in GAIT 1 and GAIT 2 (150.74 ± 52.36 IU/dL *vs* 163.96 ± 64.56 IU/dL,

respectively). Genetic heritability (h2) of plasma FVIII levels was similarly estimated in GAIT 1 (0.40 ± 0.09, *P* = 4.08e-08) and GAIT 2 (0.38 ± 0.07, *P* = 5.27e-10), indicating that ~40% of the phenotypic variation in this trait is due to the additive effect of genes. Since subjects with high FVIII:C are considered at risk for thrombosis, all of the GAIT subjects were grouped in three categories according to FVIII levels. Then, subjects were considered to have low FVIII levels if they exhibited FVIII:C below 100 IU/dL, they were considered to have medium or normal FVIII levels if it ranged between 100 IU/dL and 150 IU/dL and they were considered to have high FVIII levels if they had FVIII:C above 150 IU/dL. The individuals included in the latter category were considered high-risk subjects for thrombosis and they represented 44% and 53% of all of the subjects in GAIT 1 and GAIT 2, respectively (50.23% of the overall GAIT sample). The proportion of the GAIT participants included in each category was similar in GAIT 1 and GAIT 2. Also, we grouped them according to the ABO blood group genotypes (Fig 1). Since the O allele of the ABO blood group provides a protective genotype associated with low FVIII levels and decreased risk of thrombosis, we analyzed the ABO blood genotype of all the GAIT subjects and grouped them in three categories according to the presence of the O allele. As shown in Fig 1, we considered an additive genetic model of allelic effect, in which the addition of one copy of the O allele correlates with a decrease of FVIII levels. The distribution of the subjects in the three categories of the ABO blood group was similar in both GAIT samples.

Table 1. Phenotypic Characteristics of the Subjects Included in the Analysis.

Phenotypic characteristics	Phenotypic data GAIT 1	Phenotypic data GAIT 2*
No. of subjects, n (%)	394 (100)	904 (100)
No. of families	21	35
Mean age, years ± SD	37.47 ± 19.74	39.79 ± 21.40
Males, n (%)	181 (45.94)	449 (49.67)
Current smokers †, n (%)	150 (38.07)	214 (23.67)
Oral contraceptives ‡, n (%)	15 (3.81)	38 (4.20)
Mean FVIII:C levels, IU/dL ± SD	150.74 ± 52.36	163.96 ± 64.56
Heritability (h^2) of FVIII:C levels ± SD	0.40 ± 0.09	0.38 ± 0.07
FVIII:C levels:		
• [<100] IU/dL, n (%)	60 (15.23)	100 (11.06)
• [100-150] IU/dL, n (%)	159 (40.35)	327 (36.17)
• [>150] IU/dL, n (%)	175 (44.42)	477 (52.77)
Categories of the ABO blood group:		
• No O alllele, n (%)	58 (14.72)	166 (18.36)
A1/A1	33	90
A1/A2	14	19
A1/B	10	37
A2/A2	1	1
• One O alllele, n (%)	190 (48.22)	435 (48.12)
A1/O1	147	299
A2/O1	12	35
A1/O2	0	17
A2/O2	0	3
B/O1	30	77
B/O2	1	4
• Two O allleles, n (%)	138 (35.03)	283 (31.31)
O1/O1	130	254
O1/O2	8	29
• NA, n (%)	8 (2.03)	20 (2.21)

* The GAIT 2 study was used for replication and validation of the results from the association analyses. † Subjects in the study were defined as currently smokers when they smoke, independently of the number of cigarettes. ‡ Oral contraceptives use at inclusion. SD: standard deviation. NA: not analyzed.

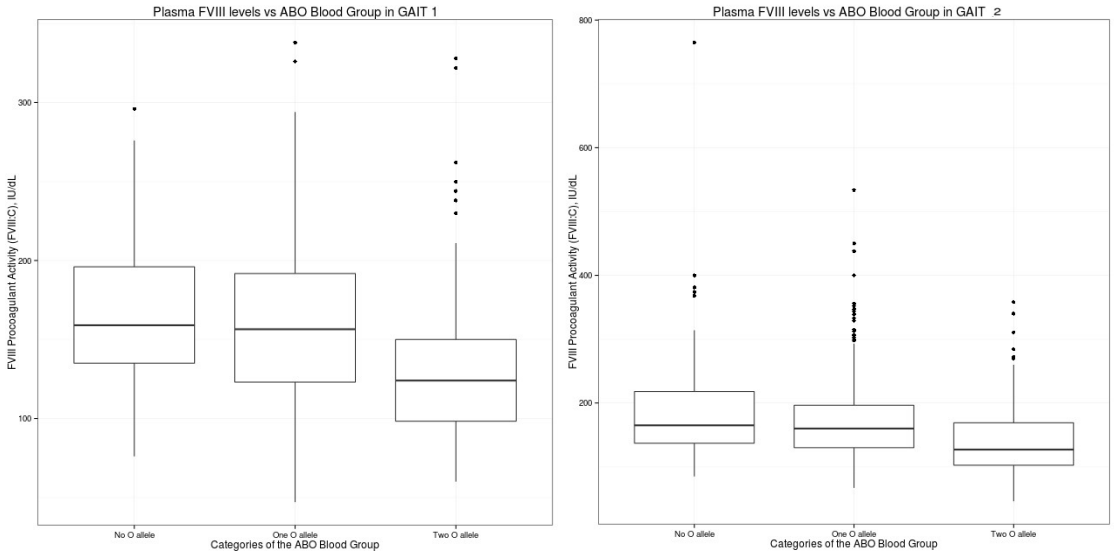

Fig 1. Box plots showing the distribution of plasma FVIII levels according to the ABO blood group in GAIT 1 and GAIT 2. The *y* axis shows the functional clotting FVIII levels and the *x* axis shows the three categories delineated for the ABO blood group according to an additive genetic model of allelic effect in which the presence of the O allele correlates with low FVIII levels.

Models of analysis and covariates

We conducted two polygenic models in GAIT 1 according to the final covariates included in the analysis. The effects of the covariates in the variance of FVIII levels were estimated simultaneously with the genetic effects. The first model was adjusted for age (*P* = 3.0e-04), age-squared (age^2) (*P* = 1.0e-08) and for the ABO blood group genotypes (*P* = 1.0e-05) since these were the covariates that showed a significant effect on the phenotype. The estimated proportion of the variance in plasma FVIII levels due to these covariates was 23%. The covariates with a significant effect on FVIII levels that were included in the second model were age (*P* = 3.9e-02), APCR (*P* = 2.0e-05) and vWF (*P* = 2.9e-49). In this case, 60% of the variance in FVIII levels in plasma was estimated to be under the influence of these covariates. This great effect may reflect the well-known influence of vWF on FVIII levels, since FVIII activity in plasma depends on the availability of its protein carrier, vWF.

Autosomal QTLs Influencing FVIII Levels

A standard multipoint variance-component method was used to assess linkage between autosomal DNA markers and plasma FVIII levels in GAIT 1 for each adjusted model. The results from these linkage scans revealed two quantitative trait loci (QTLs), which may influence the quantitative variation of FVIII levels. One of them was on the short arm of Chromosome 2 (2p) and the other one was on the short arm of Chromosome 3 (3p). The QTL on Chromosome 2 was detected through a peak LOD score of 3.41 (genome-wide *P* = 0.001) located in the interval flanked by markers D2S305 and D2S367, in a region that maps to 2p24.1-2p22.3 (Fig 2). The QTL on Chromosome 3 was detected through a peak LOD score of 3.90 (genome-wide *P* = 0.001) located between markers D3S1289 and D3S1285, in a region that maps to 3p14.3-3p14.1 (Fig 3). Specific data from the linkage analysis are shown in Table 2. A bioinformatic search in these linkage regions showed several potential candidate genes for the control of FVIII levels. The *calcium and integrin binding family member 4* (*CIB4*; Ensembl Gene ID: ENSG00000157884) was proposed as a strong candidate gene in the QTL detected on Chromosome 2 (2p23.3). The *Rho guanine nucleotide exchange factor (GEF) 3* (*ARHGEF3*; Ensembl Gene ID: ENSG00000163947) that maps to 3p14.3 and the *ADAM metallopeptidase with thrombospondin type 1 motif, 9* (*ADAMTS9*; Ensembl Gene ID: ENSG00000163638) that maps to 3p14.1 were two suggested positional genes in the QTL detected on Chromosome 3. A brief summary of these data can be found in Table 3.

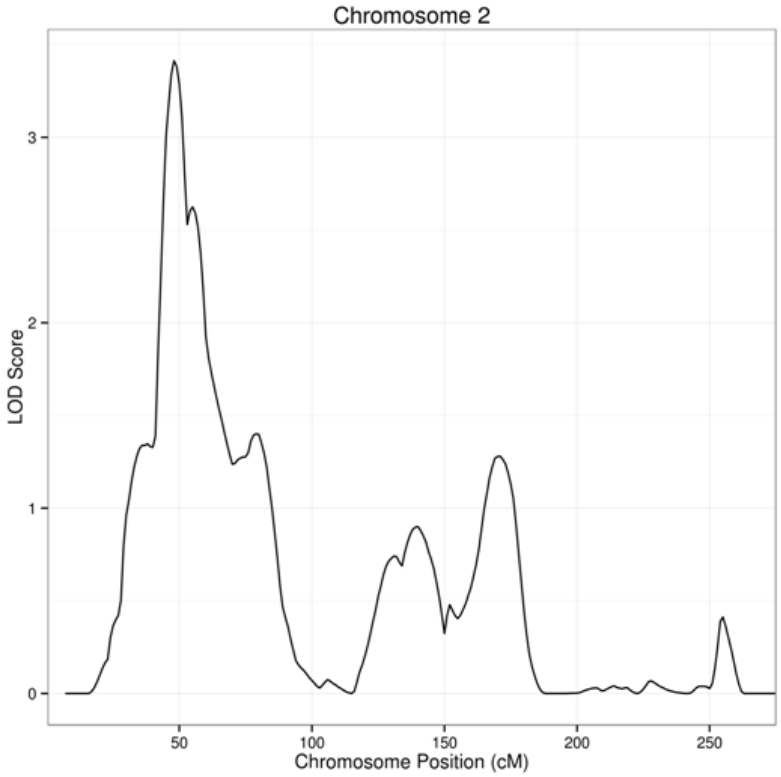

Fig 2. Detailed linkage results for Chromosome 2 in GAIT 1. The *x* axis represents the genomic position (in cM) on Chromosome 2 and the *y* axis shows the logarithm of the odds (LOD) score value. A LOD score of 3 for a linkage signal corresponds to a significant genome-wide $P \leq 0.001$. Significant linkage signal was detected by a peak LOD score of 3.41 at Chromosome position 48 cM defining a QTL for plasma FVIII levels between markers D2S305 and D2S367, in a region that maps to 2p24.1-2p22.3. *CIB4* is a positional gene in this region that could be involved in the quantitative variation of plasma FVIII.

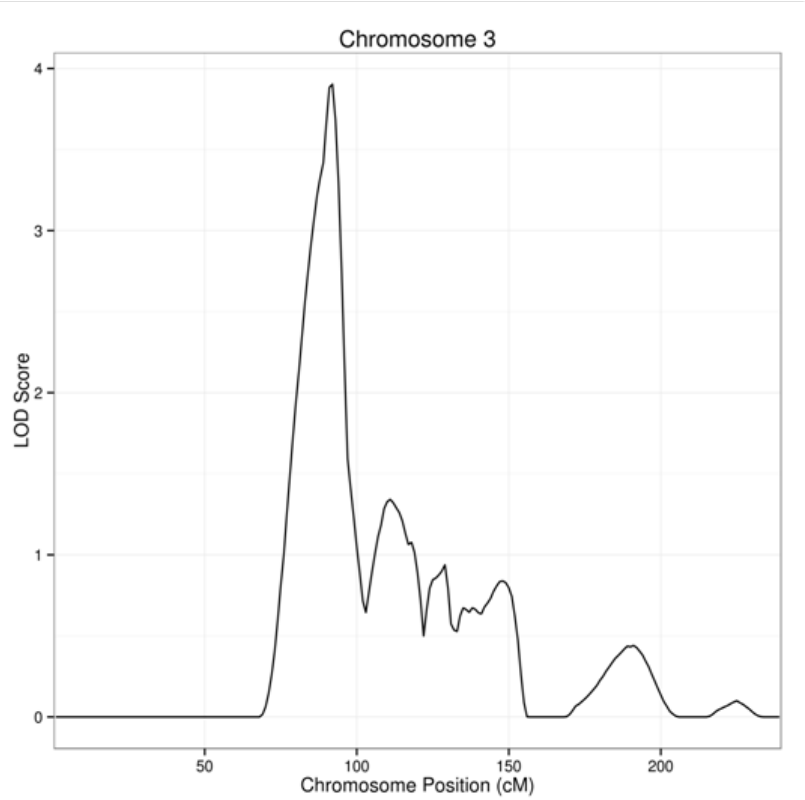

Fig 3. Detailed linkage results for Chromosome 3 in GAIT 1. The *x* axis represents the genomic position (in cM) on Chromosome 3 and the *y* axis shows the logarithm of the odds (LOD) score value. A genome-wide $P \leq 0.001$ for a linkage signal corresponds to a LOD score of 3. Significant linkage signal was detected by a peak LOD score of 3.90 at Chromosome position 92 cM defining a QTL for FVIII levels between markers D3S1289 and D3S1285, in a region that maps to 3p14.3-3p14.1. Positional genes in this linkage region that have been suggested to influence the quantitative variation of plasma FVIII levels in our sample are *ARHGEF3* and *ADAMTS9*.

Table 2. Data of the Linkage Analysis for FVIII Levels in GAIT 1.

Linkage signal	Flanking markers	Cytoband	Chr. Position* (cM)	LOD score	Genome-wide *P*-value
Chr. 2	D2S305-D2S367	2p24.1-2p22.3	48	3.41	0.001
Chr. 3	D3S1289-D3S1285	3p14.3-3p14.1	92	3.90	0.001

Chr.: chromosome. * All Chromosome positions were based on the National Center for Biotechnology Information (NCBI) build 37.

Table 3. Candidate Genes Suggested for the Variance in FVIII Levels in GAIT.

Gene symbol	Description	Cytoband	Chr. Position	Subcellular location	Biological process
CIB4	*calcium and integrin binding family member 4*	2p23.3	Chr2: 26804070-26864236	Cytoskeleton (Actin filaments)	Calcium ion and integrin binding.
ARHGEF3	*Rho guanine nucleotide exchange factor (GEF) 3*	3p14.3	Chr3: 56761446-57113357	Cytoplasm	Guanine nucleotide exchange factor (GEF) for RhoA and RhoB GTPases.
ADAMTS9	*ADAM metallopeptidase with thrombospondin type 1 motif, 9*	3p14.1	Chr3: 64501333-64673676	Extracellular matrix	Glycoprotein catabolic process: cleaves the large aggregating proteoglycans, aggrecan and versican.

Chr.: chromosome. * All Chromosome positions were based on the National Center for Biotechnology Information (NCBI) build 37.

Fine-mapping of the QTLs

We conducted a specific search for genotype-phenotype associations in the QTLs detected in our genome-wide linkage analyses to detect common variants susceptible to influence plasma FVIII levels. The most significant SNPs with their *P*-values included in these genomic regions are shown in Table 4.

Fine-mapping of the QTL on Chromosome 2

A total of 260 SNPs were typed in the linkage region detected on Chromosome 2 (Fig 4). Among all of the SNPs typed in this region, only the rs1276123 (*P* = 1.33e-04) remained significant for the association with FVIII levels after applying the adjustment for multiple testing (Table 4). This SNP association was replicated in GAIT 2 (*P*=0.011) (Table 5). The rs1276123 is located in an intronic region of the *CIB4* gene. Another intronic variant of *CIB4* appeared among the ten top SNP-associations detected with FVIII levels in GAIT 1, although this SNP did not reach the threshold for statistical significance after Bonferroni adjustment. We want to

emphasize that *CIB4* was identified by linkage and association analyses, as well as the significant SNP association with FVIII levels was replicated in an independent sample, thus giving more confidence to our results.

Fine-mapping of the QTL on Chromosome 3

A total of 2319 SNPs were genotyped to fine-map the linkage region detected on Chromosome 3 (Fig 5). The most significant polymorphism association with FVIII levels was for the rs681751 (*P*=5.2e-05), although it did not reach statistical significance after applying the adjustment for multiple comparisons. The rs681751 is located in a large intergenic region at a distance of 70 Kbp downstream of the gene *CACNA1D* (*calcium channel, voltage-dependent, L type, alpha 1D subunit*). Then, no significant SNP association with the procoagulant FVIII was obtained in this segment on Chromosome 3. This could indicate that the interval that lies beneath the linkage peak may contain a rare variant that could contribute to the strong linkage signal detected.

Table 4. Top SNP-associations with FVIII Levels in GAIT from Fine-Mapping on Chromosome 2 and Chromosome 3.

SNP	Chr	Position (bp)*	Type	Closest Gene	Alleles [a]	Minor Allele	MAF †	beta ‡	*P*-value **
rs1276123	2	26824246	Intronic	*CIB4*	C/T	T	0.2210	-17.72	$1.33x10^{-04}$ [(b)]
rs1484685	2	22843052	Intergenic	*KLHL29*	T/C	C	0.3061	12.30	$1.57x10^{-03}$
rs13387847	2	29426938	Intronic	*ALK*	A/G	A	0.1969	13.73	$3.52x10^{-03}$
rs1509577	2	30410818	Intergenic	*YPEL5*	C/G	G	0.4160	11.00	$3.56x10^{-03}$
rs2464091	2	26844925	Intronic	*CIB4*	T/A	A	0.4589	-10.85	$3.76x10^{-03}$
rs4952095	2	30320345	Intergenic	*YPEL5*	C/T	C	0.1637	-15.57	$6.41x10^{-03}$
rs1122899	2	23555150	Intergenic	*KLHL29*	T/C	T	0.2331	10.72	$1.08x10^{-02}$
rs2195122	2	19675315	Intergenic	*OSR1*	G/A	G	0.3161	9.44	$1.09x10^{-02}$
rs733771	2	23570037	Intergenic	*KLHL29*	A/G	A	0.2974	-10.06	$1.30x10^{-02}$
rs2060790	2	19562433	Intergenic	*OSR1*	T/G	G	0.3680	9.46	$1.47x10^{-02}$
rs681751	3	53458934	Intergenic	*CACNA1D*	T/G	T	0.1250	16.82	$5,20x10^{-05}$
rs11709171	3	53669904	Intronic	*CACNA1D*	A/C	C	0.2060	11.71	$6,50x10^{-04}$
rs1464349	3	54608038	Intergenic	*CACNA2D3*	G/A	G	0.4466	9.60	$8,81x10^{-04}$
rs1352008	3	56142860	Intronic	*ERC2*	C/T	C	0.0413	20.98	$1,11x10^{-03}$
rs4077115	3	54953492	Intronic	*ERC2*	A/G	A	0.0645	20.07	$1,12x10^{-03}$
rs1352009	3	56186069	Intronic	*ERC2*	C/T	T	0.0375	20.46	$1,65x10^{-03}$
rs3773016	3	57909258	Intronic	*SLMAP*	T/G	G	0.1275	-13.10	$1,76x10^{-03}$
rs2292662	3	63897215	Intronic	*ATXN7*	C/T	T	0.2014	-11.09	$2,02x10^{-03}$
rs658071	3	53509107	Intergenic	*RP11-72H11.1*	T/C	C	0.0966	14.01	$2,03x10^{-03}$
rs704364	3	63874734	Intronic	*ATXN7*	G/A	A	0.3608	-8.93	$2,42x10^{-03}$

Chr: chromosome. * All Chromosome positions were based on the National Center for Biotechnology Information (NCBI) build 37. [a] Alleles aligned to + strand. † Minor Allele Frequency of the SNP in our sample. ‡ represents the effect of one copy of the rare allele in FVIII levels. ** *P*-value of the association with FVIII levels. [(b)] indicates that statistical significance remains after Bonferroni correction for multiple testing.

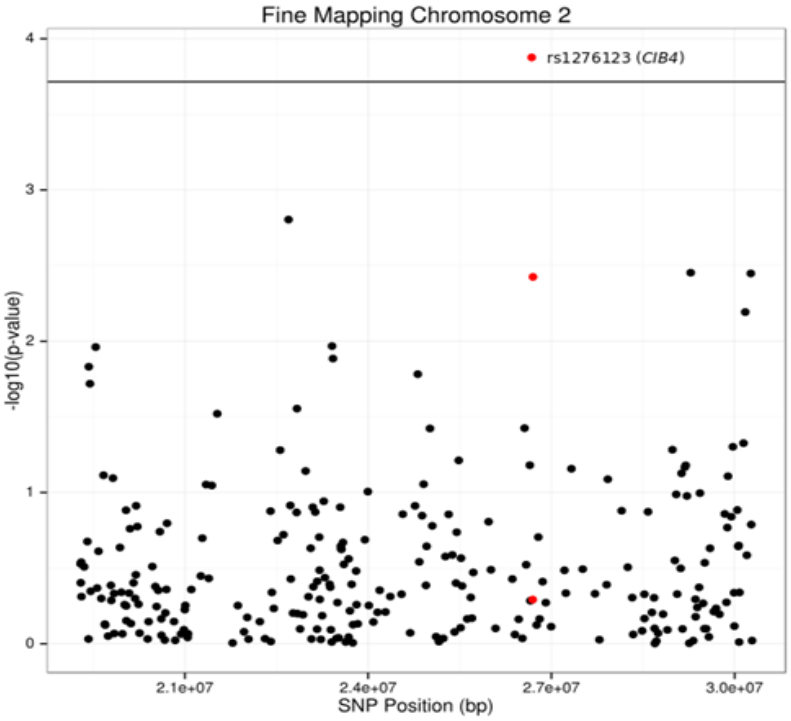

Fig 4. Fne-mapping in the Linkage Region on Chromosome 2 in GAIT 1. The *x* axis represents the genomic position (in bp) of 260 SNPs that were used to fine-map the linkage region on Chromosome 2 and the *y* axis shows –Log10 (*P*-value). The horizontal line indicates the threshold for statistical significance after Bonferroni correction at 1.93e-04. The rs1276123 (MAF=0.2210; *P* = 1.33e-04) (in red and located above the horizontal line) showed significant association with FVIII levels after correcting for multiple testing. This SNP was located within an intronic region of the *CIB4* gene, which emerges as a strong candidate gene for the control of FVIII levels. The remaining two red points below the significant threshold corresponds to other SNPs that were typed in *CIB4*.

Table 5. Replication of Significant SNP-associations with FVIII Levels in GAIT.

		GAIT 1				GAIT 2			
SNP	Alleles[a]	MAF †	beta ‡	VE (%) Φ	*P*-value **	MAF †	beta ‡	VE (%) Φ	*P*-value **
rs1276123	C/T	0.2	-17.7	4.3	$1.3x10^{-04}$ (b)	0.17	-10.3	1.1	$1.1x10^{-02}$

VE: variance explained.[a] Alleles aligned to + strand. * Chromosome position was based on the National Center for Biotechnology Information (NCBI) build 37. † Minor Allele Frequency of the SNP in our sample. ‡ represents the effect of one copy of the rare allele in FVIII levels. Φ represents the variance in FVIII levels that is explained by the effect of one copy of the rare allele. ** *P*-value of the association with FVIII levels. Statistical significance was set when a *P*-value was 0.05. [(b)] indicates that statistical significance remains after Bonferroni correction for multiple testing.

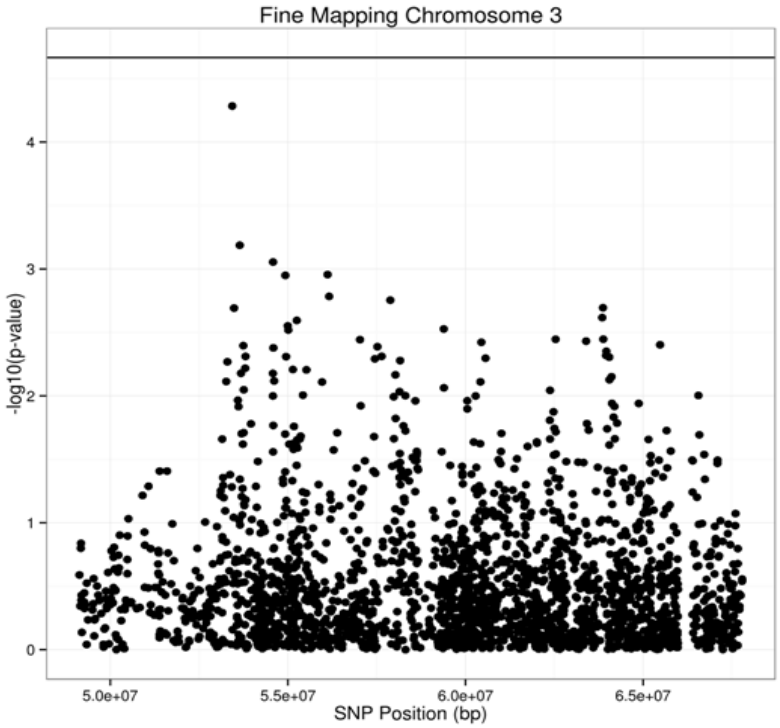

Fig 5. Fine-mapping in the Linkage Region detected on Chromosome 3 in GAIT 1. The figure shows the –Log10 (*P*-value) in the *y* axis and the genomic position (bp) of each SNP used for the fine-mapping in the *x* axis. The significant *P*-value after Bonferroni correction was set at 2.16e-05 (horitzontal line). The QTL detected on Chromosome 3 by the genome-wide linkage analysis in GAIT 1 was fine-mapped using 2319 SNPs in order to find SNP associations with plasma FVIII levels. No significant SNP association with FVIII levels remained after the adjustment for multiple comparisons. This may indicate that a rare variant in one or more GAIT families could be responsible of the linkage signal detected on Chromosome 3.

NGS

Statistics

We selected 105 individuals of the GAIT 2 Project to amplify 63,264 bp encompassing *CIB4*, which were analysed by NGS. Briefly, the length of *CIB4* region was 61,887 nucleotides and 10,832,926 reads covered this region. Interestingly, the 99% of the positions had coverage above 30x and the median coverage per individual was 171.87x.

A total of 470 unique biallelic variants were identified in the discovery sample. In detail, the 10.2% (48) were indels and the 2.1% (10) were exonic variants. Moreover, the 22.3% (105) of the genetic variants had a MAF from all populations of 1000 Genomes (Phase 1 released April 2012 version 3 http://www.1000genomes.org/) >1% and also from four populations of 1000

Genomes (American, East Asian, African and European) and the 44.3% (208) among the 470 variants had been reported in dbSNP version 137.

Association Analysis

After LD based pruning, 64 common variants and 200 low frequency variant sets remained among the 470 "variants of interest". We identified 7 common variants that showed significant association with plasma FVIII levels (p-value <0.05) using single linear association (Table 6). In detail, these common variants were located within intronic regions and among them 5 variants were related with high FVIII levels. Moreover, the variant rs1275934 was revealed as the most significant variant associated with FVIII levels. In addition, we identified 2 significant low frequency variant sets (p-value <0.05) using the collapsing method. Briefly, the first region (chr2:26,840,013-26,842,013) is located in intron 3 and the second region (chr2:26,859,563-26,861,563) is located in intron 2.

Table 6. Significant common genetic variants in *CIB4* (NM_001029881.1) associated with plasma FVIII levels.

Genome Location	Nucleotide Change	Intron	dbSNP v137	Tested Allele	P-value	Beta
chr2:26,810,043	c.329-3277A>C	4	-	A	0.040	42.62
chr2:26,810,128	c.329-3362T>C	4	-	C	0.026	-32.67
chr2:26,812,469	c.328+5575A>G	4	-	G	0.036	-31.04
chr2:26,822,857	c.187-4672A>G	3	-	G	0.020	-33.91
chr2:26,823,109	c.187-4925delT	3	-	A	0.027	-39.15
chr2:26,823,234	c.187-5049A>G	3	rs1275934	A	0.003	-42.42
chr2:26,832,357	c.187-14172G>A	3	rs1275967	G	0.050	-23.80

Discussion

High plasma FVIII levels are a strong and independent contributor to VT risk in a dose-dependent manner [2] that also predispose to the recurrence of the disease [26]. Attempts to identify major genetic loci associated with high FVIII levels have failed until now. Unfortunately, the causes of elevated FVIII are not yet fully understood and no therapeutic approach does exist. The identification of the biological mechanism involved in the control of plasma FVIII

concentration would be clue for the management of raised levels of FVIII in subjects at risk of thrombosis and would contribute to drastically reduce the prevalence and mortality rate of this vascular disease.

Current data suggest that high levels of FVIII may be controlled by additional genetic factors other than *ABO*, *VWF* and *F8*. Accordingly, other loci [28,29] and polymorphisms [24,43–47] with an effect on FVIII levels have been suggested although they have shown poor concluding data. A summary of the studies addressed to identify common variants associated with FVIII levels in different populations is shown in S2 Table.

In order to identify new genomic regions that co-segregate in families along with variable levels of FVIII in plasma we conducted a genome-wide linkage scan and fine-mapping of the linkage regions using 398 subjects from the GAIT Project (GAIT 1). Our results represent the first genome-wide scan conducted using families with idiopathic thrombophilia to identify genes implicated in the variance in FVIII levels, with a final effect on the susceptibility to thrombosis. We identified two QTLs involved in the variation of FVIII levels. One significant linkage signal was located on Chromosome 2 (LOD = 3.41) and the other linkage region was detected on Chromosome 3 (LOD = 3.90). The QTL on Chromosome 2 contained one positional gene. This gene was *CIB4* that maps to 2p23.3 and encodes a protein involved in calcium and integrin binding. The QTL on Chromosome 3 contained two suggested candidate genes for the quantitative variation of plasma FVIII levels. One of them was the *ARHGEF3*, which maps to 3p14.3 and encodes a Rho guanine nucleotide exchange factor (GEF). The other one was the gene *ADAMTS9* that maps to 3p14.1 and encodes a protease involved in the catabolic process of glycoproteins.

Fine-mapping of the Chromosome 2 linkage region in GAIT 1 revealed one significant SNP association with FVIII levels after applying the most stringent correction for multiple testing. This SNP was the rs1276123 (*P* = 1.33e-04) an intron-variant of the *CIB4* gene. Interestingly, *CIB4* was identified by using linkage and association analyses, two distinct and complementary statistical methods, thus giving more confidence to our finding. In addition, it is especially important to highlight that the association between the rs1276123 and FVIII levels was replicated (*P* = 1.14e-02) in GAIT 2 (an independent sample consisting of a new set of

thrombophilic families including 935 individuals from 35 Spanish extended pedigrees). These data validate our results and suggests a functional effect of the rs1276123 on FVIII levels. The minor allele frequency of the rs1276123 was 0.2210 in GAIT 1 and 0.1699 in GAIT 2 for the T allele. This is in accordance with the allele frequencies of the rs1276123 in Europeans from phase 1 of the 1000 Genomes Project http://www.1000genomes.org/ (C: 0.846 and T: 0.154). Notably, the presence of the rs1276123-T allele was associated with 18% and 10% reduction of FVIII levels in GAIT 1 and GAIT 2, respectively. This suggests that the rs1276123-T allele of *CIB4* would have a protective role against thromboembolic disease. *CIB4* is a member of the calcium- and integrin-binding protein (CIB) family, whose functional role *in vivo* has not been established. The CIB family consists of four isoforms CIB1, CIB2, CIB3 and CIB4. Very little information does exist about CIB4 protein; however, CIB1 is widely expressed in human tissues and it is known to interact with multiple effector proteins. The functional role of many of these interactions are unknown, although the implication of CIB1 in haemostasis and thrombosis, apoptosis, embryogenesis, DNA damage response and regulation of Ca2+ signals has been described [48]. The functional role of CIB1 in haemostasis is supported by its specific interaction with the cytoplasmic domain of the platelet αIIbβ3 integrin (or GPIIb/IIIa), a platelet receptor for fibrinogen that mediates platelet aggregation and platelet spreading on the subendothelium at sites of vascular injury for the arrest of bleeding [49–51]. Ablation of CIB1 impairs mouse tail bleeding time, favours a rebleeding phenotype, difficults the generation of an occlusive thrombus in an arterial thrombosis model and decreases platelet spreading [52]. These data indicate that CIB1 plays a key role in normal thrombus formation. More interestingly, CIB1 has been shown to influence in vitro expression and bioactivity of FVIII by an unknown mechanism [53], although it has been suggested that CIB1 may act as a regulatory molecule that modulates FVIII secretion and activity by the interaction with the A1 domain of FVIII. The A1 domain has been reported to inhibit the secretion of FVIII, so the binding of CIB1 might suppress this inhibitory effect leading to an increase of FVIII antigen levels and its procoagulant activity. CIB4 contains 3 EF-hand domains and shares 64% sequence homology with CIB1 [54]. One study have reported that platelet function in *Cib1*$^{-/-}$ mice remained unaltered and suggested that other family members can compensate for CIB1 loss, thus preventing altered thrombus generation

[55]. This indicates that these isoforms may be able to exert similar functions to CIB1 *in vivo* under certain conditions. Our results may contribute to describe for the first time a functional role of CIB4 in humans in haemostasis and thrombosis through an unknown mechanism involving FVIII clotting factor. We hypothesize that CIB4 could mediate the synthesis or secretion of FVIII by a signaling pathway or direct interaction with the clotting factor, thus modulating FVIII levels in plasma. Functional analyses are needed to test direct effect of CIB4 on the levels of FVIII antigen and its procoagulant activity. The finding of the *CIB4* gene by both linkage and association analyses, together with the biological function in haemostasis of a member of the CIB family, reinforce the role of *CIB4* as prime candidate gene in the control of FVIII levels. Moreover, we performed NGS of the target gene *CIB4* to identify additional variants associated with the variability of FVIII levels over the whole spectrum of allele frequency. This cost-effective methodology showed 7 common variants located in intronic regions in association with FVIII levels. In addition, 2 low frequency variant sets were associated also with FVIII levels. Thus, these regions within the introns 2, 3 and 4 are suggested to be related with the variability of the levels of this coagulation protein. Interestingly, this supports the analysis of introns and not only exonic regions as introns can be involved also into regulatory functions.

None of the SNPs used to fine-map the Chromosome 3 linkage region remained significant for the association with FVIII levels after applying the Bonferroni adjustment. Thus, we were not able to detect common variants involved in quantitative variation of FVIII, although we do not rule out the possibility of an effect of a rare variant on this clotting phenotype. This is supported by the fact that among the 21 families of GAIT 1 that were included in the genome-wide linkage analysis, three out families showed a clear strong contribution to the linkage signal identified on Chromosome 3. According to our search for candidate genes in this linkage region, we suggest *ARHGEF3* and *ADAMTS9* as positional genes for the regulation of plasma FVIII levels. *ARHGEF3* encodes the rho guanine nucleotide exchange factor 3 (RhoGEF3), which specifically activates RhoA and RhoB, two members of the Rho GTPase family, by catalyzing the exchange of GDP for GTP. Rho GTPases play an important role in many cellular processes such as regulation of cell morphology, cell aggregation, cytoskeletal rearrangements, and transcriptional activation. They have been implicated in the pathogenesis of several

cardiovascular disorders including hypertension, coronary and cerebral vasospasm, artheriosclerosis and diabetes [56]. RhoA is the most characterized small GTPase and it has been related to endothelial dysfunction, inflammation and vascular smooth muscle cell proliferation in artheriosclerosis [57]. Also, RhoA is an important regulator of platelet function in thrombosis and haemostasis [58]. *ARHGEF3* is expressed in platelets, macrophages, megakaryocytes and erythroblasts, among other tissues. Interestingly, *ARHGEF3* has been associated with the variance in the number and volume of platelets [59–62]. In particular, increased mean platelet volume (MPV) has been associated with myocardial and cerebral infarction and is considered an independent and strong predictor for postevent morbidity and mortality. In addition, other members of the *ARHGEF* gene family have been associated with the susceptibility of atherothrombotic stroke [63,64]. Also, the SNP rs6445834 in *ARHGEF3* has been associated with the susceptibility to the metabolic syndrome [65], an important cardiovascular risk factor including VT. We hypothesize that *ARHGEF3* could be involved in an intracellular signaling pathway with an effect on FVIII levels.

The other potential candidate gene is *ADAMTS9*, which is highly expressed in embryonic and adult tissues, with highest expression in heart, placenta and skeletal muscle. This gene encodes an enzyme of the ADAMTS (a disintegrin and metalloproteinase with thrombospondin motifs) protein family [66]. Members of this family share an ADAM protease domain and a variable number of thrombospondin type 1 motif. The ADAMTS9 protein is the most highly conserved member of the ADAMTS family. Proteins of this family have been involved in the cleavage of proteoglycans, the control of organ shape during development, and the inhibition of angiogenesis. *ADAMTS9* has been considered a thrombosis-related gene up-regulated in placenta from a group of smoker women [67]. Interestingly, a protein of the same family, which is encoded by the *ADAMTS13* gene, is known to be responsible for cleaving at the site of Tyr842-Met843 of the vWF, and thus it plays a role in thrombosis [68]. Another protein containing an ADAM protease domain, which is encoded by the gene *ADAM-like, decysin 1* (*ADAMDEC1*), has been characterized for an association with FVIII levels in venous thromboembolism [46]. In particular, the *ADAMDEC1* haplotype tgtgg/tgtgg has been moderately associated with high FVIII levels in thrombosis patients. In addition to the protease

activity of *ADAMTS9*, it has been recently shown *in vitro* to have a role in protein transport from the endoplasmic reticulum to the golgi, although clinical implications of a potential defect in this function have not been reported [69]. We speculate that *ADAMTS9* could modulate FVIII levels by influencing either its clearance from the circulation or its cellular secretion by an unknown mechanisms that could be related with a proteolytic process. Further investigations are essential to confirm an effect of *ADAMTS9* on FVIII plasma levels.

Our results from a family-based genome-wide scan provide data of potential autosomal candidate genes that could influence the quantitative variation of FVIII plasma levels in the Spanish population. This shed light to new genetic factors with an effect on FVIII levels variance. The knowledge of these factors could help to elucidate the biological mechanisms underlying the thrombosis risk associated with elevated FVIII levels and contribute to reduce the high prevalence and mortality of the thrombotic disease.

Acknowledgments

We would like to acknowledge the advice and helpful discussion of Professor W.H. Stone. Also, we thank R. Pérez for her technical support as well as A. Cárdenas, J. Nicolau and O. Solà for their administrative help with regard to the inclusion of all of the subjects. Finally, we are indebted to all of the families who participated in the GAIT Project.

Financial Disclosure

This study was supported by the ‘Instituto de Salud Carlos III-Fondo de Investigación Sanitaria’ (ISCIII-FIS) (PI 11/0184), ‘Red de Investigación Cardiovascular (RIC) (RD12/0042/0032) and ‘Agència de Gestió d’ajuts Universitaris i de Recerca’ (AGAUR) (SGR-01068 and SGR-1240). Sonia López was supported by ‘Contratos Posdoctorales de Perfeccionamiento Sara Borrell’ from ISCIII-FIS (CD08/00059). Laura Martin-Fernandez was supported by Ayudas Predoctorales de Formación en Investigación en Salud (PFIS) FI12/00322.

References

1 Kraaijenhagen RA, in't Anker PS, Koopman MM, Reitsma PH, Prins MH, van den Ende A, Buller HR. High plasma concentration of factor VIIIc is a major risk factor for venous thromboembolism. Thromb Haemost 2000/02/11 ed. 2000; 83: 5–9.

2 Koster T, Blann AD, Briet E, Vandenbroucke JP, Rosendaal FR. Role of clotting factor VIII in effect of von Willebrand factor on occurrence of deep-vein thrombosis. Lancet 1995/01/21 ed. 1995; 345: 152–5.

3 Kyrle PA, Minar E, Hirschl M, Bialonczyk C, Stain M, Schneider B, Weltermann A, Speiser W, Lechner K, Eichinger S. High plasma levels of factor VIII and the risk of recurrent venous thromboembolism. N Engl J Med 2000/08/19 ed. 2000; 343: 457–62.

4 Folsom AR, Wu KK, Rosamond WD, Sharrett AR, Chambless LE. Prospective study of hemostatic factors and incidence of coronary heart disease: the Atherosclerosis Risk in Communities (ARIC) Study. Circulation 1997/08/19 ed. 1997; 96: 1102–8.

5 Tracy RP, Arnold AM, Ettinger W, Fried L, Meilahn E, Savage P. The relationship of fibrinogen and factors VII and VIII to incident cardiovascular disease and death in the elderly: results from the cardiovascular health study. Arterioscler Thromb Vasc Biol 1999/07/09 ed. 1999; 19: 1776–83.

6 Folsom AR, Rosamond WD, Shahar E, Cooper LS, Aleksic N, Nieto FJ, Rasmussen ML, Wu KK. Prospective study of markers of hemostatic function with risk of ischemic stroke. The Atherosclerosis Risk in Communities (ARIC) Study Investigators. Circulation 1999/08/18 ed. 1999; 100: 736–42.

7 Rumley A, Lowe GD, Sweetnam PM, Yarnell JW, Ford RP. Factor VIII, von Willebrand factor and the risk of major ischaemic heart disease in the Caerphilly Heart Study. Br J Haematol 1999/05/08 ed. 1999; 105: 110–6.

8 O'Donnell J, Mumford AD, Manning RA, Laffan M. Elevation of FVIII: C in venous thromboembolism is persistent and independent of the acute phase response. Thromb Haemost 2000/02/11 ed. 2000; 83: 10–3.

9 Kamphuisen PW, Eikenboom JC, Vos HL, Pablo R, Sturk A, Bertina RM, Rosendaal FR. Increased levels of factor VIII and fibrinogen in patients with venous thrombosis are not caused by acute phase reactions. Thromb Haemost 1999/06/12 ed. 1999; 81: 680–3.

10 De Lange M, Snieder H, Ariens RA, Spector TD, Grant PJ. The genetics of haemostasis: a twin study. Lancet 2001/02/24 ed. 2001; 357: 101–5.

11 Orstavik KH, Magnus P, Reisner H, Berg K, Graham JB, Nance W. Factor VIII and factor IX in a twin population. Evidence for a major effect of ABO locus on factor VIII level. Am J Hum Genet 1985/01/01 ed. 1985; 37: 89–101.

12 Rosendaal FR, Bovill EG. Heritability of clotting factors and the revival of the prothrombotic state. Lancet 2002/03/07 ed. 2002; 359: 638–9.

13 Souto JC, Almasy L, Borrell M, Gari M, Martinez E, Mateo J, Stone WH, Blangero J, Fontcuberta J. Genetic determinants of hemostasis phenotypes in Spanish families. Circulation 2000/04/04 ed. 2000; 101: 1546–51.

14 Souto JC, Almasy L, Borrell M, Blanco-Vaca F, Mateo J, Soria JM, Coll I, Felices R, Stone W, Fontcuberta J, Blangero J. Genetic susceptibility to thrombosis and its relationship to physiological risk factors: the GAIT study. Genetic Analysis of Idiopathic Thrombophilia. Am J Hum Genet 2000/10/20 ed. 2000; 67: 1452–9.

15 Kamphuisen PW, Houwing-Duistermaat JJ, van Houwelingen HC, Eikenboom JC, Bertina RM, Rosendaal FR. Familial clustering of factor VIII and von Willebrand factor levels. Thromb Haemost 1998/03/11 ed. 1998; 79: 323–7.

16 Mansvelt EP, Laffan M, McVey JH, Tuddenham EG. Analysis of the F8 gene in individuals with high plasma factor VIII: C levels and associated venous thrombosis. Thromb Haemost 1998/11/03 ed. 1998; 80: 561–5.

17 Kamphuisen PW, Eikenboom JC, Rosendaal FR, Koster T, Blann AD, Vos HL, Bertina RM. High factor VIII antigen levels increase the risk of venous thrombosis but are not associated with polymorphisms in the von Willebrand factor and factor VIII gene. Br J Haematol 2001/11/28 ed. 2001; 115: 156–8.

18 De Visser MC, Sandkuijl LA, Lensen RP, Vos HL, Rosendaal FR, Bertina RM. Linkage analysis of factor VIII and von Willebrand factor loci as quantitative trait loci. J Thromb Haemost 2003/08/13 ed. 2003; 1: 1771–6.
19 Scanavini D, Legnani C, Lunghi B, Mingozzi F, Palareti G, Bernardi F. The factor VIII D1241E polymorphism is associated with decreased factor VIII activity and not with activated protein C resistance levels. Thromb Haemost 2005/03/01 ed. 2005; 93: 453–6.
20 Nossent AY, Eikenboom JC, Vos HL, Bakker E, Tanis BC, Doggen CJ, Bertina RM, Rosendaal FR. Haplotypes encoding the factor VIII 1241 Glu variation, factor VIII levels and the risk of venous thrombosis. Thromb Haemost 2006/05/30 ed. 2006; 95: 942–8.
21 Viel KR, Machiah DK, Warren DM, Khachidze M, Buil A, Fernstrom K, Souto JC, Peralta JM, Smith T, Blangero J, Porter S, Warren ST, Fontcuberta J, Soria JM, Flanders WD, Almasy L, Howard TE. A sequence variation scan of the coagulation factor VIII (FVIII) structural gene and associations with plasma FVIII activity levels. Blood 2007/01/09 ed. 2007; 109: 3713–24.
22 Campos M, Buchanan A, Yu F, Barbalic M, Xiao Y, Chambless LE, Wu KK, Folsom AR, Boerwinkle E, Dong JF. Influence of single nucleotide polymorphisms in factor VIII and von Willebrand factor genes on plasma factor VIII activity: the ARIC Study. Blood 2012/01/06 ed. 2012; 119: 1929–34.
23 Shen W, Gu Y, Zhu R, Zhang L, Zhang J, Ying C. Copy number variations of the F8 gene are associated with venous thromboembolism. Blood Cells Mol Dis 2013/02/14 ed. 2013; 50: 259–62.
24 Antoni G, Oudot-Mellakh T, Dimitromanolakis A, Germain M, Cohen W, Wells P, Lathrop M, Gagnon F, Morange PE, Tregouet DA. Combined analysis of three genome-wide association studies on vWF and FVIII plasma levels. BMC Med Genet 2011/08/04 ed. 2011; 12: 102.
25 Souto JC, Almasy L, Muniz-Diaz E, Soria JM, Borrell M, Bayen L, Mateo J, Madoz P, Stone W, Blangero J, Fontcuberta J. Functional effects of the ABO locus polymorphism on plasma levels of von Willebrand factor, factor VIII, and activated partial thromboplastin time. Arterioscler Thromb Vasc Biol 2000/08/11 ed. 2000; 20: 2024–8.
26 Kamphuisen PW, Eikenboom JC, Bertina RM. Elevated factor VIII levels and the risk of thrombosis. Arterioscler Thromb Vasc Biol 2001/05/23 ed. 2001; 21: 731–8.
27 De Visser MC, Rosendaal FR, Bertina RM. A reduced sensitivity for activated protein C in the absence of factor V Leiden increases the risk of venous thrombosis. Blood 1999/02/09 ed. 1999; 93: 1271–6.
28 Soria JM, Almasy L, Souto JC, Buil A, Martinez-Sanchez E, Mateo J, Borrell M, Stone WH, Lathrop M, Fontcuberta J, Blangero J. A new locus on chromosome 18 that influences normal variation in activated protein C resistance phenotype and factor VIII activity and its relation to thrombosis susceptibility. Blood 2002/10/24 ed. 2003; 101: 163–7.
29 Berger M, Mattheisen M, Kulle B, Schmidt H, Oldenburg J, Bickeboller H, Walter U, Lindner TH, Strauch K, Schambeck CM. High factor VIII levels in venous thromboembolism show linkage to imprinted loci on chromosomes 5 and 11. Blood 2004/09/09 ed. 2005; 105: 638–44.
30 Camacho M, Martinez-Perez A, Buil A, Siguero L, Alcolea S, Lopez S, Fontcuberta J, Souto JC, Vila L, Soria JM. Genetic determinants of 5-lipoxygenase pathway in a Spanish population and their relationship with cardiovascular risk. Atherosclerosis 2012/07/28 ed. 2012; 224: 129–35.
31 Miller SA, Dykes DD, Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res 1988/02/11 ed. 1988; 16: 1215.
32 Soria JM, Almasy L, Souto JC, Bacq D, Buil A, Faure A, Martinez-Marchan E, Mateo J, Borrell M, Stone W, Lathrop M, Fontcuberta J, Blangero J. A quantitative-trait locus in the human factor XII gene influences both plasma factor XII levels and susceptibility to thrombotic disease. Am J Hum Genet 2002/01/24 ed. 2002; 70: 567–74.
33 Dyke B. PEDSYS, a pedigree data management system. User's manual. Technical Report No 2 Population Genetics Laboratory, Department of Genetics, Southwest Foundation for Biomedical research, San antonio, TX 1995; .
34 Almasy L, Blangero J. Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet 1998/05/23 ed. 1998; 62: 1198–211.
35 Boehnke M, Lange K. Ascertainment and goodness of fit of variance component models for pedigree data. Prog Clin Biol Res 1984/01/01 ed. 1984; 147: 173–92.
36 Boerwinkle E, Chakraborty R, Sing CF. The use of measured genotype information in the analysis of quantitative phenotypes in man. I. Models and analytical methods. Ann Hum Genet 1986/05/01 ed. 1986; 50: 181–94.

37 Corrales I, Ramírez L, Altisent C, Parra R, Vidal F. Rapid molecular diagnosis of von Willebrand disease by direct sequencing. Detection of 12 novel putative mutations in VWF gene. Thromb Haemost. 2009;101:570–6.

38 Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IMJ, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar S V, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DMD, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Chiara E Catenazzi M, Chang S, Neil Cooley R, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo K V, Scott Furey W, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. Macmillan Publishers Limited. All rights reserved; 2008;456:53–9.

39 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet [Internet]. 2007;81:559–75. Available from: http://pngu.mgh.harvard.edu/purcell/plink/

40 Lee S, Miropolsky L, Wu M. SKAT: SNP-Set (Sequence) Kernel Association Test. R package version 1.0.9. [Internet]. 2015. Available from: https://cran.r-project.org/web/packages/SKAT/

41 Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011;89:82–93.

42 International HapMap Consortium. A haplotype map of the human genome. Nature. 2005;437:1299–320.

43 Morange PE, Tregouet DA, Frere C, Saut N, Pellegrina L, Alessi MC, Visvikis S, Tiret L, Juhan-Vague I. Biological and genetic factors influencing plasma factor VIII levels in a healthy family population: results from the Stanislas cohort. Br J Haematol 2004/12/21 ed. 2005; 128: 91–9.

44 Marchetti G, Lunghi B, Legnani C, Cini M, Pinotti M, Mascoli F, Bernard F. Contribution of low density lipoprotein receptor-related protein genotypes to coagulation factor VIII levels in thrombotic women. Haematologica 2006; 91: 1261–3.

45 Vormittag R, Bencur P, Ay C, Tengler T, Vukovich T, Quehenberger P, Mannhalter C, Pabinger I. Low-density lipoprotein receptor-related protein 1 polymorphism 663 C > T affects clotting factor VIII activity and increases the risk of venous thromboembolism. J Thromb Haemost 2006/12/13 ed. 2007; 5: 497–502.

46 Berger M, Moscatelli H, Kulle B, Luxembourg B, Blouin K, Spannagl M, Lindhoff-Last E, Schambeck CM. Association of ADAMDEC1 haplotype with high factor VIII levels in venous thromboembolism. Thromb Haemost 2008/05/02 ed. 2008; 99: 905–8.

47 Smith NL, Chen MH, Dehghan A, Strachan DP, Basu S, Soranzo N, Hayward C, Rudan I, Sabater-Lleal M, Bis JC, de Maat MP, Rumley A, Kong X, Yang Q, Williams FM, Vitart V, Campbell H, Malarstig A, Wiggins KL, Van Duijn CM, et al. Novel associations of multiple genetic loci with plasma levels of factor VII, factor VIII, and von Willebrand factor: The CHARGE (Cohorts for Heart and Aging Research in Genome Epidemiology) Consortium. Circulation 2010/03/17 ed. 2010; 121: 1382–92.

48 Yamniuk, A.P. and Vogel HJ. Insights into the structure and function of calcium- and integrin-binding proteins. Calcium Bind Proteins 2006; 1: 150–5.

49 Naik UP, Naik MU. Association of CIB with GPIIb/IIIa during outside-in signaling is required for platelet spreading on fibrinogen. Blood 2003; 102: 1355–62.

50 Huang H, Ishida H, Yamniuk AP, Vogel HJ. Solution structures of Ca2+-CIB1 and Mg2+-CIB1 and their interactions with the platelet integrin alphaIIb cytoplasmic domain. The Journal of biological chemistry 2011; 286: 17181–92.

51 Huang H, Vogel HJ. Structural basis for the activation of platelet integrin αIIbβ3 by calcium- and integrin-binding protein 1. Journal of the American Chemical Society 2012; 134: 3864–72.

52 Naik MU, Nigam A, Manrai P, Millili P, Czymmek K, Sullivan M, Naik UP. CIB1 deficiency results in impaired thrombosis: the potential role of CIB1 in outside-in signaling through integrin alpha IIb beta 3. Journal of thrombosis and haemostasis : JTH 2009; 7: 1906–14.
53 Fang X, Chen C, Wang Q, Gu J, Chi C. The interaction of the calcium- and integrin-binding protein (CIBP) with the coagulation factor VIII. Thromb Res 2001/04/27 ed. 2001; 102: 177–85.
54 Gentry HR, Singer AU, Betts L, Yang C, Ferrara JD, Sondek J, Parise L V. Structural and biochemical characterization of CIB1 delineates a new family of EF-hand-containing proteins. J Biol Chem 2004/12/03 ed. 2005; 280: 8407–15.
55 Denofrio JC, Yuan W, Temple BR, Gentry HR, Parise L V. Characterization of calcium- and integrin-binding protein 1 (CIB1) knockout platelets: potential compensation by CIB family members. Thrombosis and haemostasis 2008; 100: 847–56.
56 Loirand G, Scalbert E, Bril A, Pacaud P. Rho exchange factors in the cardiovascular system. Current opinion in pharmacology 2008; 8: 174–80.
57 Pleines I, Hagedorn I, Gupta S, May F, Chakarova L, van Hengel J, Offermanns S, Krohne G, Kleinschnitz C, Brakebusch C, Nieswandt B. Megakaryocyte-specific RhoA deficiency causes macrothrombocytopenia and defective platelet activation in hemostasis and thrombosis. Blood 2012; 119: 1054–63.
58 Aslan JE, McCarty OJT. Rho GTPases in platelet function. Journal of thrombosis and haemostasis : JTH 2013; 11: 35–46.
59 Meisinger C, Prokisch H, Gieger C, Soranzo N, Mehta D, Rosskopf D, Lichtner P, Klopp N, Stephens J, Watkins NA, Deloukas P, Greinacher A, Koenig W, Nauck M, Rimmbach C, Völzke H, Peters A, Illig T, Ouwehand WH, Meitinger T, et al. A genome-wide association study identifies three loci associated with mean platelet volume. American journal of human genetics 2009; 84: 66–71.
60 Soranzo N, Spector TD, Mangino M, Kühnel B, Rendon A, Teumer A, Willenborg C, Wright B, Chen L, Li M, Salo P, Voight BF, Burns P, Laskowski RA, Xue Y, Menzel S, Altshuler D, Bradley JR, Bumpstead S, Burnett M-S, et al. A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. Nature genetics 2009; 41: 1182–90.
61 Gieger C, Radhakrishnan A, Cvejic A, Tang W, Porcu E, Pistis G, Serbanovic-Canic J, Elling U, Goodall AH, Labrune Y, Lopez LM, Mägi R, Meacham S, Okada Y, Pirastu N, Sorice R, Teumer A, Voss K, Zhang W, Ramirez-Solis R, et al. New gene functions in megakaryopoiesis and platelet formation. Nature 2011; 480: 201–8.
62 Li J, Glessner JT, Zhang H, Hou C, Wei Z, Bradfield JP, Mentch FD, Guo Y, Kim C, Xia Q, Chiavacci RM, Thomas KA, Qiu H, Grant SFA, Furth SL, Hakonarson H, Sleiman PMA. GWAS of blood cell traits identifies novel associated loci and epistatic interactions in Caucasian and African-American children. Human molecular genetics 2013; 22: 1457–64.
63 Matsushita T, Ashikawa K, Yonemoto K, Hirakawa Y, Hata J, Amitani H, Doi Y, Ninomiya T, Kitazono T, Ibayashi S, Iida M, Nakamura Y, Kiyohara Y, Kubo M. Functional SNP of ARHGEF10 confers risk of atherothrombotic stroke. Human molecular genetics 2010; 19: 1137–46.
64 Yin Y-Y, Zhang B, Zhou M-K, Guo J, Lei L, He X-H, Xu Y-M, He L. The functional SNP rs4376531 in the ARHGEF gene is a risk factor for the atherothrombotic stroke in Han Chinese. Neurology India 59: 408–12.
65 Kim K, Yang YJ, Kim K, Kim MK. Interactions of single nucleotide polymorphisms with dietary calcium intake on the risk of metabolic syndrome. The American journal of clinical nutrition 2012; 95: 231–40.
66 Clark ME, Kelner GS, Turbeville LA, Boyer A, Arden KC, Maki RA. ADAMTS9, a novel member of the ADAM-TS/ metallospondin gene family. Genomics 2000; 67: 343–50.
67 Bruchova H, Vasikova A, Merkerova M, Milcova A, Topinka J, Balascak I, Pastorkova A, Sram RJ, Brdicka R. Effect of maternal tobacco smoke exposure on the placental transcriptome. Placenta 2010/01/23 ed. 2010; 31: 186–91.
68 Zheng XL. Structure-function and regulation of ADAMTS-13 protease. Journal of thrombosis and haemostasis : JTH 2013; 11 Suppl 1: 11–23.
69 Yoshina S, Sakaki K, Yonezumi-Hayashi A, Gengyo-Ando K, Inoue H, Iino Y, Mitani S. Identification of a novel ADAMTS9/GON-1 function for protein transport from the ER to the Golgi. Molecular biology of the cell 2012; 23: 1728–41.

Artículo 3

Título

The Central Role of KNG1 Gene as a Genetic Determinant of Coagulation Pathway-Related Traits: Exploring Metaphenotypes.

Autores

Helena Brunel, Raimon Massanet, Angel Martinez-Perez, Andrey Ziyatdinov, Laura Martin-Fernandez, Juan Carlos Souto, Alexandre Perera y José Manuel Soria.

Referencia

PloS one 2016. 11(12): e0167187. doi: 10.1371/journal.pone.0167187. PMID: 28005926.

Resumen

Los estudios genéticos tradicionales que analizan fenotipos individualmente pueden no tener suficiente poder estadístico para detectar fenómenos pleiotrópicos implicados en las enfermedades complejas, como la VTE. Teniendo en cuenta las correlaciones existentes entre los distintos fenotipos que participan en un mismo proceso biológico, hemos desarrollado una nueva metodología para aplicar el GWAS en grupos de fenotipos relacionados entre sí e involucrados en la cascada de la coagulación sanguínea. Esta metodología se divide en dos fases, consistiendo inicialmente en la creación de nuevas variables denominadas “metafenotipos” a partir de combinaciones lineales de los fenotipos originales. Este proceso se ha llevado a cabo a partir del análisis de componentes independientes (ICA). Por lo tanto, dichos “metafenotipos” integrarán la información del proceso biológico en el que estén implicados. En la

segunda fase, estos “metafenotipos” se han utilizado en el GWAS realizado en el Proyecto GAIT-1 para detectar los factores genéticos que influyen en su variabilidad. Como resultado de este trabajo se han obtenido 15 “metafenotipos” con heredabilidades estadísticamente significativas que varían del 20% al 70%, lo que demuestra la relevancia de su base genética. De éstos, 4 muestran asociaciones significativas con SNPs y cobra especial interés una de las regiones que abarca los genes *HRG*, *FETUB* y *KNG1*. En concreto, se destaca el papel de *KNG1* como determinante genético de la cascada de la coagulación y del riesgo de VTE. La metodología presentada en este trabajo permite capturar la información de distintos fenotipos relacionados entre sí en “metafenotipos” para esclarecer nuevos mecanismos genéticos implicados en las enfermedades complejas.

Material suplementario

Tablas suplementarias de la publicación (ver anexo II).

PLOS ONE

RESEARCH ARTICLE

The Central Role of *KNG1* Gene as a Genetic Determinant of Coagulation Pathway-Related Traits: Exploring *Metaphenotypes*

Helena Brunel[1], Raimon Massanet[2], Angel Martinez-Perez[1], Andrey Ziyatdinov[1], Laura Martin-Fernandez[1], Juan Carlos Souto[3], Alexandre Perera[3], José Manuel Soria[1]*

1 Unit of Genomics of Complex Diseases, Sant Pau Institute of Biomedical Research (IIB-Sant Pau), Barcelona, Spain, **2** B2SLab, Departament d'Enginyeria de Sistemes, Automàtica i Informàtica Industrial, Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, **3** Thrombosis and Haemostasis Unit, Sant Pau Institute of Biomedical Research (IIB-Sant Pau), Barcelona, Spain

* jsoria@santpau.cat

OPEN ACCESS

Citation: Brunel H, Massanet R, Martinez-Perez A, Ziyatdinov A, Martin-Fernandez L, Souto JC, et al. (2016) The Central Role of *KNG1* Gene as a Genetic Determinant of Coagulation Pathway-Related Traits: Exploring *Metaphenotypes*. PLoS ONE 11(12): e0167187. doi:10.1371/journal.pone.0167187

Editor: Shenying Fang, University of Texas MD Anderson Cancer Center, UNITED STATES

Received: July 25, 2016

Accepted: November 9, 2016

Published: December 22, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This study was supported by funds from the Instituto de Salud Carlos III Fondo de Investigación Sanitaria PI 11/0184 and PI 14/00582, Red Investigación Cardiovascular RD12/0042/0032 and AGAUR 2009 SGR 1147 from Generalitat de Catalunya. Laura Martin-Fernandez was supported by Ayudas Predoctorales de Formación en Investigación en Salud (PFIS) FI12/

Abstract

Traditional genetic studies of single traits may be unable to detect the pleiotropic effects involved in complex diseases. To detect the correlation that exists between several phenotypes involved in the same biological process, we introduce an original methodology to analyze sets of correlated phenotypes involved in the coagulation cascade in genome-wide association studies. The methodology consists of a two-stage process. First, we define new phenotypic meta-variables (linear combinations of the original phenotypes), named *metaphenotypes*, by applying Independent Component Analysis for the multivariate analysis of correlated phenotypes (i.e. the levels of coagulation pathway–related proteins). The resulting *metaphenotypes* integrate the information regarding the underlying biological process (i.e. thrombus/clot formation). Secondly, we take advantage of a family based Genome Wide Association Study to identify genetic elements influencing these *metaphenotypes* and consequently thrombosis risk. Our study utilized data from the GAIT Project (Genetic Analysis of Idiopathic Thrombophilia). We obtained 15 *metaphenotypes*, which showed significant heritabilities, ranging from 0.2 to 0.7. These results indicate the importance of genetic factors in the variability of these traits. We found 4 *metaphenotypes* that showed significant associations with SNPs. The most relevant were those mapped in a region near the *HRG*, *FETUB* and *KNG1* genes. Our results are provocative since they show that the *KNG1* locus plays a central role as a genetic determinant of the entire coagulation pathway and thrombus/clot formation. Integrating data from multiple correlated measurements through *metaphenotypes* is a promising approach to elucidate the hidden genetic mechanisms underlying complex diseases.

00322. This work has been partially supported by the Supported by the Spanish National Grants from Ministry of Economy and Competitiveness with grant TEC2014-60337-R, and the Generalitat de Catalunya, under the grant 2014 SGR 1063. CIBER-BBN is an initiative of the ISCIII.014 SGR 1063. CIBER-BBN is an initiative of the ISCIII. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

Considerable efforts have been invested to evaluate hundreds of genetic variants associated with human traits. Despite these efforts, the loci that have been identified only explain a small proportion of the total phenotypic variance. Thus, there is the question of where the remaining heritability resides. For a complex disease, such as thrombosis, traditional single-trait genetic studies may be unable to detect the pleiotropic effect that a given genetic variant could have on the intermediate phenotypes involved with the disease. In particular, the normal physiological process underlying thrombosis is complex and many of its components are involved in the coagulation and fibrinolysis pathways. These components form a collection of intermediate phenotypes that are generally measured in the study of thrombosis. These intermediate phenotypes may reflect more directly the effects from causal genes than disease status. They are also less genetically complex and more strongly associated with susceptibility loci.

So far, the genetic analyses of thrombosis have been carried out using one or more intermediate traits separately [1–7]. However, if a locus is associated with two or more traits, i.e. it is pleiotropic, a single-trait study may lose the power to detect this pleiotropic effect. However, finding disease risk indexes would contribute to a greater understanding of the pathogenesis of disease, and ultimately will develop better diagnostic, prevention and treatment strategies. In addition, the simultaneous analyses of multiple traits may uncover regulating elements such as master regulators or variants belonging to transcription factor binding sites. Genetic analyses have been performed using aPTT (Activated Partial Thromboplastin Time) as a phenotype to improve the understanding of the biological mechanisms underlying thrombotic disease [8,9]. Although aPTT measures the combined activity of several clotting factors in the intrinsic and common coagulation pathways [10] (including factors FII, FV, FVIII, FIX, FX, FXI and FXII), the present genetic studies on aPTT consider it as an univariate model without considering pleiotropic effects [11]. Another example of exploiting the genetic information of different traits comes from the GAIT (Genetic Analysis of Idiopathic Thrombophilia) Project, where we demonstrated that coagulation factors FVIII and vWF are genetically correlated with thrombotic disease [12]. Also, in a previous study, we identified common variants associated with the plasma levels of several proteins and consequently the risk of thrombosis [13]. However, the pleiotropic effects of loci in the coagulation cascade have not been explored fully.

Both genetic association and linkage research have focused on statistical and computational techniques to investigate the genetic effects between one genotype and one phenotype including polygenic and multiphenotypic approaches. Several strategies have been applied for the analysis of multiple and correlated traits. These can be divided into three categories: p-value correction methods, regression models and data reduction methods. P-value correction methods consist on combining several univariate tests, one for each trait, accounting for the observed correlational structure of the traits [14,15].

Regression models make use of mixed effects models for modelling the covariance structure of the phenotypes, as well as population structure [16]. These two approaches have a limited practical use since with a large number of correlated traits, they require the simultaneous estimation of too many parameters [17]. As an alternative, data reduction methods based on the transformation of the original traits to a reduced number of canonical traits have been proposed [18–20] with the intent of applying the traditional single trait analyses to these new variables. Generally, the canonical variables are obtained through a given mathematical model that transforms the original phenotypic data in a new space of reduced dimensionality where the new coordinate axes (also called components) define new phenotypic quantities obtained synthetically. In particular, Principal Components Analysis (PCA) has been applied for this purpose [17, 21, 22].

In this study, we explore an original methodology to determine the inner correlation within a set of related traits involved in the coagulation cascade, to help understanding the genetic bases of the coagulation cascade consequently of thrombosis risk. We apply Independent Component Analysis, a data reduction method, original in this field, to derive new phenotypic variables, called *metaphenotypes*, which integrate information regarding the underlying biological variability on the thrombus/clot formation. Then, we take advantage of our GWAS to identify genetic elements influencing these *metaphenotypes* and their relationship with thrombosis risk.

Materials and Methods

The GAIT Project

The GAIT (Genetic Analysis of Idiopathic Thrombophilia) Project has been described in Souto et al 2000 [13]. Briefly, the GAIT Project included 398 individuals from 21 extended Spanish families (mean pedigree size = 19) [12]. Twelve of these families were selected on the basis of a proband with idiopathic thrombophilia, whereas the remaining nine families were unaffected and selected randomly. The ages of the subjects ranged from <1 to 88 years (mean = 37.7 years) and the male to female sex ratio was 0.85. The study was performed according to the Declaration of Helsinki. All procedures were reviewed by the Institutional Review Board of the Hospital de la Santa Creu i Sant Pau, Barcelona, Spain. Adult subjects gave written consent for themselves and for their minor children.

Genotypes and Data Cleaning. A genome-wide set of 307,984 SNPs was typed for all of the participants using the Infinium® 317k Beadchip on the Illumina platform (San Diego, CA, USA). Individuals with a low call rate (<0.5%), a too high IBS (>0.95%) and a too high heterozygosity (FDR <1%) were removed from the sample. In addition, markers with a low call rate (<0.95%) and a low MAF (<0.0064%) were discarded also. A total of 34 individuals and 30,793 SNPs were removed from the study. A clean dataset containing $n = 364$ individuals and 277,191 SNPs was obtained for further analyses. This procedure was implemented in R using the GenABEL package [23].

Phenotypes. Among the 80 phenotypes in the GAIT sample, $m = 27$ phenotypes involved in the coagulation pathway were selected to study their joint biological activity within this metabolic process. These phenotypes were selected as they are defined in the literature [24]. The original phenotypes are described in S1 Table.

To properly apply the mathematical methods that we used, phenotypic data were freed of missing values. To guarantee this condition, the phenotypic dataset was imputed using a bPCA, a Bayesian method for missing value imputation [25].

"Metaphenotypes" as a Concept

A *metaphenotype* is defined as a new phenotypic variable obtained synthetically from a set of traits (phenotypes) using a given mathematical model of dimensionality reduction. *Metaphenotypes* should be able to capture the original structure of the data to describe them as a whole. Therefore, identifying genetic variants related to these *metaphenotypes* may help to ascertain the genetic bases of the observed variability of the set of phenotypes, here the coagulation pathway.

The coagulation factors in the coagulation cascade show related patterns of activity. It is known that the genes coding for the different coagulation factors share a joint ancestry [13], so there may exist also some regulatory elements jointly regulating their activity. We consider analyzing the 27 coagulation pathway-related phenotypes measured in the GAIT project under the concept of *metaphenotypes*.

Metaphenotypes are computed from the correlation among factors. There are several algorithms in the literature that are able to decompose the variability under different criteria. We applied an ICA (Independent Component Analysis), an algorithm based on a criterion of minimum shared information. ICA was compared to PCA(Principal Component Analysis) a reference method for studying the genetic association of correlated phenotypes [18, 21].

Statistical Analyses

Both PCA and ICA methods apply a linear transformation to the original phenotypic data and obtain a new system of coordinates of reduced dimensionality, following the expression in Eq 1.

$$X = M \cdot W + E \quad (1)$$

where X (n × m) are the original phenotypes, M (n × m) are the *metaphenotypes* and W (m × m) are the weights of the model and E is the error of the model. Note that the *metaphenotypes* correspond to the axes of the new system of coordinates, and are called "components". The maximum number of components obtained is the same as the original phenotypes but generally, only a few of them are informative and therefore are taken into account.

The *metaphenotypes* are determined by the characterization of the weights of this linear transformation, either using PCA or using ICA.

Independent Component Analysis. In ICA, the weights W are optimized to guarantee the statistical independence of the *metaphenotypes*. The independence of the components is guaranteed by finding W that maximizes the non-gaussianity of the *metaphenotypes* (M).

Among the several ICA algorithms, the fastICA procedure was applied, using a particular approximation of the negentropy measure for maximizing the nongaussianity [26]. In particular, this method was applied with an optimal number of *metaphenotypes* (components) of $k = 15$, according to a criterion based on cross-validation approximations [27]. As other ICA implementations, fastICA previously applies a PCA to the data in order to ensure that the components are uncorrelated. The number of components to be used are then determined from the PCA procedure using a cross-validation model.

Principal Component Analysis. In PCA, the weights W are optimized so that the *metaphenotypes* capture the maximum covariance existing between the original phenotypes. In this case, the *metaphenotypes* explore the correlation that exists among the original traits to capture the variability shared by the collection of original phenotypes.

PCA was used as a reference method. By default, PCA obtained as many components as original variables (k = 27).

Differences between PCA and ICA. As the structure of the interrelations among phenotypes is hidden and unknown, both techniques are complementary to unravel the cascade of physiological relationships.

PCA and ICA answer different biological questions.

PCA obtains *metaphenotypes* that explain the greatest overall variability or correlation between the original phenotypes. In other words, *metaphenotypes* built with PCA are a new set of indexes of jointly altered levels of the original phenotypes, capturing the common activity of the original phenotypes.

In contrast, ICA was chosen because it obtains *metaphenotypes* that are statistically independent. Thus, ICA is able to separate the different (independent) sources of variability captured by the original phenotypes. Let us consider the original traits as statistical mixtures of different sources of variability (genetic, environmental, or experimental). If there was a genetic source of variability captured by the set of phenotypes (pleiotropy), the *metaphenotypes*

obtained using ICA will capture it. In other words, ICA is especially useful to detect pleiotropic effects.

Heritability Estimation. The heritabilities of the *metaphenotypes* were estimated using the variance component method implemented in SOLAR [28]. This method partitions the total phenotypic variance into a proportion due to polygenic (additive) effects and a proportion due to environmental effects. The heritability (*h2r*) estimates the total variance of a trait due to additive genetic effects.

Genetic Association. Genome-Wide Association Analyses with the SNPs of the GAIT project were performed using a Likelihood Ratio test based on a linear mixed effects (Variance Components) model described in Eq 2.

The model provides a vector of fitted values of the phenotype and an estimate of the variance-covariance matrix for each family [28, 29, 30].

The polygenic mixed model defined in Eq 2 was applied for each *metaphenotype* M_i with the age and gender co-variables for testing the association as they present a significant correlation with almost all of the *metaphenotypes*.

$$M_i \sim \mu + \Sigma_j \beta_j c_{ji} + G_i + \varepsilon_j \quad (2)$$

where i is the individual index, M_i is the *metaphenotype*, μ is the overall mean, β_j is the regression coefficient of the j-th covariate, c_{ji} is the j-th covariate, G_i is the random additive polygenic effect (breeding value) which variance is defined as $\Phi\sigma_G$ where Φ is the kinship matrix and σ_G is the additive genetic variance due to polygenes. Finally ε_i are the residuals of the model.

With a comparative purpose, GWAS of the original phenotypes were also computed.

All the p-values where corrected using the Bonferroni critertion, with a significance criterion set at $\alpha = 0.05$ after adjustment.

Results

A total of 15 *metaphenotypes* were obtained with our methodology. All of the *metaphenotypes* showed a significant heritability ranging from 0.15 to 0.7 (Table 1). Significant findings obtained in GWAS are shown in Table 2. To illustrate the relevance of these findings, they were compared with metaphenotypes obtained with a PCA-based approach and with univariate GWAS applied to the original phenotypes. Heritabilities of PCA-based metaphenotypes are shown in S2 Table. Table 2 presents SNPs significantly associated with ICA-based metaphenotypes in comparison with PCA-based *metaphenotypes* and univariate phenotypes. For two particular SNPs (*rs9898* and *rs27311672*), concordant results were found among the three GWAS approaches. Both SNPs were significantly associated with both an ICA-based and a PCA-based *metaphenotype* as well as with the univariate phenotypes corresponding to the proteins coded by their respective closest gene (*HRG* and *F12*). Concordant ICA-based and PCA-based metaphenotypes were compared as follows. Associations p-values with all the original phenotypes with the SNPs reported in Table 2 are presented in S3 Table.

To obtain a clear and interpretable view of *metaphenotypes*, we plotted them in a simple graph (Fig 1). Non-directed graphs were used to express the existing interaction among the 27 original phenotypes, represented by nodes whose colors represent their weights in the resulting *metaphenotype*. This is interpretable as the contribution of the original phenotype to the corresponding *metaphenotype*. Numerical values for the weights are included in S4 and S5 Tables.

It is observed in Table 2 that ICA and PCA obtained concordant in two cases. For instance, SNPs *rs2731672* and *rs9898* were significantly associated with metaphenotypes coming from different methodologies.

Table 1. Heritabilities of ICA-based metaphenotypes (components 1 to 15 from the ICA model).

Metaphenotype	h2r
C1	0.48***
C2	0.17*
C3	0.53***
C4	0.15*
C5	0.22*
C6	0.61***
C7	0.24**
C8	0.55***
C9	0.35***
C10	0.7***
C11	0.45***
C12	0.58***
C13	0.32***
C14	0.24***
C15	0.59***

Significant thresholds for heritability estimation:
* <0.05,
**<0.005,
*** <0.0005

doi:10.1371/journal.pone.0167187.t001

It is observed that the two metaphenotypes significantly associated with SNP *rs2731672* (ICA-C10 and PCA-C10) are influenced clearly by the trait corresponding to the FXII levels (dark nodes in Fig 1.b and 1.d). This SNP is an intergenic variant ~5.8kb upstream of the *F12* gene. In both cases the FXII levels have an important loading in the metaphenotypes indicating the variability captured by the *metaphenotype* is due highly to the variability in the FXII levels. As expected, this SNP was also significantly associated with the FXII levels with the univariate GWAS approach.

The *metaphenotypes* ICA-C3 and PCA-C9, significantly associated with SNP *rs9898* are shown in Fig 1.a and 1.c. SNP *rs9898* is a nonsynonymous SNP in exon 5 of the *HRG* gene. While the PCA-based *metaphenotype* is oriented clearly to the HRG trait due to the weight of HRG levels in the *metaphenotype* (dark red HRG node in Fig 1.c), this specific trait does not present a high weighting value in the ICA-based *metaphenotype* (Fig 1.a).

Table 2. GWAS significant SNPs for the three approaches (univariate phenotypes, ICA-based metaphenotypes and PCA-based metaphenotypes. For each SNP, the Chromosome where it is located, its physically closest gene and its MAF are shown as well as the adjusted p-value.

SNP ID	Chr	Gene	MAF	HRG	FXII	P-value						
						ICA—C3	ICA—C4	ICA—C5	*ICA—C10*	PCA—C8	PCA—C9	PCA—C10
rs9898	3	HRG	0.35	1.9×10^{-16}		9×10^{-18}				1×10^{-07}	4.3×10^{-08}	
rs3733159	3	FETUB	0.34	3.3×10^{-13}		6.6×10^{-09}						
rs1621816	3	KNG1	0.24	1.5×10^{-09}		5×10^{-08}						
rs1403694	3	KNG1	0.32	1.1×10^{-08}		6.7×10^{-07}						
rs17255413	3	BOC	0.007				2.6×10^{-08}					
rs3113727	4	COL25A1	0.24					3.8×10^{-07}				
rs27311672	5	F12	0.17		7.6×10^{-36}				1.1×10^{-14}			1.5×10^{-11}

doi:10.1371/journal.pone.0167187.t002

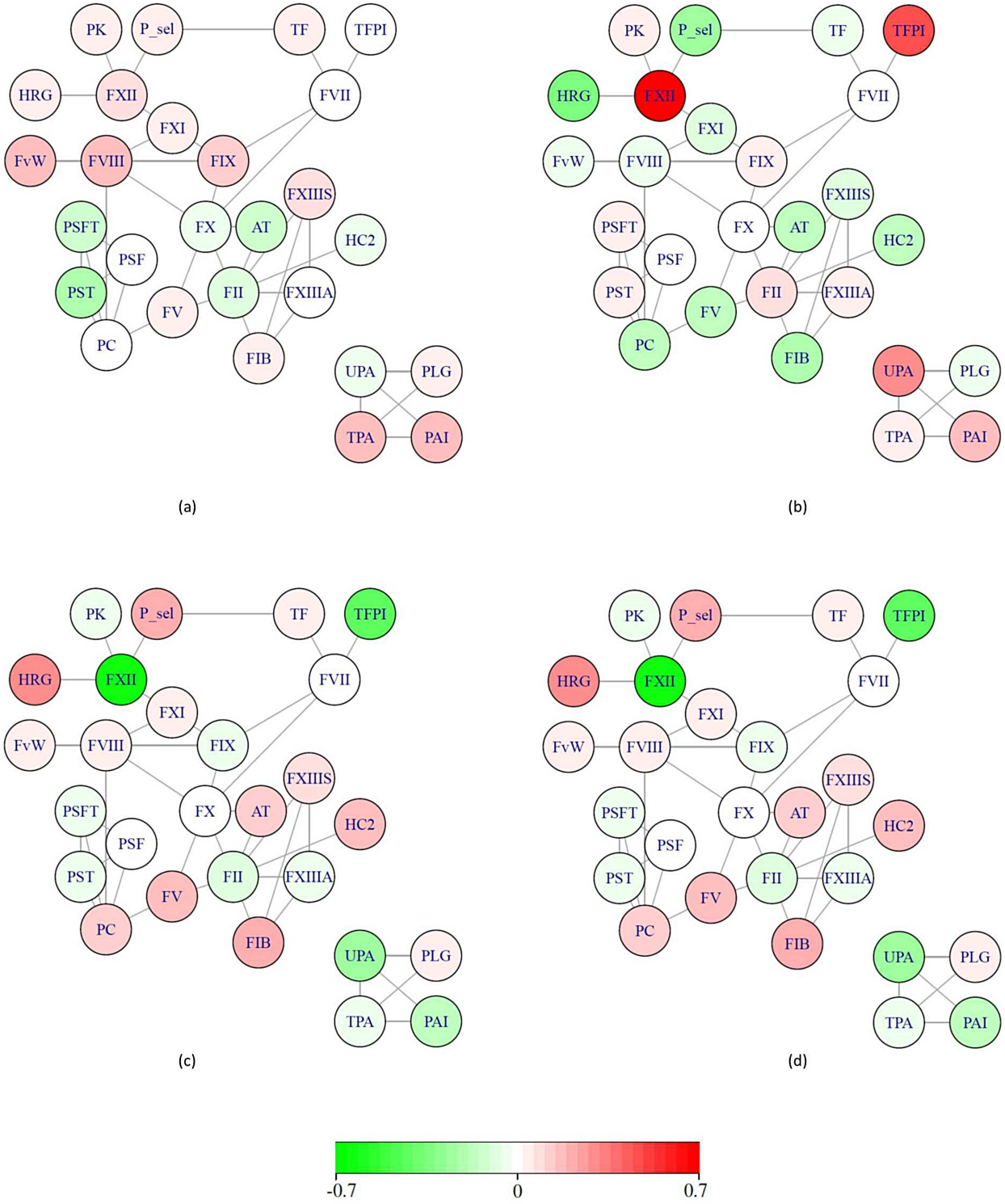

Fig 1. Metaphenotype graphical representation using a simple graph. (a) ICA-based metaphenotype corresponding to the 3rd component (ICA-C3), (b) ICA-C10, (c) PCA-C9 (d) PCA-C10.

doi:10.1371/journal.pone.0167187.g001

In addition, Fig 2 compares directly both *metaphenotypes* in terms of their loadings (the weight of each trait on the *metaphenotypes*) and their scorings (the projection of each individual on the *metaphenotypes*). For the metaphenotypes associated with SNP *rs2731672* (ICA-C10 and PCA-C10) (Fig 2.a), a clear correlation between both loadings and scorings from both *metaphenotypes* was observed. This confirms that the common variability captured by both *metaphenotypes* is the same in both cases and is due highly to the variability of the FXII. By contrast, as shown in Fig 2.b, no correlation was observed between the loadings or the scoring of the metaphenotypes associated with SNP *rs9898* (ICA-C3 and PCA-C9). This indicates that both metaphenotypes capture different information from the original phenotypes.

In addition, the ICA-C3 metaphenotype was significantly associated with three other SNPs (*rs3733159*, *rs1621816* and *rs1403694*) on Chromosome 3. The former one corresponds to an intronic SNP in the *FETUB* gene, whereas the latter two are intronic SNPs in the *KNG1* gene. It is important to note that *KNG1* is located at a distance of around 40Kb from *HRG*. However, the SNPs *rs1621816* and *rs1403694* in the *KNG1* gene showed a low amount of Linkage Disequilibrium with the SNP *rs9898* in the *HRG* gene ($r^2 = 0.22$ and $r^2 = 0.21$). These four SNPs showed a significant association with the HRG trait with the univariate GWAS approach.

For the metaphenotypes associated with SNP *rs2731672* (Fig 2.a), a clear correlation between both loadings and scorings from both *metaphenotypes* was observed. This confirms that the common variability captured by both *metaphenotypes* is the same in both cases and is due highly to the variability of the FXII. By contrast, as shown in Fig 2.b, no correlation was observed between the loadings or the scoring indicating that both metaphenotypes capture different information from the original phenotypes.

Discussion

To date, there are no genetic studies of the coagulation pathway as a whole. Since single-trait genetic studies explain only a small proportion of the phenotypic variability of thrombotic disease, it is prudent to explore other sources of heritability, such as pleiotropy. In our study, we propose a methodology to capture the correlation that exists between a set of intermediate phenotypes involved in the coagulation cascade to elucidate the hidden genetic causes of thrombosis. For doing that, we introduced the concept of *metaphenotype* consisting on new phenotypic indices that gather the observed variability of a collection of related phenotypes. *Metaphenotypes* are obtained through mathematical models of data dimensionality reduction. In this study, we applied ICA for the *metaphenotype* construction. This method is original in this field and was compared to PCA, a reference method for the combined analysis of correlated traits in genetic linkage and association studies[18, 19, 22]. ICA was chosen because it is especially useful to detect pleiotropy. By contrast, PCA is characterized by being able to capture the common variability existing among the phenotypes. Because they answer different biological questions, both methodologies may be complementary.

Metaphenotypes were obtained from a collection of coagulation-related phenotypes from the GAIT project[12] with the aim of identifying genetic variants underlying the whole biological process of blood coagulation. The final goal was to propose genetic markers as candidate regulators of the coagulation cascade and consequently of thrombosis risk.

Even if *metaphenotypes* are not intuitively informative, biologically speaking, they may enable the identification of possibly important loci for a more integrated coagulation index and thus be able to characterize the genetic baseline of coagulation function or thrombosis risk. *Metaphenotypes* can be graphically represented by use of simple graphs (Fig 1) where the original phenotypes involved in their construction are represented by nodes whose colors represent their weights in the resulting *metaphenotype*. This is interpretable as the contribution of

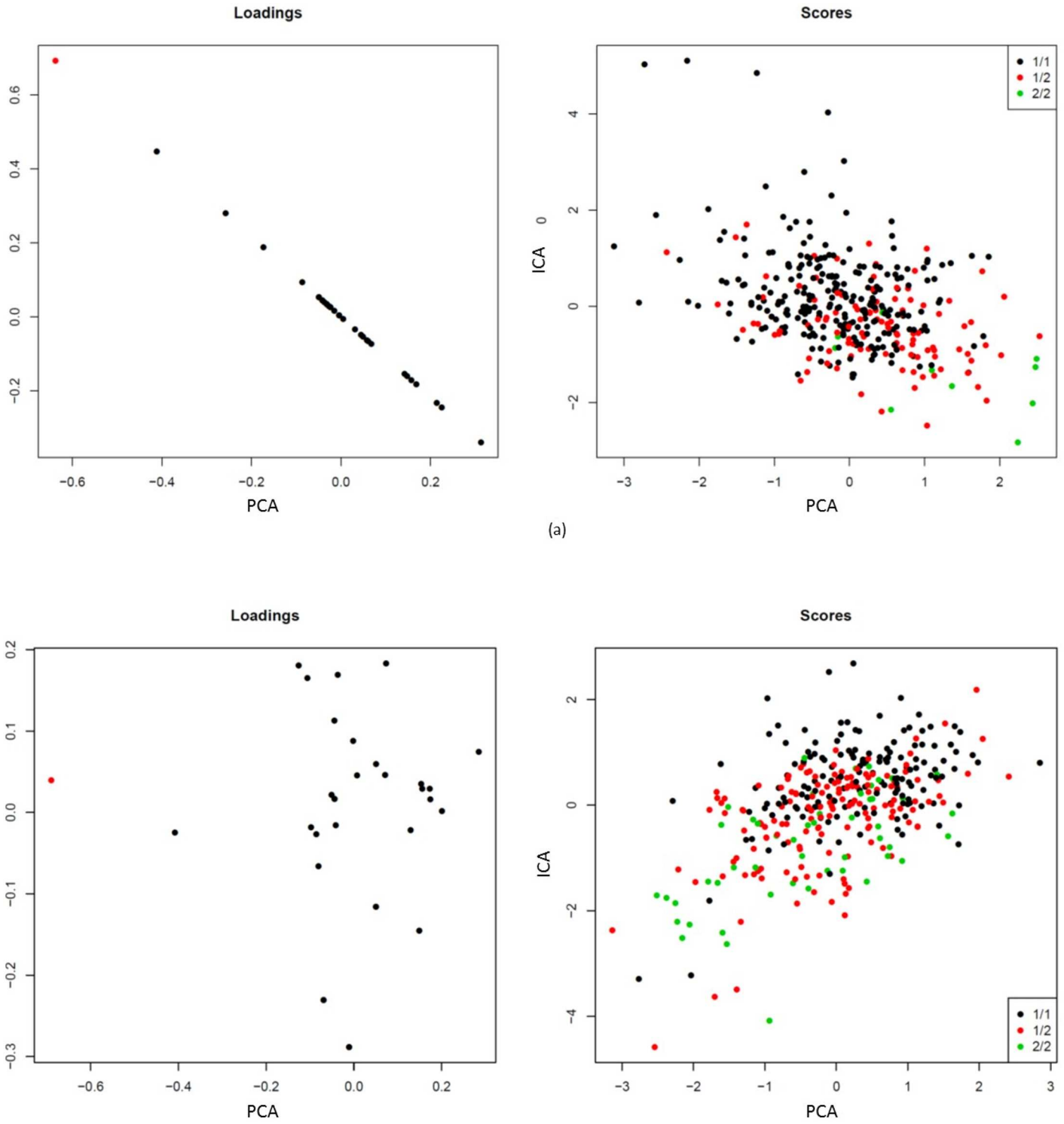

Fig 2. Comparison between the metaphenotypes obtained with both PCA and ICA models. (a) metaphenotype associated with the SNP *rs2731672* at the *F12* locus and (b) metaphenotypes associated with the SNP *rs9898* at the *HRG* locus.

doi:10.1371/journal.pone.0167187.g002

the original phenotypes to the corresponding *metaphenotype*. In addition, obtaining metaphenotypes with a heritability greater than 50% (Table 1), allows to assume that the variability of these new phenotypic entities may be highly due to genetic variants. This justifies performing GWAS to *metaphenotypes*.

Results from GWAS with both ICA-based and PCA-based metaphenotypes were concordant in two cases. For instance, SNPs *rs2731672* and *rs9898* were significantly associated with metaphenotypes coming from different methodologies.

In both cases, we compared graphically the *metaphenotype* obtained with ICA and the one obtained with PCA (Fig 2).

As shown in Table 2, the SNP *rs2731672* in the *F12* locus on Chromosome 5 was significantly associated with an ICA-based *metaphenotype* (p-value of $1.1e^{-14}$) and with a PCA-based *metaphenotype* (p-value: $1.48x10^{-11}$). It is observed in Fig 1 that both the PCA-based *metaphenotype* (Fig 1.b) and the ICA-based *metaphenotype* (Fig 1.d) are influenced clearly by the FXII levels in blood. In addition, Fig 2.a shows a clear correlation between both loadings and scorings of both *metaphenotypes*. This suggests that the common variability captured by both *metaphenotypes* is the same in both cases and is due highly to the variability of the FXII. This observation is in agreement with the univariate association between this particular *locus* (encoding the structural *F12* gene) and FXII levels [1]. This result confirms that the ICA method also captures non-pleiotropic effects.

Secondly, SNP *rs9898* at the *HRG* locus at chromosome 3 was significantly associated with an ICA-based (p-value: $9e^{-18}$) and two PCA-based (pvalues: $1e^{-07}$ and $4.3e^{-08}$) *metaphenotypes*. Comparisons were carried out with the metaphenotype showing a lower p-value. In this case, no correlation was observed between the loadings or the scoring (Fig 2.b). This suggests that both *metaphenotypes* capture different information of the original traits. Thus, the biological interpretation of the results may be done separately. The univariate GWAS confirmed that SNP *rs9898* is associated with Histidine Rich Glycoprotein (HRG) levels, but previous results also reported that it was associated with Activated Prothrombin Time (aPTT) trait and consequently with thrombosis risk [8, 31]. This explains why, as observed in Fig 1.c, the *metaphenotype* obtained with PCA is oriented clearly to the HRG trait. However, the HRG levels do not have a high weighting value in the *metaphenotype* obtained using ICA (Fig 1.a). In other words, whereas the *metaphenotype* obtained with PCA captures the variance due to the more weighted trait (that is HRG), the result obtained through ICA extend our knowledge about the implication of this genetic variant, indicating that this locus has a pleiotropic effect on the set of coagulation-related traits involved with this *metaphenotype*.

In addition, the same ICA-based *metaphenotype*, showed a significant association with three other SNPs located in the same genomic region of *HRG* on chromosome 3 (rs3733159 at the *FETUB* locus and *rs1621816* and *rs1403694* at the *KNG1* locus). These 3 SNPs were also associated with HRG levels in univariate analyses. The proteins coded by these three genes are Histidine Rich Glycoprotein (HRG), Fetuin-B (FB) and the High Molecular Weight Kininogen (HMWK). All of these proteins are structurally related to a fourth protein, the fetuin A- Heremams Schmide-glycoprotein [32]. Together, they form a subgroup (denoted type 3) within the cystatin superfamily of cysteine inhibitors. Among the several physiological roles associated to type 3 cystatins, the most relevant is the regulation of coagulation and platelet functions, controlled mainly by HRG and kininogen proteins. Furthermore, High-molecular-weight kininogen (HMWK) (encoded by *KNG1*), as well as coagulation Factor FXII (encoded by F12) are, together with prekallikrein (PK), important constituents of the plasma contact-kinin system. This system was first recognized as a surface-activated coagulation system, also known as the Coagulation Intrinsic Pathway (CIP). CIP is activated when blood or plasma interacts with artificial surfaces. A better understanding of this system may lead to insight into mechanisms

for thrombosis and, therefore, the contact-kinin system represents a promising multifunctional target for potential thromboembolic therapies, since blocking of distinct members of the kallikrein-kinin system has the potential to become an effective and safe strategy to combat cardiovascular diseases such as myocardial infarction.

Focusing on SNPs located at the *KNG1* locus we observed that SNPs *rs1621816* and *rs1403694* showed a low degree of linkage disequilibrium with the SNP *rs9898* at the *HRG* locus ($r^2 = 0.22$ and $r^2 = 0.21$). This indicates that they represent an distinct genetic signal. Thus, it is reasonable to suggest that among the results obtained, the association between SNPs located at the *KNG1* locus and an ICA-based *metaphenotype* may have more relevant biological and clinical implications. Our results indicate that KNG1 plays a relevant role in the CIP not only at a molecular level, but also at a genetic level. This result is particularly interesting since allelic variants in *KNG1* were previously associated with risk of thrombosis [33]. Our result strengthens previous conclusions concerning the association of *KNG1* with thrombosis suggesting that *KNG1* plays a role in the regulation of CIP, even without the influence of the FXI or the FXII levels, since neither FXI nor FXII levels show a specific weight within this *metaphenotype* (Fig 1.c).

In conclusion, the methodology proposed in this study complemented existing tools for detecting genetic associations in correlated phenotypes. This strategy explores the potential mechanisms and pathways underlying complex diseases and helps to interpret how they are associated with genetic variants. Our approach is based on the assumption that pleiotropy may occur in many complex diseases and more particularly in thrombosis diseases. The proposed mathematical approach is especially addressed to capture several aspects of the correlated activity of a set of original traits, here blood levels of the proteins involved in the coagulation cascade. Applying this original concept helped to identify two candidate SNPs in the *KNG1* gene susceptible to have an important role in the genetic regulation of the coagulation pathway as a whole and consequently of thrombosis disease.

Supporting Information

S1 Table. Description of the phenotypes.
(XLSX)

S2 Table. Heritabilities ofPCA-based metaphenotypes (components 1 to 27 from the PCA model). Significant thresholds for heritability estimation: * <0.05, **<0.005, *** <0.0005.
(XLSX)

S3 Table. P-values of associations of the relevant SNPs of this study with all the original phenotypes.
(XLSX)

S4 Table. Loadings (weights) of the coagulation phenotypes in the obtained ICA-based metaphenotypes.
(XLSX)

S5 Table. Loadings (weights) of the coagulation phenotypes in the obtained PCA-based metaphenotypes.
(XLSX)

Acknowledgments

We are deeply grateful to the families who participated in this study. Also, we would like to thank Professor Bill Stone for reviewing the manuscript.

Author Contributions

Conceptualization: HB RM AM-P JCS AP JMS.

Data curation: HB AM-P AZ AP.

Formal analysis: HB RM AM-P AZ AP.

Funding acquisition: AP JMS.

Investigation: JCS JMS.

Methodology: HB RM AM-P AZ LM-F JCS AP JMS.

Project administration: HB AP JMS.

Resources: JMS.

Software: HB RM AM-P AZ AP.

Supervision: AP JMS.

Validation: HB RM AM-P AZ LM-F JCS AP JMS.

Visualization: HB RM AZ.

Writing – original draft: HB.

Writing – review & editing: HB RM AM-P AZ LM-F JCS AP JMS.

References

1. Soria JM, Almasy L, Souto JC, Bacq D, Buil A, Faure A et al. A quantitative-trait locus in the human factor XII gene influences both plasma factor XII levels and susceptibility to thrombotic disease. Am. J. Hum. Genet. 2002; 70(3):567–74. PMID: 11805911
2. Soria JM, Almasy L, Souto JC, Sabater-Lleal M, Fontcuberta J, Blangero J. The F7 Gene and Clotting Factor VII Levels: Dissection of a Human Quantitative Trait Locus. Hum. Biol. 2009; 81(5):853–867.
3. Souto JC, Almasy L, Soria JM, Buil A Stone W, Lathrop M et al. Genome-wide linkage analysis of von Willebrand factor plasma levels: results from the GAIT Project. Thromb. Haemost. 2003; 89(3):468–74. PMID: 12624629
4. Athanasiadis G, Buil A, Souto JC, Borrell M, López S, Martinez-Perez A et al. A genome-wide association study of the Protein C anticoagulant pathway. PloS One. 2011; 6(12):e29168. doi: 10.1371/journal.pone.0029168 PMID: 22216198
5. Soria JM, Almasy L, Souto JC, Buil A, Lathrop M, Blangero J et al. A genome search for genetic determinants that influence plasma fibrinogen levels. Arterioscler. Thromb. Vasc. Biol. 2005; 25(6):1287–92. doi: 10.1161/01.ATV.0000161927.38739.6f PMID: 15761192
6. Viel KR, Machiah DK, Warren DM, Khachidze M, Buil A, Fernstrom K et al. A sequence variation scan of the coagulation factor VIII (FVIII) structural gene and associations with plasma FVIII activity levels. Blood. 2007; 109(9):3713–24. PMID: 17209060
7. Khachidze M, Buil A, Viel KR, Porter S, Warren D, Machiah DK et al. Genetic determinants of normal variation in coagulation factor (F) IX levels: genome-wide scan and examination of the FIX structural gene. J. Thromb. Haemost. 2006; 4(7):1537–45. PMID: 16839351
8. Park KJ, Kwon EH, Ma Y, Park IA, Kim SW, Kim SH et al. Significantly different coagulation factor activities underlying the variability of "normal" activated partial thromboplastin time. Blood Coagul. Fibrinolysis. 2012; 23(1):35–8. PMID: 22027757
9. Tang W, Schwienbacher C, Lopez LM, Ben-Shlomo Y, Oudot-Mellakh T, Johnson AD et al. Genetic associations for activated partial thromboplastin time and prothrombin time, their gene expression profiles, and risk of coronary artery disease. Am. J. Hum. Genet. 2012; 91(1):152–62. PMID: 22703881

10. Houlihan LM, Davies G, Tenesa A, Harris SE, Luciano M, Gow AJ et al. Common variants of large effect in F12, KNG1, and HRG are associated with activated partial thromboplastin time. Am. J. Hum. Genet. 2010; 86(4):626–31. PMID: 20303064

11. Tang W, Schwienbacher C, Lopez LM, Ben-Shlomo Y, Oudot-Mellakh T, Johnson AD et al., Genetic Associations for Activated Partial Thromboplastin Time and Prothrombin Time, their Gene Expression Profiles, and Risk of Coronary Artery Disease. Am J Hum Genet. 2012 July 13; 91(1): 152–162. PMID: 22703881

12. Souto JC, Almasy L, Borrell M, Blanco-Vaca F, Mateo J, Soria JM et al. Genetic susceptibility to thrombosis and its relationship to physiological risk factors: the GAIT Study. Am. J. Hum. Genet. 2000; 67:1452–9. PMID: 11038326

13. Souto JC, Almasy L, Blangero J, Stone W, Borrell M, Urrutia T et al., Genetic regulation of plasma levels of vitamin K-dependent proteins involved in hematostatis: results from the GAIT Project. Genetic Analysis of Idiopathic Thrombophilia. Thromb Haemost. 2001 Jan; 85(1):88–92. PMID: 11204594

14. Yang Q, Wu H, Guo CY, Fox CS. Analyze multivariate phenotypes in genetic association studies by combining univariate association tests. Genetic Epidemiology. 2010. 34:444–54. doi: 10.1002/gepi.20497 PMID: 20583287

15. Xu X, Tian L, Wei LJ. Combining dependent tests for linkage or association across multiple phenotypic traits. Biostatistics. 2003; 4(2):223–9. doi: 10.1093/biostatistics/4.2.223 PMID: 12925518

16. Stephens M. A unified framework for association analysis with multiple related phenotypes. PLoS ONE. 2013. 8:e65245. doi: 10.1371/journal.pone.0065245 PMID: 23861737

17. Knott SA, Haley CS. Multitrait Least Squares for Quantitative Trait Loci Detection. Genetics. 2000. 156 (2):899–911. PMID: 11014835

18. Klei L, Luca D, Devlin B, Roeder K. Pleiotropy and principal component of heritability combine to increase power for association analysis. Genetic Epidemiology. 2010. 34:444–54.

19. Mei H, Chen W, Dellinger A, He J, Wang M, Yau C et al. Principal-component-based multivariate regression for genetic association studies of metabolic syndrome components. BMC Genet. 2010. 11 (1):100.

20. Weller JI, Wiggans GR, VanRaden PM, Ron M. Application of a canonical transformation to detection of quantitative trait loci with the aid of genetic markers in a multi-trait experiment. Theor. Appl. Genet. 1996. 92(8):998–1002. doi: 10.1007/BF00224040 PMID: 24166627

21. Aschard H, Vilhjalmsson BJ, Greliche N, Morange PE, Tregouet DA, Kraft P. Maximizing the power of principal-component analysis of correlated phenotypes in genome wide association studies. American Journal of Human Genetics. 2014. 94:662–76. doi: 10.1016/j.ajhg.2014.03.016 PMID: 24746957

22. Mathias RA, Kim Y, Sung H, Yanek LR, Mantese VJ, Hererra-Galeano JE et al. A combined genome-wide linkage and association approach to find susceptibility loci for platelet function phenotypes in European American and African American families with coronary artery disease. BMC Med. Genomics. 2010. 3:22. doi: 10.1186/1755-8794-3-22 PMID: 20529293

23. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. GenABEL: an R library for genome-wide association analysis. Bioinformatics. 2007. 23(10):1294–6. doi: 10.1093/bioinformatics/btm108 PMID: 17384015

24. Lefkowitz Jerry B. "Coagulation pathway and physiology." An Algorithmic Approach to Hemostasis Testing. Northfield, IL: College of American Pathologists. 2008;3–12.

25. Stacklies W, Redestig H, Scholz M, Walther D, Selbig J. pcaMethods—a bioconductor package providing PCA methods for incomplete data. Bioinformatics. 2007. 23(9):1164–1167. doi: 10.1093/bioinformatics/btm069 PMID: 17344241

26. Hyvärinen A, Oja E. Independent component analysis: algorithms and applications. Neural Netw. 2000; (4–5:):411–30. PMID: 10946390

27. Josse J, Husson F. Selecting the number of components in principal component analysis using cross-validation approximations. Comput. Stat. Data Anal. 2012; 56(6):1869–1879.

28. Almasy L, Blangero J. Multipoint quantitative-trait linkage analysis in general pedigrees. Am. J. Hum. Genet. 1998; 62(5):1198–211. doi: 10.1086/301844 PMID: 9545414

29. Aulchenko Yurii S., de Koning Dirk-Jan, and Haley Chris. Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. Genetics. 2007; 177(1): 577–585. doi: 10.1534/genetics.107.075614 PMID: 17660554

30. Ziyatdinov A, Brunel H, Martinez-Perez A, Buil A, Perera A, Soria JM. solarius: an R interface to SOLAR for variance component analysis in pedigrees. Bioinformatics. 2016.

31. Morange PE, Oudot-Mellakh T, Cohen W,et al. KNG1 Ile581Thr and susceptibility to venous thrombosis. Blood. 2001; 117(13): 3692–3694.

32. Lee C., Bongcam-Rudloff E., Sollner C., Jahnen-Dechent W. and Claesson-Welsh L. Type 3 cystatins; fetuins, kininogen and histidine-rich glycoprotein. Frontiers in Bioscience. 2009; 14:2911–2922.

33. Sabater-Lleal M, Martinez-Perez A, Buil A, Folkersen L, Souto JC, Bruzelius M et al. A genome-wide association study identifies KNG1 as a genetic determinant of plasma factor XI level and activated Partial Thromboplastin Time. Arterioscler. Thromb. Vasc. Biol. 2012; 32:2008–2016. doi: 10.1161/ATVBAHA.112.248492 PMID: 22701019

Artículo 4

Título

Next Generation Sequencing to Dissect the Genetic Architecture of KNG1 and F11 Loci using Factor XI Levels as an Intermediate Phenotype of Thrombosis.

Autores

Laura Martin-Fernandez, Giovana Gavidia-Bovadilla, Irene Corrales, Helena Brunel, Lorena Ramírez, Sonia López, Juan Carlos Souto, Francisco Vidal y José Manuel Soria.

Referencia

PloS one 2016. Artículo sometido.

Resumen

La VTE es compleja y con una heredabilidad estimada del 61%. En base a los resultados publicados sobre el GWAS realizado en el Proyecto GAIT-1, en el cual se detectaron asociaciones estadísticamente significativas entre los niveles del FXI y diversos SNPs localizados en los genes *KNG1* y *F11*, nos hemos propuesto identificar la variabilidad genética del *KNG1* y *F11* implicada en la variación de los niveles del FXI. Para esto, se han secuenciado completamente ambos genes mediante NGS (plataforma MiSeq de Illumina), lo que implica la inclusión de exones, intrones y la región promotora, en un grupo de 110 individuos no relacionados genéticamente entre sí del Proyecto GAIT-2. Cabe destacar que, de los 110 individuos estudiados, se han seleccionado 40 individuos situados en las colas de distribución normal de los niveles de FXI como muestra de cribado. En este subgrupo se han identificado 762 variantes

genéticas que se han denominado “variantes genéticas de interés” para los posteriores análisis de asociación y de exploración de mutaciones patogénicas. Respecto a los análisis de asociación, en los que se han considerado los genotipos de dichas “variantes genéticas de interés” de la muestra total de 110 individuos para aumentar su poder estadístico, se han detectado diversas asociaciones estadísticamente significativas entre los niveles de FXI y variantes genéticas comunes o grupos de variantes genéticas de baja frecuencia alélica y raras mediante las herramientas bioinformáticas PLINK y SKAT. En concreto, la variante común rs710446 y 5 grupos de variantes genéticas raras y de baja frecuencia localizadas en *KNG1* se asociaron de forma significativa con los niveles de FXI incluso después de correcciones por test múltiples y permutaciones. Por otra parte, se han identificado dos potenciales mutaciones patogénicas relacionadas tanto con niveles altos de FXI como bajos, las cuales han sido identificadas mediante un proceso de filtrado de datos y predicciones *in silico*. Así pues, el estudio exhaustivo de *KNG1* y *F11* mediante NGS es de gran utilidad para esclarecer las relaciones existentes entre los niveles de FXI y las variantes genéticas en estas regiones de distintas frecuencias alélicas. En especial, las variantes genéticas funcionales halladas podrían contribuir en la práctica clínica como marcadores del riesgo de padecer VTE.

Material suplementario

Tablas suplementarias de la publicación (ver anexo III).

Next Generation Sequencing to Dissect the Genetic Architecture of *KNG1* and *F11* Loci using Factor XI Levels as an Intermediate Phenotype of Thrombosis.

Laura Martin-Fernandez[1], Giovana Gavidia-Bovadilla[1,2], Irene Corrales[3,4], Helena Brunel[1], Lorena Ramírez[3,4], Sonia López[1], Juan Carlos Souto[6], Francisco Vidal[3,4,5] and José Manuel Soria[1]*.

[1] Unit of Genomics of Complex Diseases, Biomedical Research Institute Sant Pau (IIB-Sant Pau), Barcelona, Spain.

[2] Department of ESAII, Center for Biomedical Engineering Research (CREB), Universitat Politècnica de Catalunya, Barcelona, Spain.

[3] Congenital Coagulopathies, Blood and Tissue Bank, Barcelona, Spain.

[4] Molecular Diagnosis and Therapy, Vall d'Hebron Research Institute, Universitat Autònoma de Barcelona (VHIR-UAB), Barcelona, Spain.

[5] CIBER de Enfermedades Cardiovasculares, Spain.

[6] Unit of Hemostasis and Thrombosis, Department of Hematology, IIB-Sant Pau, Hospital de la Santa Creu i Sant Pau, Barcelona, Spain.

***Corresponding author:**

Email: JSoria@santpau.cat (JMS)

Abstract

Venous thromboembolism is a complex disease with a high heritability. There are significant associations among Factor XI (FXI) levels and SNPs in the *KNG1* and *F11* loci. Our aim was to identify the genetic variation of *KNG1* and *F11* that might account for the variability of FXI levels. The *KNG1* and *F11* loci were sequenced completely in 110 unrelated individuals from the GAIT-2 (Genetic Analysis of Idiopathic Thrombophilia 2) Project using Next Generation Sequencing on an Illumina MiSeq. The GAIT-2 Project is a study of 935 individuals in 35 extended Spanish families selected through a proband with idiopathic thrombophilia. Among the 110 individuals, a subset of 40 individuals was chosen as a discovery sample for identifying variants. A total of 762 genetic variants were detected. Several significant associations were established among common variants and low-frequency variants sets in *KNG1* and *F11* with FXI levels using the PLINK and SKAT packages. Among these associations, those of rs710446 and five low-frequency variant sets in *KNG1* with FXI level variation were significant after multiple testing correction and permutation. Also, two putative pathogenic mutations related to high and low FXI levels were identified by data filtering and *in silico* predictions. This study of *KNG1* and *F11* loci should help to understand the connection between genotypic variation and variation in FXI levels. The functional genetic variants should be useful as markers of thromboembolic risk.

Introduction

Venous thromboembolism (VTE) includes pulmonary embolism and deep vein thrombosis. It results from changes in blood composition, blood flow and/or changes in the vessel wall. VTE is a common disease with an annual incidence of approximately 1 in 1,000 individuals in developed countries. It involves genetic and environmental risk factors and their interactions [1]. Previous studies [2–4] have estimated a heritability of approximately 60% for the risk of VTE. Well-established genetic risk factors for VTE are unfavourable genotypes at the *SERPINC1, PROC, PROS1, F2, FGG, F5* and *ABO* loci [5]. The known genetic variants involved in the risk of VTE explain only a small proportion of the genetic variance (heritability). Factor V Leiden (rs6025) and *F2* G20210A (rs1799963) mutations are the most commonly used markers in

clinical practice. The variance in the risk of VTE explained by these genetic variations is approximately 7% [6]. Recently, some genetic scores have been designed to predict the occurrence of VTE [6,7]. Soria *et al*. (2014) explained 15% of the variance in the risk of VTE and included 12 variants located in *F2*, *F5*, *ABO*, *F12*, *F13*, *SERPINA10* and *SERPINC1*. Some of these variants were rare variants.

Currently, both low-frequency and rare variants are probable sources of the unexplained heritability in complex traits [8]. Targeted sequencing in individuals selected from the tails of the normal distribution of quantitative phenotypes can be used to identify variants over the full minor allele frequency (MAF) spectrum. These individuals are more likely to carry alleles that cause loss or gain of function [8,9]. Of note, intermediate phenotypes (those correlated with disease) are statistically more powerful and closer to the action of genes than the liability of diseases [10]. Haemostasis parameters have been used to determine the underlying genetic component of the risk of VTE [11,12]. Specifically, plasma coagulation Factor XI (FXI) levels were considered as an intermediate quantitative phenotype in the Spanish families included in the Genetic Analysis of Idiopathic Thrombophilia 1 (GAIT-1) Project to identify new genetic risk factors that contribute to thrombotic disease [2,13]. Interestingly, FXI is a zymogen of a serine protease that participates in the intrinsic coagulation pathway. This glycoprotein has been described [14] as a potential target for a new strategy for VTE treatment. Plasma FXI levels have a heritability of about 45% and a significant genetic correlation with the risk of VTE [2,13]. The genome-wide association studies (GWAS) of plasma FXI levels performed in the GAIT-1 Project showed significant associations with variants in the *KNG1* and *F11* loci. Those results were replicated in a population-based Swedish cohort [15]. The *KNG1* encodes for a high molecular weight kininogen (HMWK). This gene is located on Chromosome 3q27.3 and is 25,581 bp in length. *F11* is the structural gene of coagulation FXI protein and is located on Chromosome 4q35.2 with a length of 23,718 bp (UCSC, Feb. 2009 GRCh37/hg19 release (http://genome.ucsc.edu/) [16]). Also, there is an association of *KNG1* with activated partial thromboplastin time (aPTT) [17,18] and the risk of VTE [19]. The *F11* locus has been associated with aPTT [18] and FXI levels [20,21]. In addition, there is an association of *F11* with the risk of VTE [20,21] that is explained at least in part by an association with FXI levels after pairwise linkage disequilibrium (LD).

We designed a targeted gene Next Generation Sequencing (NGS) strategy to dissect the genetic variability of *KNG1* and *F11* loci. We studied 110 genetically unrelated individuals from the GAIT-2 Project and we identify the genetic variants that might contribute to the variation of plasma FXI levels.

Material and Methods

Study Subjects

The GAIT-2 Project included 935 individuals in 35 extended Spanish families. These families were selected through a proband with idiopathic thrombophilia. A detailed description of the recruitment and the criteria used for inclusion have been given previously [4,22]. Our study was performed according to the Declaration of Helsinki and adult subjects gave written informed consent for themselves and for their minor children. All procedures were evaluated and approved by the Institutional Review Board of the Hospital de la Santa Creu i Sant Pau, Barcelona, Spain.

A total of 110 individuals genetically unrelated from the GAIT-2 Project were chosen to be sequenced using NGS. A subset sample of 40 unrelated individuals was selected as a discovery sample. Among these 40 individuals, there were 20 with low plasma FXI levels (36-80%) and 20 with high plasma FXI levels (158-250%) compared to the normal range in the clinical diagnosis (55-185%). There were 17 males and 23 females. This discovery sample provided more than 98% probability of detecting variants with a MAF greater than or equal to 0.05 [9]. Thus, variants identified by the NGS in the discovery sample were denoted as “variants of interest” for subsequent association analyses and for putative pathogenic mutation screening. Importantly, the genotypes of the “variants of interest” of individuals from the whole sample of 110 individuals were used for association analyses to increase the statistical power to identify statistically significantly genetic variants.

Blood Collection and Phenotype Determinations

Blood was collected by venipuncture following a 12-hour fast. Samples were collected in an anticoagulant consisting of 1/10 volume containing 0.129 mol/L sodium citrate. None of the

participants was using oral anticoagulants or heparins at the time of blood collection. Platelet-poor plasma was obtained by centrifugation at 2,000 *g* for 20 minutes at room temperature (22±2ºC) and used for the phenotype determinations that required fresh samples. The remaining plasma was stored at −80°C. FXI activity levels were assayed with deficient plasma (Diagnostica Stago, Asnières, France). DNA was extracted from whole blood using a standard salting out procedure [23].

PCR Primer Design

One short PCR and 9 Long Range (LR) PCRs were carried out to amplify 60,052 bp of genomic DNA encompassing the *KNG1* (GRCh37/hg19 chr3:186,435,098-186,460,678) and *F11* (GRCh37/hg19 chr4:187,187,118-187,210,835) loci. All the PCR amplicons were overlapped within each gene region including all exons, introns, 5'-UTR, 3'UTR and approximately 1,500 bp of the promoter region. Primer sequences, primer positions and PCR amplicon sizes in *KNG1* and *F11* are shown in S1 Table.

The PCR amplicons were tested for target specificity by Sanger sequencing of both strands as described below.

PCR Amplification and Normalization

Short PCR was performed with the FastStart High Fidelity PCR System, dNTPack (Roche Diagnostics, Mannheim, Germany). LR-PCR amplifications were performed using the SequalPrep Long PCR Kit with dNTPs (Invitrogen, Thermo Fisher Scientific Inc., MA, USA). PCR master mix conditions for Short and LR-PCR amplifications are described in S2 Table and thermocycling conditions used to obtain the optimum short and LR-PCR amplifications are shown in S3 and S4 Table.

Short and LR-PCR amplicons of the 110 individuals were separated on 0.7% agarose gel electrophoresis and visualized by SYBR safe (Invitrogen, Thermo Fisher Scientific Inc.) staining. Also, all of the PCR amplicons were quantified using the fluorometer Qubit (Invitrogen, Thermo Fisher Scientific Inc.). A normalized pool of the 10 PCR amplicons was obtained for each individual by combining equimolar amounts. The fluorometer Qubit was used to adjust the 110 PCR pools at 0.2 ng/µl for library preparation.

Library Preparation, Sequencing and Data Analysis

The sequencing libraries were prepared from PCR pools using the Nextera XT DNA Sample Preparation kit (Illumina, San Diego, CA, USA) with double indexing, according to the manufacturer's protocol. Paired-end sequencing was used to improve the mapping quality [24]. We obtained 110 paired-end libraries that were pooled and simultaneously run on an Illumina Miseq sequencing system (Illumina) by the Miseq sequencing reagent kit v2 of 300 cycles (2x150 bp paired end) (Illumina).

Indexed sequences were de-multiplexed and analyzed individually. Paired sequence files in fastq format were analysed with CLC Genomic Workbench version 6.5 software (CLC Bio - Qiagen, Aarhus, Denmark). The raw data were trimmed with length (minimum 25 bp; maximum 500 bp), ambiguous nucleotide (maximum 2) and quality score (0.05) filters. This software aligns the trimmed reads against a human genome reference (hg19). The read mapping was performed with specific parameter setting (mismatch count, 2; indel count, 3; length fraction, 0.7; similarity fraction, 0.9). Parameters were used for quality-based variant detection (minimum coverage, 30x; minimum variant frequency, 25%). Results in variant call format (VCF) file were used as input for the Illumina VariantStudio Data Analysis version 2.1 Software (Illumina) to annotate the genetic variants.

The amino acid numbering and the nomenclature used to describe the sequence variations followed the international recommendations of the Human Genome Variation Society (HGVS; http://www.HGVS.org).

Association Analysis for Common and Low-Frequency Variants

We treated our data (FXI levels) as normally distributed over the 110 samples. The normality of the distribution was tested using the Shapiro-Wilk normality test (W = 0.97623; p-value = 0.052) First, MAFs in the 110 individuals were calculated using the PLINK package version 1.07 [25]. For this association analysis, the common variants were defined as genetic variants with MAF ≥10%. Thus, low-frequency variants were defined as genetic variants with MAF <10% [26]. Then, LD based variant pruning was applied for each variant group (common and low-frequency) using the PLINK package to find informative variants. For this, we used the variance inflation factor (VIF) criterion to check for multi-collinearity of variants and recursively remove

variants within a sliding window. The VIF was calculated as $1/(1-R^2)$ where R^2 is the multiple correlation coefficient for a variant regressed on all other variants simultaneously at each step. This method considers the correlation between variants and between linear combinations of them. We used a variant windows size of 30, a number of variants to shift the windows at each step equal to 3 and the VIF threshold equal to 2 (i.e. implies R^2 of 0.5). This allowed variants greater than this VIF to be removed. Association with plasma FXI levels was performed by using two different approaches: (a) a single linear association for common variants using PLINK package, and, (b) a collapsing method based on sliding window for low-frequency variants using the SNP-set (Sequence) Kernel Association Test (SKAT version 1.0.9) available in R [27]. A single variant test analysis is the standard approach to testing for association between genetic variants. However, this approach is less powerful for low-frequency variants. Thus, we analyze common and low-frequency variants separately in order to guarantee the robustness of the results. In the collapsing method, the sets of variants were obtained by shifting one low-frequency variant to the right within a sliding 2-kb window. SKAT aggregates individual score test statistics of each variant and computes variant-set level p-values, while adjusting for covariates. This allowed different variants to have different directions and magnitudes, including no effects [28]. Both methods were adjusted by age and gender. The association of common variants with plasma FXI levels was adjusted for multiple testing by controlling for family-wise error rate (FWER). Multiple testing included Bonferroni single-step correction, Holm step-down correction, Sidak single-step correction, Sidak step-down correction, Benjamini& Hochberg false discovery rate (FDR) control and Benjamini&Yekutieli FDR control. The FWER in low-frequency variants was controlled using a resampling method. We applied a permutation strategy using 1,000 permutations to determine the empirical p-value for denoting a particular variant-plasma FXI level association as statistically significant. Permutation shuffled the FXI values by maintaining the correlation structure among variants. Finally, we used the clump function in PLINK to identify independent signals among common variants.

Putative Pathogenic Mutations Screening

The following criteria were used to identify putative pathogenic mutations in the data of genetic variation in the discovery sample: a) whether the variant was rare (allele frequency <1% in 1000

Genomes April 2012 version 3 [29] and NHLBI Exome Variant Server June 2013 ESP6500SI-V2), b) location referring to Variant Effect Predictor version 2.8 data base (intronic mutations located >30 bp into flanking intronic regions of each exon were rejected), and c) MAF in our sample of 110 unrelated individuals <5%, for variants not annotated in 1000 Genomes April 2012 version 3 [29].

Sanger Sequencing Validations

Putative pathogenic mutations were validated by conventional Sanger sequencing. Briefly, enzymatic purification was performed using ExoSAP-IT treatment (USB Corporation, Cleveland, OH, USA) and Sanger sequencing of both strands was carried out using the BigDye Terminator version 3.1 Cycle Sequencing Kit (Applied Biosystems, Thermo Fisher Scientific Inc., MA, USA). The ABI Prism 3130 Genetic Analyzer (Applied Biosystems, Thermo Fisher Scientific Inc.) was used for capillary electrophoresis. The sequences were mapped against *KNG1* (GRCh37/hg19 chr3:186,435,098-186,460,678; NM_001102416.2) and *F11* (GRCh37/hg19 chr4:187,187,118-187,210,835; NM_000128.3) loci using CLC Genomic Workbench version 6.5 software. We performed also a co-segregation analysis between validated putative pathogenic mutations and plasma FXI levels in available family members by Sanger sequencing.

In Silico Prediction Analysis

Functional effects of putative pathogenic mutations that co-segregated with plasma FXI levels were evaluated using the *in silico* prediction software Alamut Visual version 2.6.1 (Interactive Biosoftware, Rouen, France). Potential splicing alterations in intron variants included SpliceSiteFinder, MaxEntScan, NNSplice, GeneSplicer and Human Splicing Finder algorithms. The threshold employed was a variation between the native and the mutation score of more than 10% in at least two different algorithms [30]. Pathogenic impact of the missense variants was analysed using the programs SIFT, PolyPhen-2, Align GVGD and Mutation Taster. Conservation phyloP scores were obtained also. Interpretation of predictive structural effects of the missense mutations was evaluated by using the Project HOPE software [31].

Replication Subjects and Genotyping

A total of 250 unrelated patients who had suffered thrombosis from a case-control study of Spanish population samples [32] were designated as independent replication subjects. They

were used to genotype putative pathogenic mutations from the discovery sample that might be involved in the risk of VTE. Genotyping was performed using TaqMan technology (Applied Biosystems, Thermo Fisher Scientific Inc.) and run in an ABI 7500 instrument (Applied Biosystems, Thermo Fisher Scientific Inc.).

Results

NGS Statistics

We amplified 60,052 bp of genomic DNA from 110 individuals of the GAIT-2 Project to study the candidate genes *KNG1* and *F11*. Libraries were prepared and pooled for NGS.

The length of the *KNG1* target region was 29,535 nucleotides. The number of reads that cover this region was 7,746,180 and the percentage of the mapped positions with a depth of coverage above 30x was 98%. Further, the median coverage of the gene per individual was 242x. The length of *F11* region was 25,358 nucleotides, the number of reads covering this region was 3,504,328 and the 99% of the positions had coverage above 30x. The median coverage per individual in the *F11* target region was 131x.

A subset of 40 individuals was selected as a discovery sample. We identified a total of 762 unique biallelic variants from this discovery sample (S5 Table). We identified 504 genetic variants in the *KNG1* locus and 258 genetic variants in the *F11* locus. The percentage of indels in the *KNG1* locus was 7.9% (n=40) and 8.1% (n=21) in the *F11* locus. The percentage of exonic variants in the *KNG1* locus was 2.2% (n=11) and 6.2% (n=16) in the *F11* locus. We found 30.2% (n=152) of the variants in the *KNG1* locus that had an allele frequency >1% from all populations of 1000 Genomes data (April 2012 version 3) and from four populations of 1000 Genomes (American, East Asian, African and European). The 20.5% (n=53) of the variants in the *F11* locus had an allele frequency >1% from all populations of 1000 Genomes data (April 2012 version 3) and from four populations of 1000 Genomes (American, East Asian, African and European). Moreover, the 57.1% (n=288) of the variants in the *KNG1* locus were not reported in the dbSNP version 137 and the 61.6% (n=159) of the variants in the *F11* locus were not reported in the dbSNP version 137. The majority of these variants that were not reported in

dbSNP version 137 were found within the introns (only 2 out of the 288 variants in the *KNG1* locus were located in the promoter region).

Association Analyses

The genotypes of the 110 unrelated individuals for the 762 genetic variants in the discovery sample were used for association analyses to increase the power to detect significant variant-plasma FXI level associations. Of these 762 variants, 41.47% were common (MAF in our population of 110 individuals ≥10%) and the 58.53% were of low frequency (MAF in our population of 110 individuals <10%). After LD based pruning, 125 common variants and 275 low-frequency variants remained.

Using single linear association, the common genetic variant rs710446 was significantly associated with plasma FXI levels after correction for multiple testing. This variant was located in the *KNG1* region. Also, we identified 12 common variants, in addition to rs710446, that showed nominally statistically significant association with plasma FXI levels before correction for multiple testing (p-value <0.05) (Table 1). Of all nominally significantly associated 13 common variants, 10 common variants were located at the *KNG1* locus and 3 common variants were located at the *F11* locus. In detail, 9 out of the 10 common variants in the *KNG1* locus were located in intronic regions and 1 common variant (top SNP in *KNG1* region rs710446) was located within exon 10. In the *F11* region, 2 common variants were located in intronic regions and one variant was located downstream of the locus. The intronic variant rs56810541 was the top SNP in this region. Fig 1 and 2 show plots of the association of common variants with plasma FXI levels. We found 11 independent signals after clumping (rs5030062 and rs3856930 are grouped with the top SNP).

Table 1. Significant associations of common genetic variants in *KNG1* (NM_001102416.2) and *F11* (NM_000128.3) with plasma FXI levels.

Gene	Genome Location	Nucleotide Change	dbSNP v137	Tested Allele[1]	MAF (%)[2]	β	P-value	Bonferroni -adjusted p-value
KNG1	chr3:186,441,823	c.392-1054A>G	rs1656915	A	26.64	-10.9	0.034	1.000
KNG1	chr3:186,442,707	c.392-170T>C	rs1648711	T	50.00	-12.1	0.016	1.000
KNG1	chr3:186,443,756	c.564+707C>G	rs266724	C	50.00	-12.1	0.016	1.000
KNG1	chr3:186,444,831	c.565-195T>G	rs62294376	G	38.32	16.5	0.003	0.355
KNG1	chr3:186,450,895	c.930+432T>C	rs1656926	T	29.91	-16.8	0.007	0.887
KNG1	chr3:186,453,418	c.930+2955A>T	rs4686800	A	33.18	-8.4	0.045	1.000
KNG1	chr3:186,454,180	c.931-2708A>C	rs5030062	C	38.79	15.7	0.001	0.148
KNG1	chr3:186,458,322	c.1126-989C>T	rs3856930	T	35.05	12.2	0.015	1.000
KNG1	chr3:186,458,910	c.1126-401G>A	rs5030081	A	14.95	23.2	0.002	0.227
KNG1	chr3:186,459,927	c.1742T>C	rs710446	C	47.20	19.1	0.00008	0.010
F11	chr4:187,197,994	c.755+450C>T	rs4253416	T	46.73	11.5	0.043	1.000
F11	chr4:187,200,550	c.756-616A>T	rs56810541	T	36.92	11.6	0.029	1.000
F11	chr4:187,210,837	c.*1069delT	rs67843441	A	22.47	10.7	0.049	1.000

[1] The tested allele is the minor allele by default in our population of 110 individuals.
[2] The minor allele frequency from our population of 110 individuals, using PLINK package (version 1.07) [25].

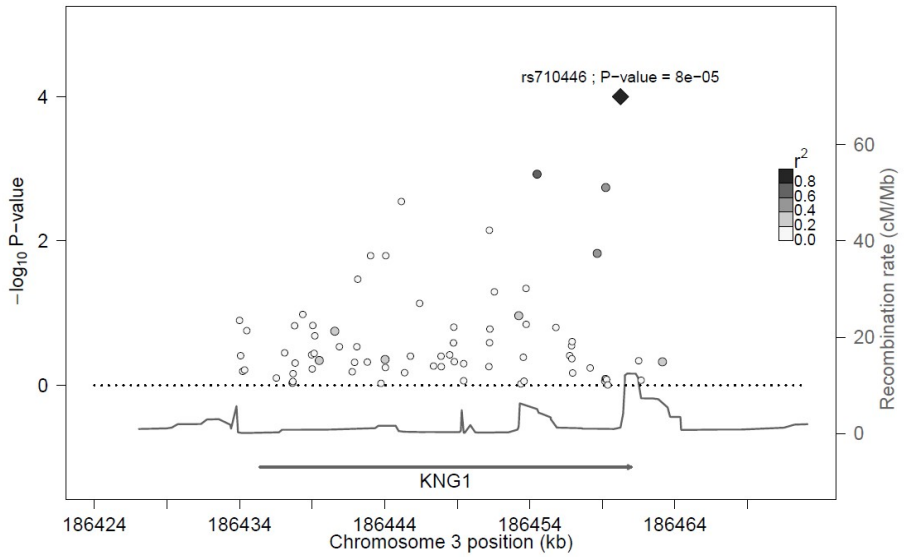

Fig 1. Plot of the association in the *KNG1* locus with plasma FXI levels. Markers represented common variants organized by genomic position. The diamond-shaped marker represented the top SNP (rs710446) which was still statistically significantly associated after adjustment for multiple testing. The left axis shows the statistical significance of the variant-plasma FXI level variation association expressed as -$\log_{10}$ of the p-values and colour intensities show the level of linkage disequilibrium between all variants and the top SNP. The recombination rate in the HapMap II sample [33] for this region is measured on the right axis.

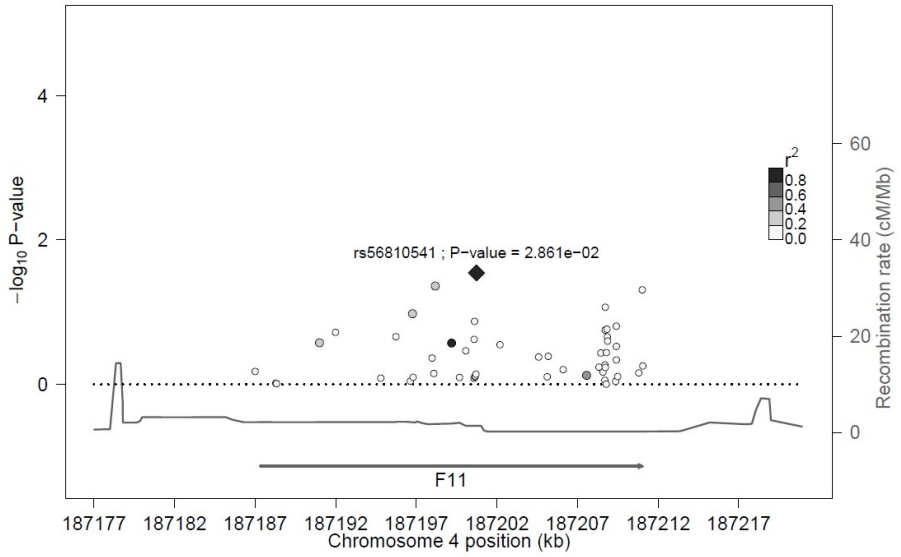

Fig 2. Plot of the association between variants within the *F11* locus with plasma FXI levels. Markers represented common variants organized by genomic position. Tthe diamond-shaped marker represented the top SNP (rs56810541) with the lowest variant-plasma FXI level association p-value. The left axis shows the statistical significance expressed as -$\log_{10}$ of the p-values and colour intensities show the level of linkage disequilibrium between all variants and the top SNP. The recombination rate in the HapMap II sample [33] for this region is measured on the right axis.

Using the collapsing method, 5 low-frequency variant sets were significantly associated with plasma FXI levels after 1,000 permutation analyses, which was used to control the family-wise error rate (FWER=0.05). Specifically, the 5 low-frequency variant sets were chr3:186,448,468-186,450,468 (p-value=0.0002), chr3:186,448,470-186,450,470 (p-value=0.0003), chr3:186,448,478-186,450,478 (p-value=0.0005), chr3:186,448,482-186,450,482 (p-value=0.0005) and chr3:186,448,484-186,450,484 (p-value=0.0004) and included the following genetic variants (NM_001102416.2): c.673-866A>T, c.673-864A>T, c.673-862A>T, c.673-860delT, c.673-856T>C, c.673-852T>C, c.673-850T>C, c.673-842T>C, c.673-838T>C, c.673-806T>C, c.673-406T>C, c.673-136T>C, c.673-67A>G, c.758-91A>C and c.758-12T>C. Also, we identified an additional 89 nominally significantly associated low-frequency variant sets (p-value <0.05) before 1,000 permutation analyses. By merging overlapped significant variant sets before and after 1,000 permutation analyses, different regions in *KNG1* and *F11* were related to plasma FXI variability. Briefly, in the *KNG1* locus, a total of 4 regions were targeted. The first region (chr3:186,441,571-186,444,798) encompassed the exon 4 and the surrounding intronic

areas. The second region (chr3:186,446,561-186,450,528) included exon 6, exon 7 and the introns around them. Specifically, the 5 low-frequency variant sets significantly associated with plasma FXI levels after 1,000 permutation analyses were located within these second region (chr3:186,446,561-186,450,528) in *KNG1*. The next area (chr3:186,454,294-186,458,652) encompassed the exons 8 and 9 and the surrounding intronic region. Finally, the downstream region of *KNG1* was the last targeted region (186,462,745-186,465,081). In the *F11* locus 3 regions were highlighted. The first one (chr4:187,189,805-187,191,805) emerged 2,000 bp of the intron 2. Another targeted region encompassed the exons 8, 9 and 10 and the surrounding intronic areas (chr4:187,200,524-187,204,276) and the last targeted region included the exons 13, 14 and 15, the 3'UTR, the intronic areas around them and the downstream region (chr4:187,207,354-187,212,318). Plot for collapsing method in *KNG1* locus is represented in Fig 3 and plot for collapsing method in *F11* locus is represented in Fig 4.

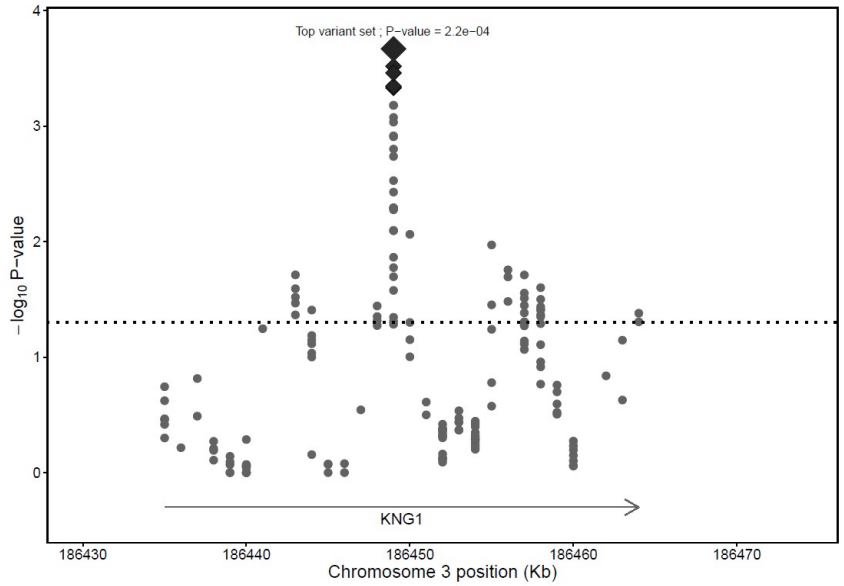

Fig 3. Plot of the collapsing method association in the *KNG1* locus with plasma FXI levels. Markers represented the mean position of low-frequency variant sets. All of the markers located above the dotted horizontal line obtained a p-value <0.05 after the collapsing method association. Diamond-shaped markers represented the five significant low-frequency variant sets with controlling FWER=0.05. The biggest diamond-shaped marker is the top low-frequency variant set.

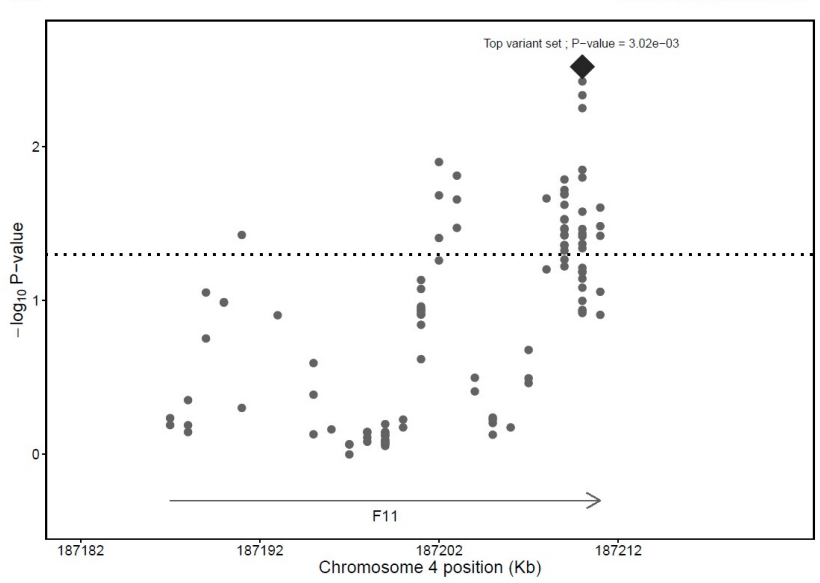

Fig 4. Plot of the collapsing method association in the *F11* locus. Markers represented the mean position of low-frequency variant sets. The diamond-shaped marker represented the top low-frequency variant set and markers located above the dotted horizontal line are the most significant low-frequency variant sets (p-value <0.05). None of the low-frequency variant sets were significantly associated after controlling FWER=0.05.

Putative Pathogenic Mutations

Among the 762 variants, 6 were selected as potentially functional mutations without association analyses (Table 2). By Sanger sequencing, all candidate mutations were validated to verify the segregation with the FXI levels. Thereby, 2 mutations remained as candidate pathogenic mutations: the intronic variant NM_001102416.2: c.758-12T>C and the missense variant NM_000128.3: c.943G>A (p.Glu315Lys).

Table 2. Validated putative pathogenic mutations in *KNG1* (NM_001102416.2) and *F11* (NM_000128.3).

Gene	Nucleotide Change	Location	dbSNP v137	MAF[1] (%)	Co-segregation Validation
KNG1	c.-1438_-1433del	Upstream	-	-	No
KNG1	c.-1294G>T	Upstream	rs2651642	NA	No
KNG1	c.758-12T>C	Intron 6	-	-	Yes
F11	c.1304+12G>A	Intron 11	rs116667976	0.18	No
F11	c.943G>A	Exon 9	rs281875257	NA	Yes
F11	c.*566G>C	3'UTR	-	-	No

NA: Not Annotated. [1] The variant allele frequency from all populations of 1000 genomes data, April 2012 version 3 (www.1000genomes.org) [29].

The putative pathogenic mutation NM_001102416.2: c.758-12T>C (in intron 6) at the *KNG1* locus was in heterozygous state within 1 subject with high plasma FXI levels (188%). This variant has been listed previously as variant 3:186450279 T / C in the Exome Aggregation Consortium (ExAC) Browser (Cambridge, MA, http://exac.broadinstitute.org) with a MAF of $4.12x10^{-5}$. The predicted variations at the natural acceptor splice site of exon 7 between the native and the mutation score were -3.7% (SpliceSiteFinder score native-mutated: 88.56-85.25 [0-100], threshold: ≥70), -6.6% (MaxEntScan score native-mutated: 6.28-5.86 [0-16], threshold: ≥0), -13.7% (NNSplice score native-mutated: 0.66-0.57 [0-1], threshold: ≥0.4), -12.2% (GeneSplicer score native-mutated: 7.43-6.52 [0-15], threshold: ≥0) and -1.4% (Human Splicing Finder score native-mutated: 91.23-89.98 [0-100], threshold: ≥65). These results suggested that the intronic variant NM_001102416.2: c.758-12T>C might interfere with the recognition of the natural acceptor splice site. Notably, this variant was identified in an individual with high FXI levels from the GAIT-2 Project, so the new putative pathogenic mutation NM_001102416.2: c.758-12T>C in the *KNG1* was genotyped in the case-control study of thrombosis. However, it was not identified in any of the 250 patients.

The variation NM_000128.3: c.943G>A (p.Glu315Lys) in the exon 9 of *F11* was in heterozygous state in one individual with low plasma FXI levels (36%). The predicted effects were tolerated (SIFT score: 1 [1-0], median: 3.45), possibly damaging (PolyPhen-2 score: 0.830 [0-1], sensitivity: 0.84, specificity: 0.93), likely pathogenic (Align GVGD score: C55 [C0-C65], GV: 0.00, GD: 56.87) and polymorphism (Mutation Taster p-value: 1 [0-1]). It was predicted that this putative pathogenic mutation would have a weakly conserved nucleotide (phyloP: 1.09 [-14.1;6.4]). According to Project HOPE, the native glutamic acid and the lysine residue resulted from the mutation differ in size and charge. The lysine residue has the opposite charge (positive) at this position and thus contacts with other molecules are likely disturbed. In addition, the lysine residue is bigger than the glutamic acid residue and, as it is located on the surface of the protein, the mutation affects interactions with other parts of the protein and with other molecules. The p.Glu315Lys variation was located within a domain named “Apple 4”, which is important for binding other molecules. Thus, the mutation could disturb the interaction between domains and could ultimately affect the normal protein function.

Discussion

We report that the genetic variability of *KNG1* and *F11* loci influences the variation in plasma FXI levels.

Our approach was based on target sequencing from a set of high-fidelity polymerase amplifications using NGS. We performed whole gene sequencing to examine not only exonic regions but also non-coding regions that might have regulatory functions [9,34,35].

The *KNG1* and *F11* loci were covered by a large number of high quality reads (median coverage per individual of 242x for *KNG1* and 131x for *F11*), as it has been suggested that a sequencing depth between 30-35x is already optimal to determine variants [36]. We detected a total of 762 genetic variants at the *KNG1* and *F11* loci in the 40 discovery sample individuals. Most of these variants (58.7%) had not been annotated previously in the dbSNP version 137, similar to what has been reported for other loci [26,37], In addition, only the 26.9% of the variants had an allele frequency from 1000 Genomes (April 2012 version 3) >1%. Thus, most of the 762 variants would not have been included in a GWAS panel.

We analyzed common variants for association with plasma FXI levels in the 110 unrelated individuals. Most importantly, the missense variant rs710446 (p.Ile581Thr) in *KNG1* was associated significantly with high FXI levels after correction for multiple testing. Of note, this genetic variant has been already described [15] in association with plasma FXI levels as a top SNP in the *KNG1* region. Also, the rs710446 variant has been associated previously [17–19] with aPTT and the risk of VTE. In addition, we found 12 additional variants in *KNG1* and *F11* loci that showed nominal significant associations with FXI levels. Of them, the rs5030062 variant in the *KNG1* locus has been described previously [15] in association with plasma FXI levels. Our results are consistent with these results and extents them.

In our study, the low-frequency variants were grouped according to their position across regions of 2 kb and we could identify different sets of low-frequency variants in association with FXI levels. They could be exposing independent regions that contribute to the phenotypic variability.

Two studies [38,39] have explored the risk of VTE using NGS targeted gene strategy. They support the hypothesis that low-frequency and rare variants may contribute to the risk of VTE. We evaluated a combination of common and low-frequency variants at two different loci that

affect plasma FXI levels. Our results demonstrated the complexity of phenotypic variation in a single trait.

The 2 putative mutations that were associated with low and high FXI levels indicated that these mutations could be risk factors of FXI deficiency or of thrombosis. We evaluated whether the intronic variation NM_001102416.2: c.758-12T>C in *KNG1* might increase the risk of disease using 5 *in silico* bioinformatics programs. All 5 of them showed that the mutation might reduce splicing and 2 of them predicted an alteration of more than 10% at the acceptor site. To our knowledge, this is the first time that this putative pathogenic mutation has been described in association with high FXI levels. Our genotype assay in patients who had suffered thrombosis from a case-control study did not identify this mutation in this study of Spanish case-control population as it might be restricted to this family. The relatively small number of patients included in our replication study limits the confidence of our conclusions and larger studies need to be performed. Interestingly, the *in silico* prediction of the missense variant NM_000128.3: c.943G>A in the *F11* locus was not conclusive. According to previously published association data [40], this mutation in heterozygous state is associated with low plasma FXI levels. Thus, it is possible that the *in silico* predictions could be underestimating its pathogenicity. We think it is important to emphasize that the 2 putative pathogenic mutations that we detected were located within 2 low-frequency variant sets (chr3:186,446,561-186,450,528 in *KNG1* and chr4:187,200,524-187,204,276 in *F11*) that were nominally significantly associated with plasma FXI levels. Interestingly, the intronic variation NM_001102416.2: c.758-12T>C in *KNG1* was one of the variants included in the low-frequency variant sets that were significantly associated with plasma FXI levels after permutation testing.

Most of the genetic variants associated with plasma FXI levels that we have described would not be detected by examining only the coding regions of the genes. Therefore, it is clear that the non-coding regions can contribute to the regulation of complex phenotypes. We applied *in silico* tools to clarify this issue and provide evidence that support the pathogenic role. However, because prediction programs have serious limitations, further analyses are needed to determine the functional characteristics of these putative mutations.

In conclusion, the continuous DNA sequence data reported in our study using the targeted gene NGS strategy represent one of the largest bodies of sequence data on individuals for the *KNG1* and *F11* loci. Based on the large variation that we detected in the *F11* and *KNG1* loci among 110 individuals from the GAIT-2 Project, we suggest strongly that an overall effect of several of these genetic variations modulates plasma FXI levels. Clearly, further studies are warranted. The resulting functional genetic variants should be useful to predict the risk of thromboembolic disease.

Acknowledgments

We are deeply grateful to the families who participated in this study. Also, we would like to thank Professor Bill Stone for reviewing the manuscript. CIBERCV is an initiative of ISCIII co-financed by Fondo Europeo de Desarrollo Regional (FEDER) a way to build Europe.

Financial Disclosure

This study was supported by funds from the Instituto de Salud Carlos III Fondo de Investigación Sanitaria PI 11/0184, PI 12/01494 and PI 15/00269 and Red Investigación Cardiovascular RD12/0042/0032 and RD12/0042/0053. Laura Martin-Fernandez was supported by Ayudas Predoctorales de Formación en Investigación en Salud (PFIS) FI12/00322.

References

1. Rosendaal FR. Venous thrombosis: the role of genes, environment, and behavior. Hematology Am Soc Hematol Educ Program. 2005;2005: 1–12.
2. Souto JC, Almasy L, Borrell M, Blanco-Vaca F, Mateo J, Soria JM, et al. Genetic susceptibility to thrombosis and its relationship to physiological risk factors: the GAIT study. Am J Hum Genet. 2000;67: 1452–1459.
3. Heit JA, Phelps MA, Ward SA, Slusser JP, Petterson TM, De Andrade M. Familial segregation of venous thromboembolism. J Thromb Haemost. 2004;2: 731–736.
4. Martin-Fernandez L, Ziyatdinov A, Carrasco M, Millon JA, Martinez-Perez A, Vilalta N, et al. Genetic Determinants of Thrombin Generation and Their Relation to Venous Thrombosis: Results from the GAIT-2 Project. PLoS One. 2016;11: e0146922.
5. Rosendaal FR, Reitsma PH. Genetics of venous thrombosis. J Thromb Haemost. 2009;7 Suppl 1: 301–304.
6. Soria JM, Morange P-E, Vila J, Souto JC, Moyano M, Tregouet D-A, et al. Multilocus Genetic Risk Scores for Venous Thromboembolism Risk Assessment. J Am Heart Assoc. 2014;3: e001060.
7. De Haan HG, Bezemer ID, Doggen CJM, Le Cessie S, Reitsma PH, Arellano AR, et al. Multiple SNP testing improves risk prediction of first venous thrombosis. Blood. 2012;120: 656–663.
8. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. Macmillan Publishers Limited. All rights reserved; 2009;461: 747–753.
9. Soria JM, Almasy L, Souto JC, Sabater-Lleal M, Fontcuberta J, Blangero J. The F7 Gene and Clotting Factor VII Levels: Dissection of a Human Quantitative Trait Locus. Hum Biol. 2009;81: 853–867.
10. Souto JC. Search for new thrombosis-related genes through intermediate phenotypes. Genetic and household effects. Pathophysiol Haemost Thromb. 2002;32: 338–340.
11. Soria JM, Almasy L, Souto JC, Bacq D, Buil A, Faure A, et al. A quantitative-trait locus in the human factor XII gene influences both plasma factor XII levels and susceptibility to thrombotic disease. Am J Hum Genet. 2002;70: 567–574.
12. Buil A, Trégouët D-A, Souto JC, Saut N, Germain M, Rotival M, et al. C4BPB/C4BPA is a new susceptibility locus for venous thrombosis with unknown protein S-independent mechanism: results from genome-wide association and gene expression analyses followed by case-control studies. Blood. 2010;115: 4644–4650.
13. Souto JC, Almasy L, Borrell M, Garí M, Martínez E, Mateo J, et al. Genetic determinants of hemostasis phenotypes in Spanish families. Circulation. 2000;101: 1546–1551.
14. Zhang H, Löwenberg EC, Crosby JR, MacLeod AR, Zhao C, Gao D, et al. Inhibition of the intrinsic coagulation pathway factor XI by antisense oligonucleotides: a novel antithrombotic strategy with lowered bleeding risk. Blood. American Society of Hematology; 2010;116: 4684–4692.
15. Sabater-Lleal M, Martinez-Perez A, Buil A, Folkersen L, Souto JC, Bruzelius M, et al. A genome-wide association study identifies KNG1 as a genetic determinant of plasma factor XI Level and activated partial thromboplastin time. Arterioscler Thromb Vasc Biol. 2012;32: 2008–2016.
16. Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, et al. The UCSC Genome Browser database: 2014 update. Nucleic Acids Res. 2014;42: D764–D770.
17. Houlihan LM, Davies G, Tenesa A, Harris SE, Luciano M, Gow AJ, et al. Common variants of large effect in F12, KNG1, and HRG are associated with activated partial thromboplastin time. Am J Hum Genet. 2010;86: 626–631.
18. Tang W, Schwienbacher C, Lopez LM, Ben-Shlomo Y, Oudot-Mellakh T, Johnson AD, et al. Genetic associations for activated partial thromboplastin time and prothrombin time, their gene expression profiles, and risk of coronary artery disease. Am J Hum Genet. 2012;91: 152–162.
19. Morange P, Oudot-Mellakh T. KNG1 Ile581Thr and susceptibility to venous thrombosis. Blood. 2011;117: 3692–3694.

20. Li Y, Bezemer ID, Rowland CM, Tong CH, Arellano AR, Catanese JJ, et al. Genetic variants associated with deep vein thrombosis: the F11 locus. J Thromb Haemost. 2009;7: 1802–1808.
21. Bezemer ID, Bare LA, Doggen CJM, Arellano AR, Tong C, Rowland CM, et al. Gene variants associated with deep vein thrombosis. JAMA. 2008;299: 1306–1314.
22. Camacho M, Martinez-Perez A, Buil A, Siguero L, Alcolea S, López S, et al. Genetic determinants of 5-lipoxygenase pathway in a Spanish population and their relationship with cardiovascular risk. Atherosclerosis. 2012;224: 129–135.
23. Miller SA, Dykes DD, Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res. 1988;16: 1215.
24. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. Macmillan Publishers Limited. All rights reserved; 2008;456: 53–59.
25. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81: 559–575. Available: http://pngu.mgh.harvard.edu/purcell/plink/
26. Harismendy O, Bansal V, Bhatia G, Nakano M, Scott M, Wang X, et al. Population sequencing of two endocannabinoid metabolic genes identifies rare and common regulatory variants associated with extreme obesity and metabolite level. Genome Biol. BioMed Central; 2010;11: R118.
27. Lee S, Miropolsky L, Wu M. SKAT: SNP-Set (Sequence) Kernel Association Test. R package version 1.0.9. [Internet]. 2015. Available: https://cran.r-project.org/web/packages/SKAT/
28. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011;89: 82–93.
29. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491: 56–65.
30. Théry JC, Krieger S, Gaildrat P, Révillion F, Buisine M-P, Killian A, et al. Contribution of bioinformatics predictions and functional splicing assays to the interpretation of unclassified variants of the BRCA genes. Eur J Hum Genet. 2011;19: 1052–1058.
31. Venselaar H, te Beek TA, Kuipers RK, Hekkelman ML, Vriend G. Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. BMC Bioinformatics. 2010;11: 548.
32. Tirado I, Soria JM, Mateo J, Oliver A, Souto JC, Santamaria A, et al. Association after linkage analysis indicates that homozygosity for the 46C-->T polymorphism in the F12 gene is a genetic risk factor for venous thrombosis. Thromb Haemost. 2004;91: 899–904.
33. International HapMap Consortium. A haplotype map of the human genome. Nature. 2005;437: 1299–1320.
34. Corrales I, Ramírez L, Altisent C, Parra R, Vidal F. The study of the effect of splicing mutations in von Willebrand factor using RNA isolated from patients' platelets and leukocytes. J Thromb Haemost. 2011;9: 679–688.
35. Gehring NH, Frede U, Neu-Yilik G, Hundsdoerfer P, Vetter B, Hentze MW, et al. Increased efficiency of mRNA 3' end formation: a new genetic mechanism contributing to hereditary thrombophilia. Nat Genet. 2001;28: 389–392.
36. Pelak K, Shianna K V, Ge D, Maia JM, Zhu M, Smith JP, et al. The characterization of twenty sequenced human genomes. PLoS Genet. 2010;6: e1001111.
37. Collins SC, Bray SM, Suhl JA, Cutler DJ, Coffee B, Zwick ME, et al. Identification of novel FMR1 variants by massively parallel sequencing in developmentally delayed males. Am J Med Genet A. 2010;152A: 2512–2520.
38. Lotta LA, Wang M, Yu J, Martinelli I, Yu F, Passamonti SM, et al. Identification of genetic risk variants for deep vein thrombosis by multiplexed next-generation sequencing of 186 hemostatic/pro-inflammatory genes. BMC Med Genomics. 2012;5: 7.
39. Lotta LA, Tuana G, Yu J, Martinelli I, Wang M, Yu F, et al. Next-generation sequencing study finds an excess of rare, coding single-nucleotide variants of ADAMTS13 in patients with deep vein thrombosis. J Thromb Haemost . 2013;11: 1228–1239.

40. Quélin F, Mathonnet F, Potentini-Esnault C, Trigui N, Peynet J, Bastenaire B, et al. Identification of five novel mutations in the factor XI gene (F11) of patients with factor XI deficiency. Blood Coagul Fibrinolysis. 2006;17: 69–73.

Artículo 5

Título

The Unravelling of the Genetic Architecture of Plasminogen Deficiency and its Relation to Thrombotic Disease.

Autores

Laura Martin-Fernandez, Pascual Marco, Irene Corrales, Raquel Pérez, Lorena Ramírez, Sonia López, Francisco Vidal y José Manuel Soria.

Referencia

Scientific Reports 2016; 6:39255. doi: 10.1038/srep39255. PMID: 27976734.

Resumen

El plasminógeno es el componente principal del sistema fibrinolítico y diversas alteraciones genéticas en el gen estructural *PLG* se han relacionado con la deficiencia de plasminógeno. Nuestro objetivo ha consistido en la identificación de las alteraciones genéticas en el gen *PLG* causantes de la deficiencia de plasminógeno en familias reclutadas a partir de un individuo afecto que padeció también un evento trombótico espontáneo. Además, se ha evaluado si la deficiencia de plasminógeno debería considerarse un factor de riesgo de trombosis, ya que existen publicaciones con resultados contradictorios. Para esto, se han estudiado 13 individuos de un total de 4 familias. En concreto, se ha caracterizado el perfil genético de trombofilia mediante el kit Thrombo inCode, a través del cual se ha identificado una familia portadora homocigota para alelos de riesgo en el gen *F12* (rs1801020) y en el *F13A1* (rs5985). Por otra parte, se ha secuenciado *PLG* mediante NGS a partir de LR-PCR solapadas y

se han realizado estudios *in silico* de predicción de funcionalidad, a partir de los cuales se han destacado 5 potenciales mutaciones patogénicas relacionadas con la deficiencia de plasminógeno. Estas mutaciones incluyen 3 variantes genéticas de cambio de sentido y 2 potenciales mutaciones de *splicing*. Por lo tanto, hemos podido contribuir al conocimiento de los determinantes genéticos implicados en la deficiencia de plasminógeno mediante la metodología de NGS. A pesar de que no se han identificado factores genéticos relacionados con el riesgo de padecer trombosis en 3 de las 4 familias estudiadas y de que no se ha podido establecer una cosegregación de dichas potenciales mutaciones patogénicas localizadas en el gen *PLG* con los eventos trombóticos sufridos, no podemos descartar la deficiencia de plasminógeno como factor de riesgo de trombosis, al considerar que la trombosis es una enfermedad compleja y multifactorial en la que factores de riesgo desconocidos, además de la deficiencia de plasminógeno, podrían estar implicados en la susceptibilidad para padecer trombosis en estas familias.

www.nature.com/scientificreports

OPEN

The Unravelling of the Genetic Architecture of Plasminogen Deficiency and its Relation to Thrombotic Disease

Received: 01 August 2016
Accepted: 22 November 2016
Published: 15 December 2016

Laura Martin-Fernandez[1], Pascual Marco[2], Irene Corrales[3,4], Raquel Pérez[1], Lorena Ramírez[3,4], Sonia López[1], Francisco Vidal[3,4] & José Manuel Soria[1]

Although plasminogen is a key protein in fibrinolysis and several mutations in the plasminogen gene (*PLG*) have been identified that result in plasminogen deficiency, there are conflicting reports to associate it with the risk of thrombosis. Our aim was to unravel the genetic architecture of *PLG* in families with plasminogen deficiency and its relationship with spontaneous thrombotic events in these families. A total of 13 individuals from 4 families were recruited. Their genetic risk profile of thromboembolism was characterized using the Thrombo inCode kit. Only one family presented genetic risk of thromboembolism (homozygous carrier of *F12* rs1801020 and *F13A1* rs5985). The whole *PLG* was tested using Next Generation Sequencing (NGS) and 5 putative pathogenic mutations were found (after *in silico* predictions) and associated with plasminogen deficiency. Although we can not find genetic risk factors of thrombosis in 3 of 4 families, even the mutations associated with plasminogen deficiency do not cosegregated with thrombosis, we can not exclude plasminogen deficiency as a susceptibility risk factor for thrombosis, since thrombosis is a multifactorial and complex disease where unknown genetic risk factors, in addition to plasminogen deficiency, within these families may explain the thrombotic tendency.

The plasminogen protein plays a pivotal role in fibrinolysis and wound healing. Briefly, this protein generates the active enzyme plasmin by tissue-type plasminogen activator (t-PA) or urokinase-type plasminogen activator (u-PA). Plasmin, which is a serine protease enzyme, is essential for the dissolution of blood clots in a fibrin-dependent manner. In addition, plasmin has different substrate specificities and functions that are important in wound healing, such as other matrix glycoproteins or the activation of matrix metalloproteinases[1]. The plasminogen protein is encoded by the PLG gene, which is located on Chromosome 6q26, has 19 exons and is 51,861 bp in length (http://genome.ucsc.edu/; February 2009 release of the human genome, GRCh37/hg19)[2]. Numerous mutations have been described in the *PLG* locus that cause plasminogen deficiency, such as missense, nonsense, frameshift, splice site, deletion and insertion mutations[3,4].

Plasminogen deficiency is a rare disorder that has been classified as type I or hypoplasminogenemia, described first by Hasegawa *et al.* in ref. 5, and type II or dysplasminogenemia, which was described for the first time by Aoki *et al.* in ref. 6. In type I, both plasminogen activity and antigen levels are decreased. In contrast, type II is characterized by decreased plasminogen activity but normal antigen levels[4,7]. Both functional and antigen assays have been used in clinical diagnosis. Molecular genetic analysis has supported the clinical diagnosis, and identified individuals at risk for plasminogen deficiency. It may be used also for prenatal diagnosis[8]. Currently, different approaches have been reported for testing for genetic mutations: single-strand conformation polymorphism analysis of PCR products encompassing exons and intron boundaries, restriction fragment length polymorphism technique and Sanger sequencing of both PCR amplicons of exons and flanking intronic regions and long range (LR) PCR amplicons[4,9,10].

[1]Unit of Genomics of Complex Diseases, Biomedical Research Institute Sant Pau (IIB-Sant Pau), Barcelona, Spain. [2]Department of Haematology, Hospital General Universitario, Alicante, Spain. [3]Congenital Coagulopathies, Blood and Tissue Bank, Barcelona, Spain. [4]Molecular Diagnosis and Therapy, Vall d'Hebron Research Institute, Universitat Autònoma de Barcelona (VHIR-UAB), Barcelona, Spain. Correspondence and requests for materials should be addressed to J.M.S. (email: JSoria@santpau.cat)

Family	Relationship to proband	Age (years)	Age at first episode of thrombosis (years)	Sex	Thrombotic events	Relevant data for thrombotic disease*	Plasminogen activity (%)†
A	Proband	49	49	F	DVT	Negative	24
	Son	22	—	M	No	Negative	67
	Son	13	—	M	No	Negative	58
B	Proband	46	25	F	DVT	Negative	47
	Mother	74	—	F	TIA	Positive APLA	67
	Daughter	25	—	F	No	Negative	50
	Sibling	51	—	F	No	Negative	58
	Sibling	44	—	F	No	Negative	82
C	Proband	43	43	M	DVT	Negative	68
	Sibling	31	—	F	No	Negative	64
	Mother	63	—	F	No	Negative	64
	Maternal aunt	60	—	F	No	Negative	101
D	Proband	50	48	F	IS	Positive APLA	55

Table 1. Phenotypic data of the four families with plasminogen deficiency. F: female; M: male; DVT: deep venous thrombosis; TIA: transient ischemic attack; IS: ischemic stroke; APLA: antiphospholipid antibodies. *Thrombophilia investigations included lupus anticoagulant, antiphospholipid antibodies of immunoglobulin G (IgG) and IgM types, homocysteine, antithrombin, Protein C measured with two functional tests and Protein S (free) measurements. †Normal range of plasminogen activity: 72% to 127%.

Plasminogen deficiency results in a reduction in the degradation of fibrin and thus affects wound healing. For example, ligneous conjunctivitis, congenital occlusive hydrocephalus and juvenile colloid milium have been reported[1,3]. Among these, ligneous conjunctivitis is the most common and is characterized by chronic tearing, erythema of the conjunctiva and white, yellow-white, or red thick masses with a wood-like consistency. This chronic conjunctivitis is associated with homozygous or compound heterozygous type I patients. Of note, heterozygous type I and type II patients are asymptomatic[3]. Some studies[6,11] suggested that there was an association between plasminogen deficiency and thrombosis due to impaired fibrinolysis. However, these studies involved few patients of sympthomatic thrombophilia. In addition, most family studies showed that only the probands had thrombosis. It is noteworthy that some risk factors of thrombosis such as activated protein C (APC) resistance were unknown when this association was reported[12]. More recent reports supports the hypothesis that this disorder by itself is not a risk factor of venous or arterial thrombosis[3,12,13]. Also, it is suggested that the diminished plasminogen capability of fibrinolysis may be compensated by the action of alternative enzymes in the blood[3].

In our study, we identified 4 Spanish patients that had suffered spontaneous thrombotic events. We identified a plasminogen deficiency in these patients and in several of their relatives. Our aims were to characterize the genetic risk factors of thromboembolism in these families and to detect the genetic mutations that were involved in the plasminogen deficiency. We used a new method to diagnose the deficiency based on Next Generation Sequencing (NGS) of *PLG*. We evaluated also the relation of these causal genetic mutations to the thrombotic outcomes to elucidate the role of plasminogen deficiency in thrombotic disease.

Material and Methods

Subjects. A total of 13 individuals from 4 Spanish families (A, B, C and D) were recruited at the Hospital General Universitario de Alicante. These families were selected through probands with a positive history of spontaneous thrombosis and plasminogen deficiency. The latter was defined as a functional activity below 72% (normal range, 72% to 127%). The characterization of these families is presented in Table 1. Clinical manifestations related to plasminogen deficiency were not identified and no classification of type I or type II was performed. All of the probands and one relative had experienced a spontaneous deep venous thrombosis of the legs or an arterial thrombotic event. The number of individuals per family varied from 1 to 5 and they ranged in age from 13 to 74 years. A total of 9 individuals were selected from the 4 families for genetic analysis, which included the probands and individuals with the lowest functional plasminogen deficiency values. At least 2 relatives of every family were selected for inclusion, except for one family with no relatives.

The study is part of the clinical routine of the Hospital General Universitario de Alicante for thrombophilia patients conducted in accordance with the Good Practice Guidelines that included an informed consent from all individuals prior to inclusion. The studies were conducted in accordance with the Declaration of Helsinki.

Blood Collection and Phenotype Determinations. Blood samples were obtained from the antecubital vein and anticoagulated with 3.8% sodium citrate in the proportion 1:10. The blood samples were centrifuged for 10 minutes at 3,500 g within 15 minutes after collection. The poor-platelet plasma (PPP) samples were immediately frozen and stored at −80 °C until tested. DNA was extracted from whole blood collected in EDTA using a standard salting out procedure[14]. The functional plasminogen activity was performed in PPP using a chromogenic substrate assay, activated by tissue plasminogen activator (Instrumentation Laboratory, Werfen Group, MA, USA).

PCR	PCR forward primer	Forward primer position*	PCR reverse primer	Reverse primer position*	Size (bp)
PLG_LR1	GCGCCAGCACAGAGCTCTGCTCAAC	chr6:161,121,602- 161,121,626	GCTTGCTACTTGTAAGAACTAATAACTTC	chr6:161,129,932-161,129,960	8,359
PLG_LR2	CAATTTAGCTCTCCAAACATTCTGCATCC	chr6:161,129,653-161,129,681	GAACTCCTTGAACTCACTCAAGCAATCC	chr6:161,137,405-161,137,432	7,780
PLG_LR3	AGGTACTAGATGAGTATCTTTAGGCAGG	chr6:161,137,170-161,137,197	GGCATCTCCATTGAGCTCACTGTTCC	chr6:161,144,785-161,144,810	7,641
PLG_LR4	GTCTTGCTGAACTGAGGAAGGAGACTGG	chr6:161,144,556-161,144,583	GTAGATGTCCAGCCACAGATTTCTTACC	chr6:161,152,264-161,152,291	7,736
PLG_LR5	CTGAATATTCTCCCACCTCTTGTGACC	chr6:161,152,034-161,152,060	GAACAGCTCTGTTCTGCAGTTTATTCAGG	chr6:161,159,857-161,159,885	7,852
PLG_LR6	CCCACTGCTTGGAGAAGTATGTTTAGG	chr6:161,159,629-161,159,655	TCAAGATCCAGGAGGATGGCAAAATCC	chr6:161,167,598-161,167,624	7,996
PLG_LR7	CATTTCTGAGATTCTTTCCTCAGCTTGG	chr6:161,167,413-161,167,440	TCAAAATGGGGAACAACATTTGTGTAAGG	chr6:161,175,204-161,175,232	7,820

Table 2. Primers used for LR-PCR amplification of *PLG*. *Primer positions referred to GRCh37/hg19, UCSC Genome Browser assembly, February 2009 release (http://genome.ucsc.edu/)[2].

Genetic Profile of Thromboembolism. The Thrombo inCode (TiC) kit (Ferrer in Code, Barcelona, Spain) was used to determine the genetic profile of thrombosis. The TiC kit identified the alleles of 12 genetic variants: *F5* (rs6025, rs118203906, rs118203905), *F2* rs1799963, *F12* rs1801020, *F13A1* rs5985, *SERPINC1* rs121909548, *SERPINA10* rs2232698 and the A1 blood group (rs8176719, rs7853989, rs8176743 and rs8176750). This genetic profile has been validated and reported in the literature[15]. The TiC kit was applied by Taqman assays run in an ABI 7500 instrument according to manufacturer's instructions.

Primer Design and PCR Amplification. A total of 55,184 bp of the *PLG* locus (GRCh37/hg19 chr6:161,123,225-161,175,085; NM_000301.3) were amplified using 7 LR-PCR amplicons. The primer sequences, positions and PCR amplicon size are shown in Table 2. The PCR primers were designed to specifically amplify the *PLG* and to avoid co-amplifications in view of the high homology between *PLG*, *LPA* family genes, pseudogenes and plasminogen-like genes[10]. The overlapped PCR amplicons covered all of the exons, introns, 5′-UTR, 3′-UTR and approximately 1,500 bp of the 5′-promoter region. The PCR amplicons designed were tested for target specificity by Sanger sequencing as described below.

LR-PCR amplicons were generated using the SequalPrep Long PCR Kit with dNTPs (Invitrogen, Thermo Fisher Scientific Inc., MA, USA). The LR-PCR mix solution contained ~50 ng of DNA, SequalPrep 1X reaction buffer, 0.4 μl of dimethylsulfoxide (DMSO), SequalPrep 1X enhancer B, 0.75 μM of forward and reverse primers and 1.8 units of SequalPrep Long Polymerase in a total volume of 20 μl. After initial denaturation at 94 °C for 2 minutes (min), 10 cycles of 94 °C for 10 seconds (sec), 64 °C for 30 sec, and 68 °C for 18 min were performed, followed by 22 cycles of 94 °C for 10 sec, 64 °C for 30 sec, and 68 °C for 18 min (+20 sec/cycle). In addition, an elongation step at 72 °C for 5 min was performed.

Every PCR amplicon was run on 0.7% agarose gel electrophoresis and visualized using SYBR safe (Invitrogen, Thermo Fisher Scientific Inc.). PCR amplicons were quantified by the Qubit technology (Invitrogen, Thermo Fisher Scientific Inc.) and a normalized pool of the 7 PCR amplicons was prepared for each individual by mixing equimolar amounts. Finally, the PCR pools were adjusted at 0.2 ng/μl for preparing the libraries.

Library Preparation and NGS. The sequencing libraries were prepared from pooled PCR amplicons using the Nextera XT DNA Sample Preparation kit (Illumina, San Diego, CA, USA) with double indexing, following the standard manufacturer's protocol. We obtained 9 paired-end libraries that were pooled and run simultaneously on an Illumina Miseq sequencing system (Illumina) by the Miseq sequencing reagent kit v2 of 300 cycles (2 × 150 bp paired-end) (Illumina).

Bioinformatic Analysis. Indexed sequences were de-multiplexed and analyzed individually. The NGS pipeline output, paired sequence files (fastq files format), was used as input for the analysis with the CLC Genomic Workbench (v.6.5) software (CLC Bio - Qiagen, Aarhus, Denmark). The raw data were trimmed with length (minimum 25 bp; maximum 500 bp), ambiguous nucleotide (maximum 2) and quality score (0.05) filters. CLC Genomic Workbench software permitted the alignment of the trimmed reads against the human genome reference (hg19) and *in silico* analysis. Read mapping was performed with specific parameter setting (mismatch count, 2; indel count, 3; length fraction, 0.7; similarity fraction, 0.9). Indels and structural variants tool was used to identify insertions and deletions, inversions, translocations and tandem duplications, applying standard settings. Moreover, adjusted parameters for quality-based variant detection were as follows: minimum coverage, 30x; minimum variant frequency, 25%. Quality-based variant detection results in variant call format (VCF) file were used as input for the Illumina VariantStudio Data Analysis (v.2.1) Software (Illumina) to annotate variants.

Screening Putative Pathogenic Mutations. For structural variants, we used the following filter parameters to detect pathogenic variations: a) minimum variant ratio, 0.25, b) minimum mapping scores fraction, 0.6, c) consider intronic structural variants located <30 bp from exon flanking boundaries, d) allele frequency from our own NGS variants database of 110 individuals, ≤5%, and e) co-segregation in the family.

The following criteria were applied to identify putative pathogenic mutations within single nucleotide variants (SNV) and small insertions and deletions: a) whether the variant was rare: allele frequency ≤1% from 1000 Genomes (April 2012 v.3)[16] and NHLBI Exome Variant Server (June 2013 ESP6500SI-V2), b) location referring to Variant Effect Predictor (v.2.8) database: intronic mutations located <30 bp into flanking intronic sequence of

Family	Relationship to proband	Plasminogen activity (%)*	*PLG* mutations					Major mutations involved in thrombotic risk†
			Exon	Nucleotide change	Amino acid change	Genotype	MAF (%)	
A	Proband	24	2	c.112 A>G	p.Lys38Glu	Compound het in trans	0.27	—
			18	c.2134 G>A	p.Gly712Arg		0	—
	Son	67	2	c.112 A>G	p.Lys38Glu	Het	0.27	—
	Son	58	18	c.2134 G>A	p.Gly712Arg	Het	0	—
B	Proband	47	2	c.112 A>G	p.Lys38Glu	Het	0.27	—
	Mother	67	2	c.112 A>G	p.Lys38Glu	Het	0.27	—
	Daughter	50	2	c.112 A>G	p.Lys38Glu	Het	0.27	—
	Sibling	58	2	c.112 A>G	p.Lys38Glu	Het	0.27	—
	Sibling	82	Non-carrier	Non-carrier	Non-carrier	Non-carrier		—
C	Proband	68	7	c.781 C>T	p.Arg261Cys	Het	0	—
	Sibling	64	7	c.781 C>T	p.Arg261Cys	Het	0	—
	Mother	64	7	c.781 C>T	p.Arg261Cys	Het	0	—
	Maternal aunt	101	Non-carrier	Non-carrier	Non-carrier	Non-carrier	0	—
D	Proband	55	1	c.12 G>A	p.Lys4Lys	Compound het	0	*F12* (c.-4T>T)
			—	c.1878-6T>C	—		0.14	*F13* (c.103 G>G)

Table 3. Molecular data of the four families with plasminogen deficiency. MAF: minor allele frequency from all populations of 1000 genomes data (April 2012 v.3; www.1000genomes.org)[16]; Het:: heterozygous. *Normal range of plasminogen activity: 72% to 127%. †Gene (nucleotide change).

each exon were included, c) allele frequency from our own NGS variants database of 110 individuals, ≤5%, in the case of variants not annotated in 1000 Genomes (April 2012 v.3)[16], and d) co-segregation in the family.

We defined as "new" those putative pathogenic mutations that were not reported in the literature related to the plasminogen deficiency disorder. The amino acid numbering and the nomenclature used for the description of the sequence variations follows the international recommendations of the Human Genome Variation Society (HGVS; http://www.HGVS.org).

Sanger Validations. Putative pathogenic mutations were validated by Sanger sequencing[17] in probands and family members. The sequences were mapped against the *PLG* locus (GRCh37/hg19 chr6:161,123,225-161,175,085; NM_000301.3) using the CLC Genomic Workbench (v.6.5) software.

***In Silico* Analyses.** The *in silico* prediction that evaluate functional effects of putative pathogenic mutations was performed using the Alamut Visual (v.2.6.1) software (Interactive Biosoftware, Rouen, France). Changes in splicing sites were predicted using NNSplice and Human Splicing Finder tools. For splicing variants interpretation, a splicing site change was considered as potentially deleterious when a variation between the native and the mutation score of more than 10% was observed in both algorithms[18]. Missense prediction tools included SIFT, PolyPhen-2, Align GVGD and Mutation Taster. Evolutionary conservation scores were obtained using phyloP and Grantham distances. Interpretation of predictive structural effects of new missense mutations was investigated by using the Project HOPE software[19].

Results

Genetic profile of thromboembolism. We characterized the alleles of 12 genetic variants located within *F2*, *F5*, *F12*, *F13A1*, *ABO*, *SERPINA10* and *SERPINC1* loci, included in the TiC kit. Based on these results, all of the individuals were reported as not at genetic risk of thrombotic disease except for the proband of Family D (Table 3). This individual was homozygous for the alleles c.-4T in the *F12* (rs1801020), also known as 46 T risk allele[15,20], and c.103 G>G of the *F13A1* (rs5985), also known as Val34 risk allele[15].

NGS results. We amplified 55,184 bp encompassing the *PLG* locus using LR-PCR from 9 individuals. These LR-PCR amplicons were analysed by NGS. Briefly, the percentage of the mapped positions with a depth of coverage above 30x was 95% and the median coverage per individual was 212x. In total, 237 potential structural variants and 301 unique SNV and small insertions and deletions were called. Among the latter, the percentage of indels and exonic variants was 12.6% (38) and 3.7% (11), respectively, and the 55.5% (167) had an allele frequency ≤1% from all population of 1000 Genomes (April 2012 v.3) and from four populations of 1000 Genomes (American, East Asian, African and European). In addition, the 47.8% (144) of the variants had not been reported in dbSNP (v. 137) and 143 out of 144 were within the introns.

Putative pathogenic mutations. We identified 5 putative pathogenic mutations, including 3 missense variations and 2 potential splicing site mutations (1 synonymous and 1 intronic variants) reported in Table 3. Within the family members, the 5 candidate mutations were confirmed by Sanger sequencing if genomic DNA

Nucleotide Change	Protein Change	Variant Type	dbSNP v137	NNSPLICE score*	HSF score*	SIFT score (median)	PolyPhen-2 score (sensitivity - specificity)	Align GVGD score	Mutation Taster p-value	PhyloP	Grantham distances
c.12 G>A	p.Lys4Lys	PSSM	rs4252061	1.00–1.00	97.66–97.66	—	—	—	—	0.69	—
c.112 A>G	p.Lys38Glu	Missense	rs73015965	—	—	0.005 (3.37)	0.879 (0.82–0.94)	C55 (GV:0–GD:56.87)	0†	1.78	56
c.781 C>T	p.Arg261Cys	Missense	—	—	—	0 (3.84)	0.999 (0.14–0.99)	C65 (GV:0–GD:179.53)	1	4	180
c.1878-6T>C	—	PSSM	rs192519670	0.98–0.98	84.28–84.93	—	—	—	—	−0.134	—
c.2134 G>A	p.Gly712Arg	Missense	rs202074006	—	—	0.03 (3.84)	1 (0.00–1.00)	C65 (GV:0–GD:125.13)	0.99	1.25	125

Table 4. *In silico* predictions of the *PLG* identified mutations. NNSPLICE [0-1]: threshold ≥0.4 and HSF [0-100]: Human Splicing Finder, threshold ≥65. In both programs a splice site effect was considered as potentially deleterious when a variation between the native and the mutation score was more than 10%; SIFT [1-0]: scores less than 0.05 indicate substitutions are predicted as deleterious; Polyphen-2 [0-1]: scores range from 0.000 (most probably benign) to 1 (most probably damaging); Align GDGV [C0-C65]: scores range from Class C0 (less likely deleterious) to Class C65 (most likely deleterious); Mutation Taster [0-1]: from disease causing variants (p-value = 1.0) to might not be disease causing (p-value <0.99); PhyloP [-14.1 to 6.4]: from highly conserved (score >3) to moderately conserved (score 1-3) or poorly conserved (score <1); Grantham distances [0-215]: from highly different physicochemical properties and more probably damaging (score >50) to moderately (score 25-50) or poorly different and most probably tolerated (score <25); PSSM: Potential splicing site mutation. *Score native – mutation. †This variant is listed as pathogenic in NCBI ClinVar database (http://www.ncbi.nlm.nih.gov/clinvar/). Thus, it is automatically predicted to be disease-causing in Mutation Taster but real probability is shown.

was available. Also, we performed *in silico* predictions (Table 4). Specifically, any structural variant passed through the filtering criteria for putative pathogenic structural variant detection.

In Family A, which was composed of 3 members, the proband was compound heterozygous in trans for p.Lys38Glu (in exon 2) and p.Gly712Arg (in exon 18) missense variations in *PLG*. This individual had a functional plasminogen activity level of 24%. The variant p.Lys38Glu was also heterozygous in the son with a level of functional plasminogen activity of 67% while another son who was heterozygous for the p.Gly712Arg variation had a functional plasminogen activity of 58%. Of note, both missense variations have been described previously[1,3,21] in association with plasminogen deficiency.

The p.Lys38Glu variation was identified also in Family B. This missense mutation was detected as heterozygous in 4 of the 5 members. These 4 individuals included the proband with a functional plasminogen activity level of 47% and 3 family members with functional plasminogen activity level of 58%, 50% and 67%. In contrast, the non-carrier family member showed a normal level of functional plasminogen activity (82%).

In Family C, the new putative pathogenic variation p.Arg261Cys (in exon 7) was heterozygous in 3 of the family members. The functional plasminogen activity was low in these individuals, who included the proband (68%), his sibling (64%) and his mother (64%), in comparison to the level detected in the maternal aunt of the proband, who was a non-carrier family member (101%). The effects of this missense mutation predicted by SIFT, Polyphen-2, Align GDGV and Mutation Taster were deleterious, probably damaging, most likely interfering with protein function and disease causing, respectively. In addition, it was reported as a highly conserved nucleotide and there were large physicochemical differences between Arg and Cys. Project HOPE software revealed differences between the native and the mutant residues in size and charge. The cysteine residue was smaller than the native arginine residue and the positive charge of the arginine was lost as a result of this mutation. Furthermore, the mutated cysteine was located very close to a residue that makes a cysteine bond, and despite this bond which itself is not mutated, it could be affected by the mutation located in its vicinity. The missense variation p.Arg261Cys was characterized as part of a Kringle domain with hydrolase activity, which is essential for the activity of the protein and involved in protein-protein interactions.

The last proband (Family D) had a functional plasminogen activity of 55%. The mutational analysis showed that this individual was compound heterozygous for two potential splice site mutations. Specifically, the p.Lys4Lys synonymous variant in exon 1, which was located in the signal peptide, and the c.1878-6 T>C variant in intron 15. To investigate *in silico* functional consequences of these new putative pathogenic mutations predictive changes in splicing sites were analyzed. By doing so, no alterations at the donor splice site of exon 1 or at the natural acceptor splice site of exon 16 were detected for the synonymous and the intronic variants, respectively.

Discussion

We present the molecular characterization of 13 individuals from 4 plasminogen deficiency families, in which the probands had suffered a thrombotic event. Since there is doubt as to whether plasminogen deficiency is a risk factor of thrombotic disease by itself or in combination with other abnormalities[1], we genotyped 12 genetic risk factors of thrombosis coded by *F2*, *F5*, *F12*, *F13A1*, *ABO*, *SERPINA10* and *SERPINC1* loci to evaluate the putative role of plasminogen deficiency phenotype in thrombotic disease. We identified a genetic risk profile of thromboembolism in Family D. The proband was homozygous for the risk alleles in *F12* (rs1801020) and *F13* (rs5985). Interestingly, this individual had positive antiphospholipid antibodies also that might have acted as a risk factor

in her ischemic stroke. In contrast, Families A, B and C did not show genetic risk profiles of thromboembolism despite suffering thrombotic episodes. Thus, plasminogen deficiency could not be ruled out as an additional risk factor. It is noteworthy that thromboembolism is a common disease with more than 60% of the variation due to genetic risk factors[22,23]. However, genetic scores explain only 15% of the variance[15], so there is still a "missing heritability" that might have hampered the discrimination of plasminogen deficiency as an additional risk factor of thrombotic disease. Also, whether individuals of these families other than the probands will suffer a thrombotic event in the future is not known.

We sequenced the whole *PLG* locus with good coverage of high quality reads to identify the genetic variants that might be involved in the functional plasminogen deficiency of these families using the NGS methodology. Recent advances in NGS have provided high-throughput, economical, sensitive and faster sequencing methodologies[24,25]. Because of these advantages, and the fact that many samples can be analyzed simultaneously, NGS is ideal for targeted diagnostic sequencing of monogenic diseases where genetic heterogeneity is expected. New pathogenic mutations have been discovered using NGS[26,27] but to our knowledge, no publications have explored the genetic basis of plasminogen deficiency using NGS. We performed whole gene sequencing to exhaustively examine not only exons but also non-coding regions (promoter, introns, UTR) that might be involved in regulatory functions[28–30]. In addition, LR-PCR provided a fast and cost-effective technique for avoiding the amplification of plasminogen pseudogenes[31,32]. Also, LR-PCR in combination with paired-end sequencing allowed us to use algorithms to accurately detect structural variants and avoid the use of additional analyses as MLPA (multiplex ligation-dependent probe amplification) or MAPH (multiplex amplifiable probe hybridization).

We have identified 5 putative pathogenic mutations, which are in concordance with the levels of plasminogen functional activity by intrafamilial analyses. In particular, the mutation p.Lys38Glu (K19E, old nomenclature) has been described[1,3,21] widely as the most common molecular genetic defect in association with type I. This mutation has been identified in homozygous, compound heterozygous and heterozygous state[4] but it is not known why this mutation and others in the *PLG* leads to diverse clinical conditions[8]. In addition, the missense variation p.Gly712Arg (G693R, old nomenclature) has been described in heterozygous state related to type II plasminogen deficiency[21]. Interestingly, we have identified both the p.Lys38Glu and p.Gly712Arg variants as compound heterozygous in one individual. In contrast, for the first time to our knowledge the p.Arg261Cys variation is described in association with plasminogen deficiency. This missense variation was classified as deleterious using *in silico* predictions. Several structural effects in the plasminogen protein have been suggested by Project HOPE. Specifically, changes in size and charge between the native and the mutated residue were predicted, which could cause loss of protein-protein interactions. In addition, Project HOPE software predicted that p.Arg261Cys mutation could disturb the interaction between domains, which might affect the protein's function also. This variant has been listed previously as variant 6:161137789 C / T in the Exome Aggregation Consortium (ExAC) Browser (Cambridge, MA, http://exac.broadinstitute.org) with an allele frequency of 5.771×10^{-5}. It is noteworthy that another variation in the same amino acid (c.782 G>A, p.Arg261His, rs4252187) has been annotated in NCBI dbSNP (v.146, http://www.ncbi.nlm.nih.gov/SNP/). Also, we identified the synonymous p.Lys4Lys and the intronic c.1878-6 T>C variation in one individual as new putative pathogenic mutations in association with plasminogen deficiency. Both genetic variations were not predicted by *in silico* analyses to change the splice site natural junction. Despite this, synonymous and intronic variations could cause alterations in protein expression and function[33,34]. Thus, further analyses are needed to determine the functional characterization of these mutations.

It is important to note that we identified the whole spectrum of genetic variability of the *PLG* structural gene in these families. However, we did not observe a clear association between the genotype, the number or the type of putative pathogenic mutations in *PLG* with the thrombotic phenotype of these individuals. Moreover, we can not find genetic risk factors of thrombosis in 3 of 4 families. These observations provide us with a scenario where we can not confirm neither exclude plasminogen deficiency as a susceptibility risk factor for thrombosis.

Our view is that thrombosis is a multifactorial and complex disease where several genetic risk factors with environmental situations increase the susceptibility to develop a thromboembolic event. Therefore, in these families unknown genetic risk factors, in addition to plasminogen deficiency, may explain the thrombotic tendency in concrete clinical situations. Thus, we believe that it may be useful the evaluation of plasminogen deficiency in individuals that had suffered a spontaneous thrombotic event and had a negative routing thrombophilia test.

It is important to emphasize that, from a clinical point of view, we detected putative pathogenic mutations that explain the plasminogen deficiency. In these sense, our NGS approach clearly contributed to the genetic knowledge of plasminogen deficiency. We believe that NGS has the potential to identify and characterize the molecular basis of a wide variety of disorders in addition to plasminogen deficiency.

References

1. Mehta, R. & Shapiro, A. D. Plasminogen deficiency. *Haemophilia* **14,** 1261–1268 (2008).
2. Karolchik, D. *et al.* The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res.* **42,** D764–D770 (2014).
3. Schuster, V., Hügle, B. & Tefs, K. Plasminogen deficiency. *J. Thromb. Haemost.* **5,** 2315–2322 (2007).
4. Tefs, K. *et al.* Molecular and clinical spectrum of type I plasminogen deficiency: A series of 50 patients. *Blood* **108,** 3021–3026 (2006).
5. Hasegawa, D., Tyler, B. & Edson, J. Thrombotic disease in three families with inherited plasminogen deficiency. *Blood* **60,** 213a (1982).
6. Aoki, N., Moroi, M., Sakata, Y., Yoshida, N. & Matsuda, M. Abnormal plasminogen. A hereditary molecular abnormality found in a patient with recurrent thrombosis. *J. Clin. Invest.* **61,** 1186–1195 (1978).
7. Lane, D. A. *et al.* Inherited thrombophilia: Part 1. *Thromb. Haemost.* **76,** 651–662 (1996).
8. Schuster, V. & Seregard, S. Ligneous conjunctivitis. *Surv. Ophthalmol.* **48,** 369–388 (2003).
9. Schuster, V. *et al.* Homozygous and compound-heterozygous type I plasminogen deficiency is a common cause of ligneous conjunctivitis. *Thromb. Haemost.* **85,** 1004–1010 (2001).
10. Siboni, S. M., Spreafico, M., Menegatti, M., Martinelli, I. & Peyvandi, F. Molecular characterization of an Italian patient with plasminogen deficiency and ligneous conjunctivitis. *Blood Coagul. Fibrinolysis* **18,** 81–84 (2007).

11. Leebeek, F. W., Knot, E. A., Ten Cate, J. W. & Traas, D. W. Severe thrombotic tendency associated with a type I plasminogen deficiency. *Am. J. Hematol.* **30,** 32–35 (1989).
12. Demarmels Biasiutti, F. *et al.* Is plasminogen deficiency a thrombotic risk factor? A study on 23 thrombophilic patients and their family members. *Thromb. Haemost.* **80,** 167–170 (1998).
13. Okamoto, A. *et al.* Population-based distribution of plasminogen activity and estimated prevalence and relevance to thrombotic diseases of plasminogen deficiency in Japanese: the Suita Study. *J. Thromb. Haemost.* **1,** 2397–2403 (2003).
14. Miller, S. A., Dykes, D. D. & Polesky, H. F. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res.* **16,** 1215 (1988).
15. Soria, J. M. *et al.* Multilocus Genetic Risk Scores for Venous Thromboembolism Risk Assessment. *J. Am. Heart Assoc.* **3,** e001060 (2014).
16. Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
17. Corrales, I., Ramírez, L., Altisent, C., Parra, R. & Vidal, F. Rapid molecular diagnosis of von Willebrand disease by direct sequencing. Detection of 12 novel putative mutations in VWF gene. *Thromb. Haemost.* **101,** 570–576 (2009).
18. Théry, J. C. *et al.* Contribution of bioinformatics predictions and functional splicing assays to the interpretation of unclassified variants of the BRCA genes. *Eur. J. Hum. Genet.* **19,** 1052–1058 (2011).
19. Venselaar, H., te Beek, T. A., Kuipers, R. K., Hekkelman, M. L. & Vriend, G. Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. *BMC Bioinformatics* **11,** 548 (2010).
20. Soria, J. M. *et al.* A quantitative-trait locus in the human factor XII gene influences both plasma factor XII levels and susceptibility to thrombotic disease. *Am. J. Hum. Genet.* **70,** 567–574 (2002).
21. Tefs, K. *et al.* A K19E missense mutation in the plasminogen gene is a common cause of familial hypoplasminogenaemia. *Blood Coagul. Fibrinolysis* **14,** 411–416 (2003).
22. Souto, J. C. *et al.* Genetic susceptibility to thrombosis and its relationship to physiological risk factors: the GAIT study. *Am. J. Hum. Genet.* **67,** 1452–1459 (2000).
23. Martin-Fernandez, L. *et al.* Genetic Determinants of Thrombin Generation and Their Relation to Venous Thrombosis: Results from the GAIT-2 Project. *PLoS One* **11,** e0146922 (2016).
24. Metzker, M. L. Sequencing technologies - the next generation. *Nat. Rev. Genet.* **11,** 31–46 (2010).
25. Desai, A. & Jere, A. Next-generation sequencing: ready for the clinics? *Clin. Genet.* **81,** 503–510 (2012).
26. Chen, X. *et al.* Targeted next-generation sequencing reveals novel USH2A mutations associated with diverse disease phenotypes: implications for clinical and molecular diagnosis. *PLoS One* **9,** e105439 (2014).
27. Chen, X. *et al.* Targeted sequencing of 179 genes associated with hereditary retinal dystrophies and 10 candidate genes identifies novel and known mutations in patients with various retinal diseases. *Invest. Ophthalmol. Vis. Sci.* **54,** 2186–2197 (2013).
28. Soria, J. M. *et al.* The F7 Gene and Clotting Factor VII Levels: Dissection of a Human Quantitative Trait Locus. *Hum. Biol.* **81,** 853–867 (2009).
29. Corrales, I., Ramírez, L., Altisent, C., Parra, R. & Vidal, F. The study of the effect of splicing mutations in von Willebrand factor using RNA isolated from patients' platelets and leukocytes. *J. Thromb. Haemost.* **9,** 679–688 (2011).
30. Gehring, N. H. *et al.* Increased efficiency of mRNA 3′ end formation: a new genetic mechanism contributing to hereditary thrombophilia. *Nat. Genet.* **28,** 389–392 (2001).
31. Tan, Y.-C. *et al.* A Novel Long-Range PCR Sequencing Method for Genetic Analysis of the Entire PKD1 Gene. *J. Mol. Diagnostics* **14,** 305–313 (2012).
32. De Sousa Dias, M. *et al.* Detection of novel mutations that cause autosomal dominant retinitis pigmentosa in candidate genes by long-range PCR amplification and next-generation sequencing. *Mol. Vis.* **19,** 654–664 (2013).
33. Sauna, Z. E. & Kimchi-Sarfaty, C. Understanding the contribution of synonymous mutations to human disease. *Nat. Rev. Genet.* **12,** 683–691 (2011).
34. Chorev, M. & Carmel, L.The function of introns. *Front. Genet.* **3,** 55 (2012).

Acknowledgements

We would like to thank Professor Bill Stone for reviewing the manuscript. This study was supported by funds from the Instituto de Salud Carlos III Fondo de Investigación Sanitaria PI 11/0184, PI 12/01494 and PI 14/0582 and Red Investigación Cardiovascular RD12/0042/0032 and RD12/0042/0053. Laura Martin-Fernandez was supported by Ayudas Predoctorales de Formación en Investigación en Salud (PFIS) FI12/00322.

Author Contributions

L. Martin-Fernandez contributed substantially to the acquisition, analysis and interpretation of the genetic data, critical writing and final approval of the paper version to be published. P. Marco participated in the acquisition and interpretation of clinical data. I. Corrales contributed substantially to the acquisition, analysis and interpretation of genetic data and final approval of the version to be published. R. Pérez, L. Ramírez and S. López contributed to the acquisition, analysis and interpretation of the genetic data. F. Vidal and J.M. Soria substantially contributed to the conception and design of this study, searched for funds, revised the intellectual content and final approval of the version to be published.

Additional Information

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Martin-Fernandez, L. *et al.* The Unravelling of the Genetic Architecture of Plasminogen Deficiency and its Relation to Thrombotic Disease. *Sci. Rep.* **6**, 39255; doi: 10.1038/srep39255 (2016).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Discusión

Discusión

La VTE es una enfermedad común y compleja en la que están involucrados factores de riesgo genéticos y factores de riesgo ambientales, así como la interacción entre éstos (Rosendaal 1999; Souto 2002). La predisposición a esta enfermedad tiene una heredabilidad muy alta que ha sido estimada en un 61% en el Proyecto GAIT-1 (Tabla 5). Sin embargo, a pesar del elevado componente genético que presenta, éste sigue siendo muy desconocido. Por lo tanto, existe una "heredabilidad perdida" de la que todavía no se dispone de una explicación fehaciente (Manolio et al. 2009). Por este motivo, el gran reto que se plantea actualmente en el estudio de las enfermedades complejas en general, y de la VTE en particular, es identificar todos los determinantes genéticos implicados. Esto puede suponer una mejora tanto a nivel diagnóstico como preventivo y terapéutico.

Con el propósito de contribuir en la identificación de la "heredabilidad perdida" del riesgo de VTE se han empleado varios enfoques y métodos de análisis que dan respuesta a distintas cuestiones genéticas, por lo que resultan complementarios entre sí y deben ser aplicados en función de los objetivos planteados. Estos métodos están en continua evolución y no existe una estrategia única para abordar el análisis de las bases genéticas de las enfermedades complejas. La estrategia global en la que hemos basado nuestro abordaje experimental se muestra en la Figura 9. En esta estrategia, el estudio en familias ha sido esencial para la obtención de los resultados. Cabe destacar que para los estudios aquí presentados hemos determinado una MAF ≥10% para denominar a una variante genética común, una 1%≤ MAF <10% para las variantes genéticas de baja frecuencia alélica y una MAF <1% para las variantes genéticas raras.

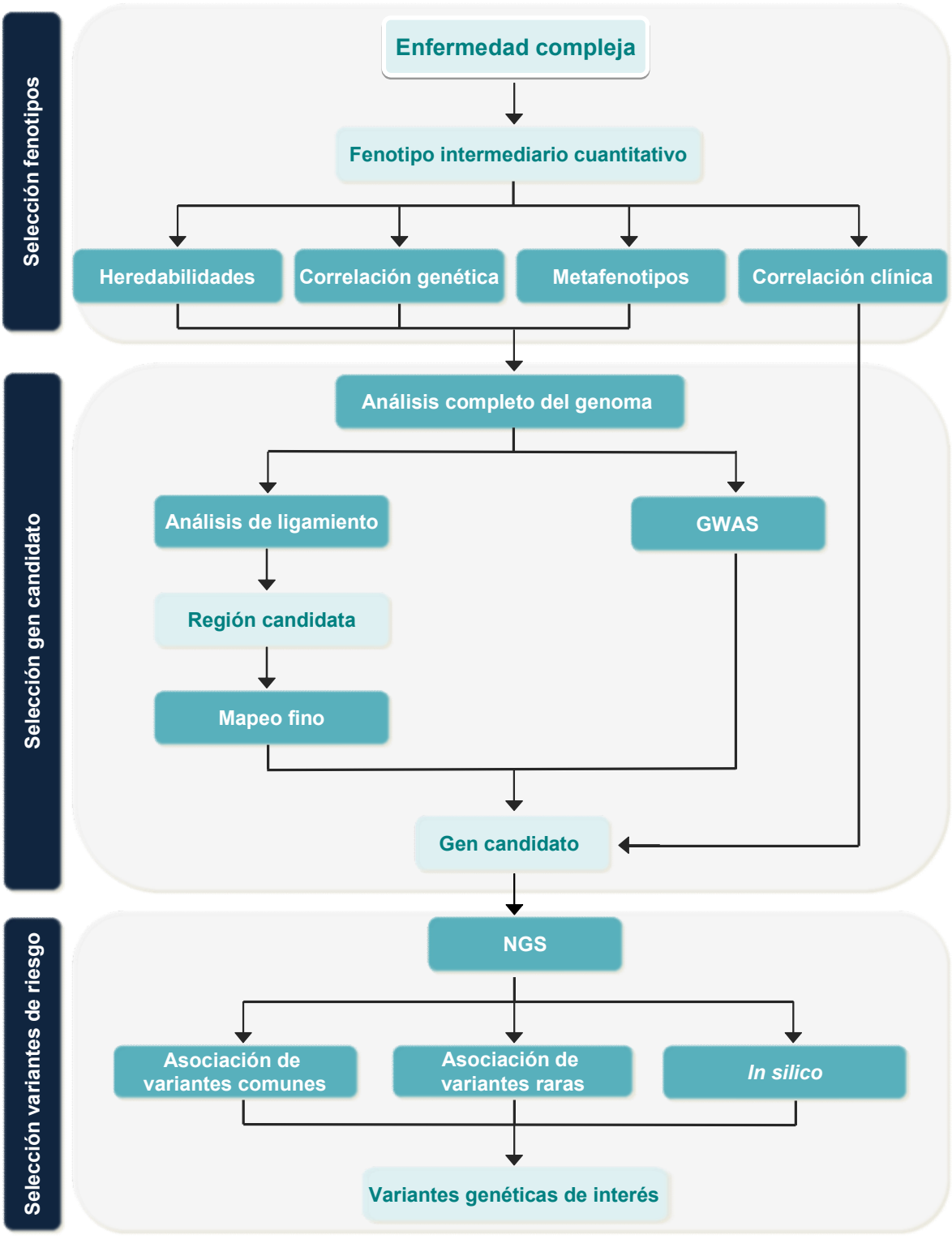

Figura 9. Esquema de la estrategia aplicada para el estudio de la enfermedad compleja VTE.

El primer paso en la aplicación de la estrategia diseñada para la identificación y la localización de determinantes genéticos implicados en el riesgo de padecer VTE debe consistir en establecer que el fenotipo de estudio, como por ejemplo el propio fenotipo enfermedad, es heredable. Es importante recordar que el concepto enfermedad se entiende en este contexto como una variable continua, por lo que los individuos tienen una determinada predisposición a padecer la VTE (Figura 4). La alta heredabilidad del riesgo de VTE comprobada en el Proyecto GAIT-1 (0,61±0,16) justifica la búsqueda de genes subyacentes que todavía no se hayan identificado o cuyo mecanismo de participación no se conozca al completo. Asimismo, hemos estimado en la presente Tesis Doctoral una heredabilidad del riesgo de VTE del 0,67±0,17 en el contexto del Proyecto GAIT-2 (Artículo 1).

Por otra parte, el análisis de fenotipos cuantitativos intermediarios asociados al estudio de una enfermedad compleja ofrece unas condiciones óptimas para el mapeo genético. La detección de genes o variantes genéticas relacionadas con la variabilidad de los fenotipos intermediarios pueden estar también relacionadas con el riesgo de VTE (Souto et al. 2000a). Para alcanzar los objetivos propuestos hemos estudiado distintos fenotipos cuantitativos intermediarios relativos a los componentes de la hemostasia, pertenecientes al sistema de la coagulación o al mecanismo de la fibrinólisis. Cabe destacar que, de la hemostasia, los componentes de la coagulación plasmática son los determinantes de riesgo de VTE mejor estudiados, tanto los factores procoagulantes como los inhibidores. En particular, resultados previos de nuestro grupo apuntan hacia la vía intrínseca de la cascada de la coagulación como una de las vías más importantes implicadas en la tendencia a la trombosis (ver apartado 4.1.1). En concreto, hemos seleccionado 3 fenotipos cuantitativos intermediarios individuales propios de la hemostasia. Éstos son niveles de FVIII y de FXI, ambos componentes de la vía intrínseca de la coagulación, y niveles de plasminógeno, proteína que está

implicada en el proceso fibrinolítico. Además, hemos seleccionado el TGT como fenotipo intermediario global y los “metafenotipos” como fenotipos integradores.

La selección de los niveles de FVIII y de FXI como fenotipos de interés se ha basado en las altas heredabilidades estimadas en el Proyecto GAIT-1, siendo aproximadamente de un 40% para los niveles de FVIII y de un 45% para los niveles de FXI (Tabla 5). Igualmente, hemos replicado en el contexto del Proyecto GAIT-2 la heredabilidad de los niveles de FVIII (38%) como parte de los resultados del Artículo 2, así como la heredabilidad de los niveles de FXI (54%) (resultado no publicado). Por otra parte, se ha descrito anteriormente una correlación genética estadísticamente significativa y positiva de los niveles de FVIII y de los niveles de FXI con la VTE (Tabla 6). Por lo tanto, se sugiere que factores genéticos que determinan la variabilidad de estos fenotipos intermediarios pueden estar implicados en el riesgo de padecer VTE.

Asimismo, se ha calculado en el Proyecto GAIT-1 la heredabilidad de los niveles de plasminógeno en un 24% (Tabla 5), aunque no se ha identificado una correlación genética con el riesgo de VTE (Souto et al. 2000a). Debemos tener en cuenta que en las recomendaciones actuales se consensúa que esta proteína no debe considerarse un factor de riesgo de trombosis (Brandt 2002). A pesar de ello, hemos seleccionado los niveles de plasminógeno funcional como fenotipo cuantitativo intermediario de interés basándonos en observaciones clínicas (Artículo 5). Además, las evidencias biológicas muestran el importante papel que juega esta proteína fibrinolítica en la degradación de las hebras de fibrina.

Por otra parte, hemos seleccionado el TGT como fenotipo intermediario global. El TGT es un test automático que mide la cantidad de trombina generada a lo largo del tiempo,

representándolo en una curva llamada trombograma. Por lo tanto, se encuentran representadas en este test las vías intrínseca, extrínseca y común.

Sin embargo, los test clínicos tradicionales son el tiempo de protrombina (PT) y el tiempo de tromboplastina parcial activada (aPTT). El PT examina a partir de la activación del FVII el tiempo que tarda la sangre en coagularse, por lo que evalúa el estado global de la vía extrínseca y de la vía común. Por otra parte, el aPTT determina el tiempo transcurrido desde la activación del FXII hasta la formación del coágulo de fibrina, evaluando la vía intrínseca y la vía común. Cabe destacar que los fenotipos globales como los test de coagulación no son fenotipos intermediarios tan próximos a la acción de los genes como los fenotipos individuales, pero suponen una mejor representación de los procesos procoagulantes o anticoagulantes sin llegar a la complejidad del fenotipo enfermedad. De hecho, el uso de PT y de aPTT como fenotipos intermediarios ha contribuido al conocimiento de los mecanismos implicados en el riesgo de VTE. Por ejemplo, identificando los genes *F7* y *PROCR* como los mayores determinantes genéticos de la variabilidad de PT. De igual modo, los genes *KNG1*, *HRG*, *F11*, *F12*, y *ABO* se han asociado al test aPTT (Morange and Oudot-Mellakh 2011; Tang et al. 2012; Sabater-Lleal et al. 2012).

En comparación, el TGT permite incluso una mayor representación del sistema de la coagulación, si tenemos en cuenta que ensayos como PT y aPTT finalizan cuando se ha generado alrededor de un 5% de la trombina (van Veen et al. 2008). En el Artículo 1 se han utilizado los individuos del Proyecto GAIT-2 para estudiar tres parámetros cuantitativos que se derivan del trombograma como fenotipos intermediarios descriptivos del TGT: el tiempo de latencia, la altura del pico de trombina y el potencial ETP (Duchemin et al. 2008). A partir de estos resultados se han estimado las heredabilidades del tiempo de latencia (49%), altura del pico de trombina (54%) y ETP (52%), aunque únicamente se ha obtenido una correlación genética estadísticamente significativa y positiva con el riesgo de VTE en el caso del pico de trombina y del ETP.

Por lo tanto, se justifica la búsqueda de genes candidatos implicados en la variabilidad de los fenotipos derivados del TGT.

Por último, a pesar de la importancia de los fenotipos intermediarios en el estudio de la base genética de las enfermedades complejas, debemos tener en cuenta que los análisis univariados explican solamente una parte de la heredabilidad estimada. Como metodología complementaria, hemos estudiado en el contexto del Proyecto GAIT-1 los fenotipos integradores denominados “metafenotipos” (Artículo 3). Éstos se han obtenido a partir de la combinación de 27 parámetros de la coagulación sanguínea y de la fibrinólisis, los cuales se han seleccionado por participar en el mismo mecanismo biológico. La principal ventaja frente al análisis de otros fenotipos globales como el TGT, que siguen representando modelos univariados, es que no sólo contemplan la participación de los distintos fenotipos intermediarios, sino que consideran la implicación de cada uno de ellos.

En concreto, hemos aplicado el modelo matemático ICA, que se basa en la premisa de que la variabilidad del conjunto de datos observados puede deberse a distintas causas independientes. Por lo tanto, cada “metafenotipo” corresponde a una combinación de los fenotipos originales, los cuales aportan distintos pesos a la construcción de manera que la variabilidad representada sea independiente a la de otros “metafenotipos”. A partir del análisis genético de un “metafenotipo” se podría detectar un determinante genético implicado simultáneamente en la variabilidad de distintos fenotipos intermediarios que no haya podido ser identificado mediante análisis univariados. En nuestro caso, los “metafenotipos” obtenidos presentan unas heredabilidades estimadas entre el 15 y el 70%.

Una vez seleccionados los fenotipos intermediarios de interés y comprobada la existencia de una base genética subyacente, la estrategia utilizada para la identificación de genes candidatos implicados en la variabilidad de estos fenotipos incluye los estudios de genoma completo mediante análisis de ligamiento genético y los análisis de asociación. Éstas son dos metodologías complementarias que ofrecen resultados de distinto tipo a pesar de basarse en el mismo principio de herencia ligada de variantes genéticas cercanas (Tabla 8).

Tabla 8. Resumen de las características básicas de los estudios de ligamiento y de asociación. Adaptado (Hirschhorn 2005; Ott et al. 2011).

Propiedades	Estudios de ligamiento	Estudios de asociación
Individuos	Familiares	No relacionados o familiares
Intervalo de detección del efecto	Amplio	Estrecho
Número de marcadores en genoma	Moderado	Grande
Detección de variantes comunes de efecto relativamente bajo	-/+	+
Detección de variantes raras	+	-/+

El estudio de ligamiento ha tenido como objetivo identificar QTLs implicados en la variabilidad del fenotipo intermediario de interés. En este caso, hemos utilizado como marcadores genéticos los microsatélites. Debido a que éstos pueden tener un gran número de alelos, los microsatélites son más idóneos para la detección de alelos de baja frecuencia alélica en comparación con otros marcadores como los SNPs (Xiong and Jin 1999). Sin embargo, debemos tener en cuenta que el poder estadístico para detectar una señal de ligamiento es dependiente de la heredabilidad o proporción de la variabilidad del fenotipo de estudio explicada por el QTL. Por lo tanto, estos estudios poseen un bajo poder estadístico para detectar variantes raras con efectos individuales moderados o bajos y un mayor poder estadístico para detectar variantes

raras con un gran efecto en el fenotipo (Williams and Blangero 1999; Zeggini and Morris 2010).

En nuestro estudio, las regiones marcadas por los QTLs sobrepasan los 7 millones de bases, por lo que es necesario realizar análisis adicionales para tratar de acotar el gen o genes potencialmente funcionales. La secuenciación podría considerarse como una metodología idónea para esto, puesto que permite obtener información sobre todo el espectro de variabilidad de la región de interés. En cambio, no siempre ha sido la metodología elegida debido al elevado coste y al tiempo requerido para estos estudios, especialmente antes de la aparición de la NGS. Una posible alternativa consiste en la limitación de la secuenciación a las zonas exónicas; no obstante, se ha descrito que las enfermedades complejas están a menudo influenciadas por variantes genéticas situadas en regiones no codificantes con funciones reguladoras y efectos moderados (Hirschhorn 2005; Manolio et al. 2009). Como metodología más simple y barata que la secuenciación de regiones candidatas hemos realizado un mapeo fino del QTL a partir del genotipado de SNPs y acompañado por un análisis de asociación. Los SNPs ofrecen numerosas ventajas frente al uso de microsatélites a pesar de su condición típicamente bialélica. Esto es debido a su gran abundancia, bajo coste, bajo ratio de mutación y facilidad de uso (Burmeister 1999; Xiong and Jin 1999).

Por otro lado, para la identificación de genes candidatos, hemos realizado un estudio de asociación mediante GWAS de fenotipos cuantitativos a partir de SNPs, los cuales pueden estar implicados en la variabilidad fenotípica o en LD con las variantes de interés. En estos análisis de asociación la penetrancia y la frecuencia alélica juegan un papel importante, siendo las variantes comunes o con mayor penetrancia genética las que aportan un mayor poder estadístico. En comparación, los estudios de asociación son más eficientes que los estudios de ligamiento en la identificación de variantes genéticas con un efecto bajo (Risch and Merikangas 1996; Hirschhorn 2005). Por lo

tanto, algunas de las limitaciones de los estudios de ligamiento se pueden superar con las técnicas de GWAS. Además, es importante tener en cuenta que el tipo de muestra requerido para un análisis de ligamiento (familias) también puede utilizarse en los análisis de asociación.

Tanto los estudios de ligamiento como los GWAS se han utilizado anteriormente en el análisis genético de la VTE (Martinelli et al. 2014; Vilalta and Souto 2014). A pesar de ello, se observa una "heredabilidad perdida" que sigue latente (Cunha et al. 2015). Esto sugiere que las enfermedades complejas como la VTE no pueden ser explicadas solamente por una serie de variantes raras con efectos individuales altos en la enfermedad, ni por variantes comunes de efecto moderado o bajo. Los resultados más recientes indican que el riesgo de padecer una enfermedad compleja parece estar influida por un conjunto de variantes genéticas con frecuencias alélicas distintas y de rangos de efecto diferentes (Figura 10). Además, también podrían estar implicados otros factores como variantes estructurales e interacciones gen-gen o gen-ambiente (Manolio et al. 2009; Zeggini and Morris 2010).

De especial interés para el estudio de la "heredabilidad perdida" son las variantes raras y de baja frecuencia. Éstas presentan una frecuencia alélica que no es lo suficientemente alta como para estar representadas en los GWAS. Además, pueden no tener un efecto suficientemente grande en la variabilidad de los niveles del fenotipo de estudio como para ser detectadas mediante análisis de ligamiento (Manolio et al. 2009). Una de las estrategias que hemos aplicado en el Proyecto GAIT-2 para afrontar estas limitaciones consiste en adaptar los métodos de GWAS mediante la imputación genotípica.

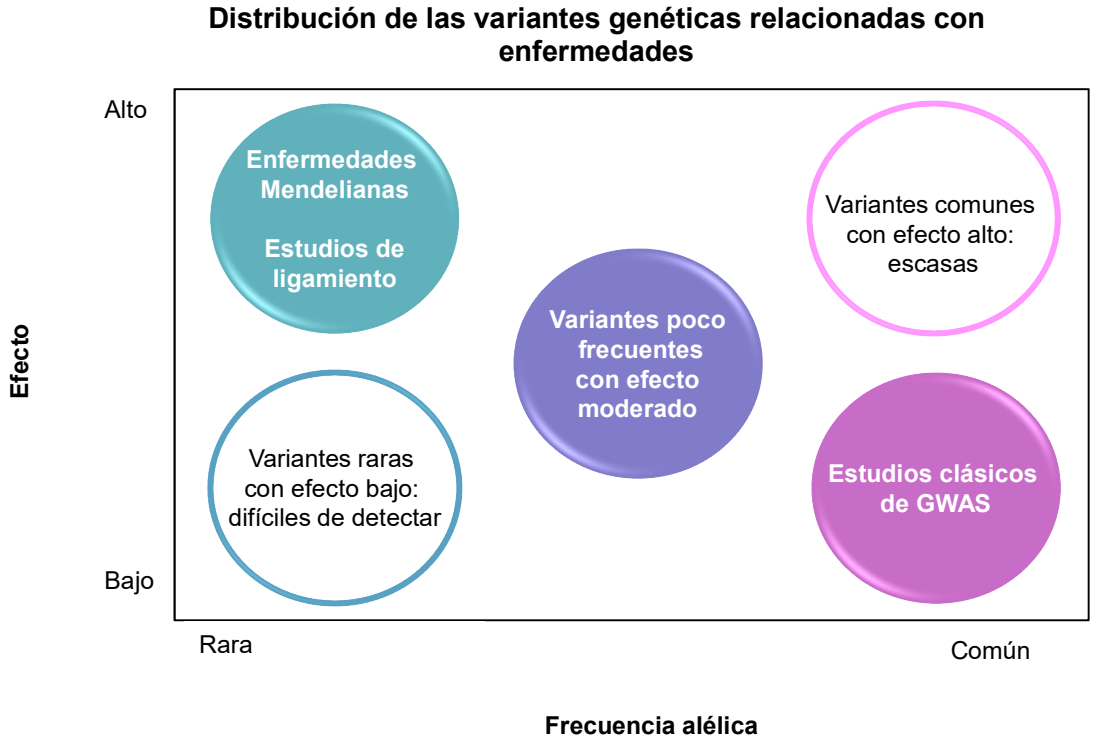

Figura 10. Distribución de las variantes genéticas relacionadas con las enfermedades complejas en función de sus efectos y frecuencias alélicas. Las variantes raras con grandes efectos en el riesgo de padecer una enfermedad son típicamente detectadas en los estudios de ligamiento, mientras que las variantes genéticas comunes con efectos relativamente bajos son detectadas con mucha mayor facilidad en los GWAS. Adaptado (Manolio et al. 2009).

La imputación consigue llevar a cabo el estudio de variantes raras o de baja frecuencia alélica a lo largo del genoma de manera más económica que mediante la técnica de secuenciación. Sin embargo, la aparición de la NGS ofrece mejoras tanto a nivel económico como de capacidad y velocidad de secuenciación (van Nimwegen et al. 2016) y además se solventa los problemas ligados a la predicción de genotipos de variantes de baja frecuencia (Tabla 4).

Una vez identificados los genes candidatos, hemos desarrollado en el marco del Proyecto GAIT-2 una metodología basada en NGS para la identificación de las variantes genéticas de riesgo. En concreto, hemos analizado individuos con fenotipos

extremos, más propensos a ser portadores de alelos causales (Manolio et al. 2009), para optimizar la detección de variantes raras y de variantes de baja frecuencia asociadas a la variabilidad del fenotipo de estudio. Las plataformas de NGS disponibles muestran distintas características que hemos valorado en función de los objetivos de estudio, puesto que su aplicación tendrá importantes repercusiones tanto en el diseño de la estrategia como en el tipo de resultados obtenidos. En nuestro caso, se ha seleccionado la tecnología de Illumina por su buena relación capacidad-coste y debido a que soslaya los problemas de secuenciación de homopolímeros (Park et al. 2015). La mayoría de variantes asociadas a enfermedades complejas en los estudios de GWAS se encuentran en regiones no codificantes (Manolio et al. 2009), por lo que en la secuenciación es especialmente importante la inclusión no sólo de regiones exónicas, sino también de intrones, regiones promotoras o UTRs que pueden tener una función reguladora. Debido a esto, en las regiones candidatas que hemos seleccionado puede ser común la presencia de zonas repetitivas o homopolímeros. A pesar de las ventajas que supone la NGS en comparación con la secuenciación tradicional, no hemos podido abordar la secuenciación de todos los individuos del Proyecto GAIT-2. Por este motivo, hemos optado por capturar la mayor variabilidad genética posible analizando individuos no relacionados entre sí de distintas familias o núcleos familiares. En el futuro, se podrán secuenciar un mayor número de individuos de forma rutinaria gracias a la continua disminución de los costes de la NGS (Figura 5).

A partir de los resultados de la NGS, hemos realizado tanto análisis de asociación entre el fenotipo de estudio y variantes comunes particulares, como análisis de asociación a partir de un método de agregación por regiones para las variantes raras y de baja frecuencia alélica. Este método permite contemplar la posibilidad de que las variantes agrupadas proporcionen efectos bidireccionales y la inclusión de variantes sin efecto (Wu et al. 2011). Además, hemos priorizado la caracterización de algunas

variantes raras, que han sido seleccionadas teniendo en cuenta su localización y posible segregación familiar con el fenotipo de estudio. Las predicciones *in silico* sobre la funcionalidad de estas variantes nos han ofrecido nuevos datos que puede dar soporte a su determinación como potenciales variantes patogénicas (Wallis et al. 2013).

En este trabajo hemos utilizado la combinación de las estrategias metodológicas discutidas hasta ahora para tratar de solventar distintas limitaciones en la identificación de los determinantes genéticos de la VTE. De esta manera, hemos analizado el genoma completo (mediante estudios de ligamiento y de GWAS) y hemos estudiado genes candidatos (con una metodología basada en NGS) para identificar factores de riesgo genéticos implicados en la variabilidad de fenotipos intermediarios relacionados con el riesgo de VTE.

En el estudio de los niveles de FVIII como fenotipo intermediario hemos realizado un estudio de ligamiento en las familias extensas del Proyecto GAIT-1, seguido por un mapeo fino y secuenciación mediante NGS de la región candidata en individuos seleccionados del Proyecto GAIT-2 (Artículo 2). En particular, a partir del análisis de la región situada en el cromosoma 2 que muestra evidencias de ligamiento, *CIB4* se ha postulado como un nuevo gen candidato relacionado con la variabilidad de los niveles de FVIII. Por lo tanto, este gen podría estar también implicado en el riesgo de VTE. En base a estos resultados y tras la secuenciación de *CIB4*, hemos obtenido un listado de variantes genéticas comunes, de baja frecuencia alélica y raras asociadas estadísticamente con los niveles de FVIII. Todas estas variantes genéticas se encuentran localizadas en intrones, por lo que nuestro diseño de secuenciación ha sido crucial para la identificación de dichas asociaciones. Estas variantes serán priorizadas en los futuros estudios de replicación. Para ello, es de gran interés el uso

de familias extensas, como las incluidas en los Proyectos GAIT-1 y GAIT-2, ya que puede facilitar el análisis de las variantes raras y de baja frecuencia al aumentar su frecuencia entre individuos relacionados y afectos.

Por el contrario, el resultado de ligamiento identificado en el cromosoma 3 no ha sido replicado mediante el mapeo fino. Esto puede deberse al tipo de marcador genético utilizado. Teniendo en cuenta que se han genotipado distintos SNPs para los análisis de asociación, en el caso de que la variante de riesgo asociada a la enfermedad sea de baja frecuencia o rara, ésta puede no estar representada en el mapeo fino. Por lo tanto, esta región es igualmente susceptible de ser analizada en futuros estudios.

Siguiendo una aproximación alternativa, en la determinación de los componentes genéticos implicados en la variabilidad de los niveles de FXI (Artículo 4) partimos de evidencias publicadas anteriormente en relación con el GWAS del Proyecto GAIT-1. A partir de este GWAS, los genes candidatos implicados en la variabilidad de los niveles de FXI fueron *KNG1* y *F11*. También se determinó que el FXI es el mecanismo principal mediante el cual estos dos genes parecen influir tanto en la variabilidad del test aPTT como en la susceptibilidad a padecer VTE. Estos resultados se replicaron en 662 individuos suecos no relacionados (Sabater-Lleal et al. 2012).

En el Artículo 4 hemos analizado mediante NGS la variabilidad de secuencia de *KNG1* y *F11* en individuos seleccionados del Proyecto GAIT-2. Por lo tanto, hemos integrado los datos obtenidos previamente mediante GWAS y los resultados obtenidos mediante estudios de asociación de datos de secuenciación. De esta manera, hemos identificado distintas variantes genéticas a lo largo de ambos genes asociadas de forma significativa con los niveles de FXI a partir del análisis tanto de variantes comunes como de agrupaciones de variantes de baja frecuencia alélica y raras. En este caso, la mayoría de las asociaciones identificadas tampoco se habrían obtenido a partir de la secuenciación de las regiones codificantes. Cabe destacar que la mayoría

de estas variantes genéticas no se han descrito previamente en asociación con este fenotipo intermediario. Sin embargo, la variante genética común que ha presentado en este estudio mayor significación estadística en relación con niveles elevados de FXI (rs710446) coincide con la variante identificada en el Proyecto GAIT-1 (Sabater-Lleal et al. 2012). Asimismo, son destacables dos variantes genéticas raras como potenciales variantes patogénicas que serán evaluadas y confirmadas en futuros estudios funcionales.

La primera variante es NM_000128.3: c.943G>A (p.Glu315Lys), está localizada en *F11* y se ha descrito previamente en relación con niveles bajos de FXI (Quélin et al. 2006). De la misma manera, en nuestro trabajo se ha detectado en un individuo con niveles de FXI del 36%, sugestivo de que podría provocar cambios en la funcionalidad de la proteína. Sin embargo, las predicciones del efecto de esta variante muestran resultados discordantes. Esto se debe a las limitaciones de las pruebas *in silico* en comparación con los ensayos funcionales existentes, aunque éstos resultan más caros y conllevan más tiempo de trabajo. Es por este motivo que las predicciones *in silico* deben considerarse como un apoyo a otras evidencias experimentales, pero no como confirmación del efecto funcional de la variante analizada. En concreto, esta variante genética podría afectar a la correcta funcionalidad de la proteína por cambios en la interacción con otras moléculas o con otras partes de la proteína.

La segunda variante seleccionada como potencialmente patogénica (NM_001102416.2: c.758-12T>C) está situada en el intrón 6 del *KNG1* y es adyacente a la región de *splicing*. En concreto, se ha detectado en un individuo que presenta niveles elevados de FXI (188%). Esta variante, no descrita previamente en relación a los niveles de FXI, podría afectar a la transcripción de *KNG1* y ser la responsable principal de este fenotipo, teniendo en cuenta que se sitúa a 12 pares de bases (bp) del corte en el sitio aceptor. Las predicciones *in silico* muestran una disminución de la

puntuación obtenida en el sitio aceptor de *splicing* respecto a la secuencia de referencia.

En conjunto, se destaca la implicación de *KNG1* en la regulación de los niveles del FXI de la coagulación que, además, pueden influenciar en el riesgo de VTE. Recientemente, se han puesto de manifiesto nuevas evidencias de su papel en la vía intrínseca de la coagulación a partir de un meta-análisis Europeo (Sennblad et al. 2017).

En el Artículo 5 se han estudiado 4 núcleos familiares con déficit de plasminógeno que fueron reclutados a partir de un individuo afecto con trombofilia espontánea. Estos individuos son independientes a los incluidos en el Proyecto GAIT-1 y GAIT-2. Por este motivo, hemos evaluado la base genética conocida de riesgo de trombosis, de la cual no teníamos información previa, mediante la genotipación de 12 variantes genéticas de riesgo de trombosis incluidas en la herramienta Thrombo inCode (Soria et al. 2014). Tras esta exploración, hemos comprobado que sólo 1 de las 4 familias es portadora de un perfil de riesgo, por lo que las causas de los eventos trombóticos sufridos en las restantes 3 familias siguen siendo desconocidas. Por lo tanto, hemos considerado que en determinadas familias, como las estudiadas en el presente trabajo, variantes genéticas en el gen estructural del plasminógeno puedan estar implicadas en la tendencia protrombótica.

A partir de la metodología de NGS desarrollada, hemos evaluado por primera vez la aplicabilidad de esta tecnología en la caracterización de mutaciones en *PLG* potencialmente patogénicas y responsables de un déficit de plasminógeno. Tras la identificación de la variabilidad genética de *PLG* presente en los individuos de estudio, sin discriminación por frecuencia alélica, hemos seleccionado 5 variantes genéticas potencialmente patogénicas en el *PLG* (NM_000301.3) que se evaluaron mediante

herramientas bioinformáticas de predicción *in silico*. Entre éstas se incluyen las 3 variantes genéticas de cambio de sentido c.112A>G (p.Lys38Glu), c.781C>T (p.Arg261Cys) y c.2134G>A (p.Gly712Arg) y dos variantes potencialmente patogénicas cercanas a la región de *splicing* c.12G>A (p.Lys4Lys) y c.1878-6T>C. En este caso, los resultados concordantes entre las predicciones *in silico* y el fenotipo observado para las 3 variantes de cambio de sentido, avalarían un efecto deletéreo sobre la función del plasminógeno. En concreto, destacamos la variante c.781C>T (p.Arg261Cys) al relacionarse por primera vez con la deficiencia de plasminógeno. Esta variante genética podría afectar a la interacción de la proteína con otras moléculas. En relación a las variantes genéticas c.12G>A (p.Lys4Lys) y c.1878-6T>C con posible afectación en el proceso de *splicing* no se predijeron cambios en la actividad del sitio donador o del sitio aceptor, respectivamente. Sin embargo, otros mecanismos reguladores podrían estar alterados.

A pesar de que los resultados obtenidos son de gran utilidad en la caracterización genética del déficit de plasminógeno, no hemos podido confirmar la determinación de estas variantes genéticas como factores de riesgo de trombosis debido a la falta de cosegregación con la VTE. A pesar de esto, no podemos refutar la implicación de estas mutaciones potencialmente patogénicas e incluso la de otras variantes genéticas situadas en el gen *PLG* como factores de riesgo implicados en la susceptibilidad a VTE, teniendo en cuenta la “heredabilidad perdida” de esta patología. Es decir, cabe la posibilidad de la existencia de una interacción de estas variantes con otros factores de riesgo aún desconocidos, por lo que su papel en el riesgo de VTE no pueda esclarecerse todavía. Incluso, existe la posibilidad de que los individuos reclutados, asintomáticos a día de hoy, sufran eventos trombóticos futuros que no han podido ser contemplados en este estudio.

De forma complementaria a estos estudios, hemos realizado un GWAS en el Proyecto GAIT-2, optimizado mediante imputación, para el análisis genético de los parámetros derivados del fenotipo global TGT (Artículo 1). A partir de esta metodología hemos identificado al gen *F2* como el mayor responsable de la variabilidad de los niveles de TGT. De hecho, el *F2* es un factor de riesgo establecido de la VTE y la asociación de TGT con la variante genética G20210A (rs1799963) es conocida (Kyrle et al. 1998; Segers et al. 2010; Rocanin-Arjo et al. 2014). Incluso, se han obtenido otras asociaciones estadísticamente significativas o sugestivas con TGT, aunque los genes implicados no muestran evidencias biológicas de la participación con dicho fenotipo intermediario, por lo que resulta un reto poder establecer una relación causal. Igualmente, debemos recordar que los genotipos resultantes de la metodología de imputación se basan en predicciones, por lo que cualquier señal estadísticamente significativa derivada de estudios de imputación debe ser replicada mediante métodos experimentales de observación. La estrategia de NGS desarrollada es una metodología atractiva y factible para el futuro análisis del *F2* con el fin de establecer el catálogo de variantes genéticas localizadas en esta región implicadas en la variabilidad del TGT.

Finalmente, hemos evaluado el análisis genético de vías completas, en nuestro caso el sistema de la coagulación y la fibrinólisis, generando el concepto “metafenotipo” (Artículo 3). En concreto, hemos comparado los resultados derivados del GWAS del GAIT-1 con los respectivos análisis univariados para poder interpretar y caracterizar los resultados obtenidos a partir de nuestra metodología. De todos los análisis realizados, destacan los resultados derivados de los “metafenotipos” ICA-C10 e ICA-C3. Éstos se han obtenido a partir de distintas combinaciones de los fenotipos seleccionados, los cuales están implicados en el mecanismo de la VTE. Las

heredabilidades estimadas en ambos casos son altas, siendo del 70% para ICA-C10 y del 53% en el caso de ICA-C3.

Respecto a ICA-C10, se observa una señal estadísticamente significativa correspondiente a la variante rs2731672 en el gen *F12*. Hemos observado que la variabilidad representada en este "metafenotipo" corresponde básicamente a la variabilidad de los niveles del FXII. Por lo tanto, concluimos que los análisis de asociación de "metafenotipos" basados en ICA también pueden identificar determinantes genéticos implicados en la variabilidad de un único factor de la coagulación. De hecho, esto resultados coinciden con los obtenidos en el análisis univariado del FXII. Anteriormente, se ha estimado en el GAIT-1 que los niveles de este factor de la coagulación presenta una heredabilidad del 67% (Tabla 5) y una correlación genética positiva con la VTE (Tabla 6). Además, se ha reportado un QTL en la región del *F12* relacionado con los niveles de FXII y de VTE (Soria et al. 2002) y se ha estudiado a partir de datos de secuenciación la variabilidad genética de este gen en relación con los niveles del FXII (Calafell et al. 2010). Por lo tanto, nuestros resultados son concordantes con los ya obtenidos para estas familias mediante otras metodologías, dando validez al método empleado en este estudio.

En relación a ICA-C3 se ha identificado una región que comprende a los genes *HRG* (rs9898), *FETUB* (rs3733159) y *KNG1* (rs1621816 y rs1403694). En este caso, ICA-C3 no representa de forma mayoritaria a un solo fenotipo, sino que representa al conjunto de parámetros relacionados con la coagulación sanguínea y la fibrinólisis. Por lo tanto, constatamos que los análisis de asociación de "metafenotipos" basados en ICA pueden detectar el efecto de genes implicados en la variabilidad de distintos parámetros de la coagulación sanguínea. En comparación, el análisis univariado muestra una asociación significativa entre estos mismos genes y la glicoproteína rica en histidina (HRG) en particular. Se ha estimado anteriormente en el Proyecto GAIT-1 una heredabilidad de HRG del 52% (Tabla 5) aunque no se ha detectado una

correlación genética con el riesgo de VTE (Souto et al. 2000a). No obstante, se ha reportado el papel de HRG en la regulación tanto de la coagulación como de la fibrinólisis (Jones et al. 2005) e incluso se ha estudiado su participación concreta en la vía intrínseca de la coagulación (MacQuarrie et al. 2011). Entre los genes identificados en ICA-C3, los cuales están situados en la misma región genómica, cabe destacar la participación de las proteínas codificadas por *HRG* (HRG) y *KNG1* (HMWK) en el mecanismo estudiado. En particular, *HRG* ya se había relacionado anteriormente con la variabilidad de los niveles de aPTT y, por lo tanto, con el riesgo de VTE (Houlihan et al. 2010; Morange and Oudot-Mellakh 2011). Asimismo, se ha demostrado anteriormente en el contexto del Proyecto GAIT-1 que *KNG1* influye en aPTT a través de la variabilidad de los niveles de FXI, así como se ha determinado su participación en la variabilidad de otros parámetros como HRG y FXII (Sabater-Lleal et al. 2012). Los resultados derivados en nuestro estudio a partir de la integración de distintos fenotipos implicados en el mismo proceso biológico confirman que esta región genómica participa en la variación de los niveles de distintos fenotipos de la cascada de la coagulación y la fibrinólisis. En concreto, *KNG1* se postula como uno de los principales determinantes genéticos implicados en la variabilidad del conjunto de fenotipos de la vía intrínseca. No obstante, se observa que el mecanismo implicado no incluye especialmente la participación de los niveles de FXI o FXII, teniendo en cuenta el bajo peso de estos fenotipos en ICA-C3. Así pues, el estudio detallado de la variabilidad de *KNG1* en relación a los distintos parámetros de la vía intrínseca puede contribuir al descubrimiento de nuevos mecanismos implicados en la regulación de la cascada de la coagulación que esclarezcan parte de la “heredabilidad perdida” del riesgo de VTE.

En nuestro estudio hemos explorado mediante “metafenotipos” la base genética del sistema de la coagulación sanguínea y la fibrinólisis, proceso en el que algunos de los mecanismos de riesgo ya están establecidos. Esto nos ha permitido demostrar la

interpretación biológica de esta metodología. Por lo tanto, esta metodología podría ser también de gran utilidad en la exploración de otras vías o patologías menos conocidas.

Los resultados de nuestro trabajo, tomados conjuntamente, avalarían la hipótesis de que la VTE estaría determinada por parámetros o fenotipos que, a su vez, estarían influidos por el conjunto de variantes genéticas de distintas frecuencias alélicas y efectos. Se han descrito múltiples ejemplos que corroboran nuestras observaciones para enfermedades complejas o fenotipos cuantitativos como la diabetes tipo 2 (Bonnefond and Froguel 2015), la enfermedad de Parkinson (Tsuji 2010), el índice de masa corporal (Harismendy et al. 2010) o el cáncer (Stratton and Rahman 2008), entre otros.

La presente Tesis Doctoral aglutina los resultados derivados de un nuevo enfoque aplicado al estudio genético de la VTE en un tipo de muestra muy poco utilizado, por su dificultad de reclutamiento, como es el uso de familias extensas. Este nuevo enfoque combina métodos de análisis de datos genéticos como los estudios de ligamiento genético de rasgos cuantitativos, los modelos estadísticos de asociación masiva o GWAS, la técnica de imputación de datos genéticos, la secuenciación mediante NGS, los métodos bioinformáticos derivados para el manejo de los datos masivos generados y las herramientas de predicción *in silico*. Además, se han empleado nuevos métodos de diagnóstico molecular como la herramienta Thrombo inCode, aplicable en la actividad clínica diaria, y se ha incorporado el estudio de fenotipos intermediarios globales (como TGT) o nuevos fenotipos integradores o "metafenotipos". Asimismo, la estrategia seguida no sólo permite la exploración de la base genética de la VTE, sino que es aplicable a cualquier enfermedad compleja con el fin de contribuir al estudio y caracterización de la "heredabilidad perdida".

En concreto, hemos analizado la base genética relacionada con la variabilidad de fenotipos cuantitativos intermediarios evaluados en familias. Los estudios a lo largo del

genoma sin hipótesis previa han permitido la detección de genes candidatos que podrían estar también implicados en el riesgo de VTE. Además, la tecnología de NGS ha permitido el análisis de toda la variabilidad genética de genes implicados en la variación de los niveles de parámetros de la vía intrínseca de la coagulación, considerando todo el catálogo de frecuencia alélica y teniendo en cuenta la heterogeneidad genética de esta enfermedad. De esta manera, hemos podido definir las regiones génicas de mayor interés para futuros estudios. Asimismo, hemos aplicado la metodología de NGS desarrollada en la caracterización genética de patologías relacionadas con el sistema de la coagulación sanguínea.

Los esfuerzos aplicados para la investigación de los determinantes genéticos de riesgo trombótico pueden tener una marcada implicación en la salud de la población por su transferibilidad diagnóstica y terapéutica. Asimismo, los resultados obtenidos pueden sugerir estrategias preventivas mediante la aplicación de la medicina personalizada que podrían reducir la mortalidad y morbilidad de la VTE. Esta enfermedad supone un coste muy alto tanto a nivel económico como social, por lo que cualquier avance en el diagnóstico, la prevención y el tratamiento de esta patología tiene probablemente una repercusión coste-eficacia muy ventajosa, lo que justifica las investigaciones en este campo.

Conclusiones

Conclusiones

- El estudio de las familias extensas incluidas en el Proyecto GAIT-2 ha permitido confirmar la estimación de la heredabilidad del riesgo de padecer VTE en un 67%, siendo éste un paso fundamental y previo a la búsqueda de genes candidatos. Esta alta heredabilidad indica que la mayor parte de la variabilidad fenotípica se atribuye a la base genética, por lo que se justifica el estudio de los determinantes genéticos implicados en el riesgo de padecer VTE.

- La estimación de la heredabilidad de los parámetros derivados del TGT y las correlaciones genéticas en relación al riesgo de padecer VTE determina que estos fenotipos intermediarios son de gran utilidad para el estudio de la "heredabilidad perdida" de la trombosis, siendo el pico de trombina y el ETP los más idóneos.

- El estudio de las familias extensas de los proyectos GAIT-1 y GAIT-2 nos ha permitido localizar genes candidatos implicados en la variabilidad de distintos fenotipos intermediarios cuantitativos y con posibles efectos en el riesgo de padecer VTE.

    - Los genes identificados tras el análisis de ligamiento o GWAS de fenotipos intermediarios individuales han sido:

        - *CIB4*, implicado en la variabilidad de los niveles de FVIII.

- Los genes detectados a partir del análisis de GWAS de fenotipos que representan o integran la variabilidad del conjunto fenotipos implicados en la cascada de la coagulación y en la fibrinólisis han sido:

    - *F2*, como mayor determinante genético del TGT.
    - *KNG1*, con participación en la base genética de la vía intrínseca de la coagulación mediante mecanismos alternativos a la regulación de los niveles de FXI y FXII.

- Se ha establecido la variabilidad de la secuencia de los genes candidatos *CIB4*, *F11* y *KNG1* tras el uso de la técnica de NGS. Mediante la optimización de los análisis de asociación se ha identificado un conjunto de variantes genéticas comunes, variantes de baja frecuencia alélica y variantes raras que está implicado en la base genética que subyace a la variabilidad de los fenotipos cuantitativos FVIII y FXI de la vía intrínseca de la coagulación. Asimismo, se ha constatado la importancia tanto de las variantes codificantes como no codificantes en la regulación de los niveles de estos fenotipos. De esta manera, se muestra la extrema complejidad genética no sólo de los fenotipos intermediarios de estudio, sino también de la VTE.

- La metodología de NGS implementada en el estudio de los genes *KNG1*, *F11* y *PLG* ha permitido la identificación de variantes genéticas potencialmente patogénicas relacionadas con los niveles de FXI y con la deficiencia de plasminógeno. Aunque la investigación traslacional de estos determinantes genéticos no ha podido establecer la implicación de las mutaciones en el riesgo de VTE, la metodología presentada aporta grandes ventajas en la investigación

genética con implicaciones diagnósticas, siendo aplicable al estudio de otras enfermedades.

- Los resultados derivados de la presente Tesis Doctoral suponen un avance en la identificación de los determinantes genéticos implicados en la VTE, sin embargo aún queda una parte importante por explicar de la base genética de la susceptibilidad a padecer esta patología, lo que se conoce como "heredabilidad perdida". En este sentido, no sólo la secuencia de nuestro DNA sino otros mecanismos como la regulación de la expresión génica pueden jugar un papel importante.

Bibliografía

Bibliografía

Abecasis GR, Auton A, Brooks LD, et al (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65.

Adzhubei IA, Schmidt S, Peshkin L, et al (2010) A method and server for predicting damaging missense mutations. Nat Methods 7:248–249.

Almasy L, Blangero J (2010) Variance component methods for analysis of complex phenotypes. Cold Spring Harb Protoc 2010:pdb.top77.

Anderson FA, Spencer FA (2003) Risk factors for venous thromboembolism. Circulation 107:I9–16.

Austin SK (2013) Haemostasis. Medicine (Baltimore) 41:208–211.

Bach RR (2005) Tissue Factor Encryption. Arterioscler Thromb Vasc Biol 26:456–461.

Bentley G, Higuchi R, Hoglund B, et al (2009) High-resolution, high-throughput HLA genotyping by next-generation sequencing. Tissue Antigens 74:393–403.

Bertina RM, Koeleman BP, Koster T, et al (1994) Mutation in blood coagulation factor V associated with resistance to activated protein C. Nature 369:64–67.

Blangero J, Williams JT, Almasy L (2003) Novel family-based approaches to genetic risk in thrombosis. J Thromb Haemost 1:1391–1397.

Bonnefond A, Froguel P (2015) Rare and Common Genetic Events in Type 2 Diabetes: What Should Biologists Know? Cell Metab 21:357–368.

Bouma BN, Meijers JC (1999) Fibrinolysis and the contact system: a role for factor XI in the down-regulation of fibrinolysis. Thromb Haemost 82:243–250.

Brandt JT (2002) Plasminogen and tissue-type plasminogen activator deficiency as risk factors for thromboembolic disease. Arch Pathol Lab Med 126:1376–1381.

Burmeister M (1999) Basic concepts in the study of diseases with complex genetics. Biol Psychiatry 45:522–532.

Calafell F, Almasy L, Sabater-Lleal M, et al (2010) Sequence variation and genetic evolution at the human F12 locus: mapping quantitative trait nucleotides that influence FXII plasma levels. Hum Mol Genet 517–525.

Carlson CS, Eberle MA, Kruglyak L, Nickerson DA (2004) Mapping complex disease loci in whole-genome association studies. Nature 429:446–452.

Carter AM, Sachchithananthan M, Stasinopoulos S, et al (2002) Prothrombin G20210A is a bifunctional gene polymorphism. Thromb Haemost 87:846–853.

Cesarman-Maus G, Hajjar KA (2005) Molecular mechanisms of fibrinolysis. Br J Haematol 129:307–321.

Chen VM, Hogg PJ (2013) Encryption and decryption of tissue factor. J Thromb Haemost 11:277–284.

Cohen AT, Agnelli G, Anderson FA, et al (2007) Venous thromboembolism (VTE) in Europe. The number of VTE events and associated morbidity and mortality. Thromb Haemost 98:756–764.

Colman RW, Marder VJ, Clowes AW, et al (2006) Hemostasis and Thrombosis: Basic Principles and Clinical Practice. Lippincott Williams & Wilkins

Comp PC, Esmon CT (1984) Recurrent venous thromboembolism in patients with a partial deficiency of protein S. N Engl J Med 311:1525–1528.

Crawley JT, Gonzalez-Porras JR, Lane DA (2011) Textbook of Pulmonary Vascular Disease. Textbook of Pulmonary Vascular Disease

Cunha MLR, Meijers JCM, Middeldorp S (2015) Introduction to the analysis of next generation sequencing data and its application to venous thromboembolism. Thromb Haemost 114:920–932.

Cushman M, Tsai AW, White RH, et al (2004) Deep vein thrombosis and pulmonary embolism in two cohorts: the longitudinal investigation of thromboembolism etiology. Am J Med 117:19–25.

Dahlbäck B (2005) Blood coagulation and its regulation by anticoagulant pathways: genetic pathogenesis of bleeding and thrombotic diseases. J Intern Med 257:209–223.

Dahlbäck B (2000) Blood coagulation. Lancet 355:1627–1632.

Dahlbäck B, Carlsson M, Svensson PJ (1993) Familial thrombophilia due to a previously unrecognized mechanism characterized by poor anticoagulant response to activated protein C: prediction of a cofactor to activated protein C. Proc Natl Acad Sci U S A 90:1004–1008.

Davie EW, Ratnoff OD (1964) Waterfall Sequence for Intrinsic Blood Clotting. Science (80-) 145:1310–1312.

De Haan HG, Bezemer ID, Doggen CJM, et al (2012) Multiple SNP testing improves risk prediction of first venous thrombosis. Blood 120:656–663.

Desmet FO, Hamroun D, Lalande M, et al (2009) Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res 37:e67.

Duchemin J, Pan-Petesch B, Arnaud B, et al (2008) Influence of coagulation factors and tissue factor concentration on the thrombin generation test in plasma. Thromb Haemost 99:767–773.

Egeberg O (1965) Inherited antithrombin deficiency causing thrombophilia. Thromb Diath Haemorrh 13:516–530.

Eid J, Fehr A, Gray J, et al (2009) Real-time DNA sequencing from single polymerase molecules. Science 323:133–138.

Evangelou E, Trikalinos TA, Salanti G, Ioannidis JPA (2006) Family-based versus unrelated case-control designs for genetic associations. PLoS Genet 2:e123.

Fedurco M, Romieu A, Williams S, et al (2006) BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic Acids Res 34:e22.

Gailani D, Renné T (2007) The intrinsic pathway of coagulation: a target for treating thromboembolic disease? J Thromb Haemost 5:1106–1112.

Gehring NH, Frede U, Neu-Yilik G, et al (2001) Increased efficiency of mRNA 3' end formation: a new genetic mechanism contributing to hereditary thrombophilia. Nat Genet 28:389–392.

George JN (2000) Platelets. Lancet (London, England) 355:1531–1539.

Glenn TC (2011) Field guide to next-generation DNA sequencers. Mol Ecol Resour 11:759–769.

Goldhaber SZ, Bounameaux H (2012) Pulmonary embolism and deep vein thrombosis. Lancet 379:1835–1846.

Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17:333–351.

Gray IC, Campbell DA, Spurr NK (2000) Single nucleotide polymorphisms as tools in human genetics. Hum Mol Genet 9:2403–2408.

Griffin JH, Evatt B, Zimmerman TS, et al (1981) Deficiency of protein C in congenital thrombotic disease. J Clin Invest 68:1370–1373.

Grupo Multidisciplinar para el Estudio de la Enfermedad Tromboembólica en España (2006) Estudio sobre la enfermedad tromboembólica venosa en España.

Harismendy O, Bansal V, Bhatia G, et al (2010) Population sequencing of two endocannabinoid metabolic genes identifies rare and common regulatory variants associated with extreme obesity and metabolite level. Genome Biol 11:R118.

Heemskerk JWM, Bevers EM, Lindhout T (2002) Platelet activation and blood coagulation. Thromb Haemost 88:186–193.

Heit JA, Phelps MA, Ward SA, et al (2004) Familial segregation of venous thromboembolism. J Thromb Haemost 2:731–736.

Herrmann J, Lerman A (2001) The endothelium: dysfunction and beyond. J Nucl Cardiol 8:197–206.

Hirschhorn JN (2005) Genetic approaches to studying common diseases and complex traits. Pediatr Res 57:74R–77R.

Hoffman M, Monroe DM (2001) A cell-based model of hemostasis. Thromb Haemost 958–965.

Houlihan LM, Davies G, Tenesa A, et al (2010) Common variants of large effect in F12, KNG1, and HRG are associated with activated partial thromboplastin time. Am J Hum Genet 86:626–631.

International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–1320.

Jackson SP (2011) Arterial thrombosis--insidious, unpredictable and deadly. Nat Med 17:1423–1436.

Jain M, Olsen HE, Paten B, Akeson M (2016) The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol 17:239.

Jick H, Slone D, Westerholm B, et al (1969) Venous thromboembolic disease and ABO blood type. A cooperative study. Lancet (London, England) 1:539–542.

Jobling L, Eyre L (2013) Haemostasis, blood platelets and coagulation. Anaesth Intensive Care Med 14:51–53.

Jones AL, Hulett MD, Parish CR (2005) Histidine-rich glycoprotein: A novel adaptor protein in plasma that modulates the immune, vascular and coagulation systems. Immunol Cell Biol 83:106–118.

Kathiresan S, Musunuru K, Orho-Melander M (2008) Defining the spectrum of alleles that contribute to blood lipid concentrations in humans. Curr Opin Lipidol 19:122–127.

Kim S, Misra A (2007) SNP Genotyping: Technologies and Biomedical Applications. Annu Rev Biomed Eng 9:289–320.

Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 4:1073–1081.

Kyrle PA, Mannhalter C, Béguin S, et al (1998) Clinical studies and thrombin generation in patients homozygous or heterozygous for the G20210A mutation in the prothrombin gene. Arterioscler Thromb Vasc Biol 18:1287–1291.

Lander ES, Linton LM, Birren B, et al (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921.

Lane DA, Mannucci PM, Bauer KA, et al (1996) Inherited thrombophilia: Part 1. Thromb Haemost 76:651–662.

Lee S, Miropolsky L, Wu M (2015) SKAT: SNP-Set (Sequence) Kernel Association Test. R package version 1.0.9.

Levy SE, Myers RM (2016) Advancements in Next-Generation Sequencing. Annu Rev Genomics Hum Genet 17:95–115.

Li L, Li Y, Browning SR, et al (2011) Performance of Genotype Imputation for Rare Variants Identified in Exons and Flanking Regions of Genes. PLoS One 6:e24945.

Liu JZ, Almarri MA, Gaffney DJ, et al (2012) Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis. Nat Genet 44:1137–1141.

Macfarlane RG (1964) An Enzyme Cascade in the Blood Clotting Mechanism, and its Function as a Biochemical Amplifier. Nature 202:498–499.

Mackman N (2012) New insights into the mechanisms of venous thrombosis. J Clin Invest 122:2331–2336.

MacQuarrie JL, Stafford AR, Yau JW, et al (2011) Histidine-rich glycoprotein binds factor XIIa with high affinity and inhibits contact-initiated coagulation. Blood 117:4134–4141.

Makris M, Preston FE, Beauchamp NJ, et al (1997) Co-inheritance of the 20210A allele of the prothrombin gene increases the risk of thrombosis in subjects with familial thrombophilia. Thromb Haemost 78:1426–1429.

Mann KG, Butenas S, Brummel K (2003) The dynamics of thrombin formation. Arterioscler Thromb Vasc Biol 23:17–25.

Manolio TA, Collins FS, Cox NJ, et al (2009) Finding the missing heritability of complex diseases. Nature 461:747–753.

Marian AJ (2012) Molecular genetic studies of complex phenotypes. Transl Res 64–79.

Martinelli I, De Stefano V, Mannucci PM (2014) Inherited risk factors for venous thromboembolism. Nat Rev Cardiol 11:140–156.

Mateo J (2013) Nuevos anticoagulantes orales y su papel en la práctica clínica. Rev Española Cardiol Supl 13:33–41.

Metzker ML (2010) Sequencing technologies - the next generation. Nat Rev Genet 11:31–46.

Morange P, Oudot-Mellakh T (2011) KNG1 Ile581Thr and susceptibility to venous thrombosis. Blood 117:3692–3694.

Morange PE, Suchon P, Trégouët DA (2015) Genetics of Venous Thrombosis: update in 2015. Thromb Haemost 114:910–919.

Morange P-E, Tregouet D-A (2010) Deciphering the molecular basis of venous thromboembolism: where are we and where should we go? Br J Haematol 148:495–506.

Morange P-E, Trégouët D-A (2013) Current knowledge on the genetics of incident venous thrombosis. J Thromb Haemost 11 Suppl 1:111–121.

Mosesson MW (2005) Fibrinogen and fibrin structure and functions. J Thromb Haemost 3:1894–1904.

Müller F, Gailani D, Renné T (2011) Factor XI and XII as antithrombotic targets. Curr Opin Hematol 18:349–355.

Neale BM, Medland SE, Ripke S, et al (2010) Meta-analysis of genome-wide association studies of attention-deficit/hyperactivity disorder. J Am Acad Child Adolesc Psychiatry 49:884–897.

Norris LA (2003) Blood coagulation. Best Pract Res Clin Obstet Gynaecol 17:369–383.

Ott J, Kamatani Y, Lathrop M (2011) Family-based designs for genome-wide association studies. Nat Rev Genet 12:465–474.

Paramo JA, Ruiz-de-Gaona E, García R, Rodríguez P (2007) Diagnóstico y tratamiento de la trombosis venosa profunda. Rev Med Univ Navarra 51:13–17.

Park SJ, Saito-Adachi M, Komiyama Y, Nakai K (2015) Advances, Practice, and Clinical Perspectives in High-throughput Sequencing.

Peck-Radosavljevic M (2007) Review article: coagulation disorders in chronic liver disease. Aliment Pharmacol Ther 26:21–28.

Perkel J (2008) SNP genotyping: six technologies that keyed a revolution. Nat Methods 5:447–453.

Pertea M, Lin X, Salzberg SL (2001) GeneSplicer: a new computational method for splice site prediction. Nucleic Acids Res 29:1185–1190.

Poort SR, Rosendaal FR, Reitsma PH, Bertina RM (1996) A common genetic variation in the 3'-untranslated region of the prothrombin gene is associated with elevated plasma prothrombin levels and an increase in venous thrombosis. Blood 88:3698–3703.

Quélin F, Mathonnet F, Potentini-Esnault C, et al (2006) Identification of five novel mutations in the factor XI gene (F11) of patients with factor XI deficiency. Blood Coagul Fibrinolysis 17:69–73.

Reese MG, Eeckman FH, Kulp D, Haussler D (1997) Improved Splice Site Detection in Genie. J Comput Biol 4:311–323.

Reitsma PH, Versteeg HH, Middeldorp S (2012) Mechanistic view of risk factors for venous thromboembolism. Arterioscler Thromb Vasc Biol 32:563–568.

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Sci 273:1516–1517.

Rocanin-Arjo A, Cohen W, Carcaillon L, et al (2014) A meta-analysis of genome-wide association studies identifies ORM1 as a novel gene controlling thrombin generation potential. Blood 123:777–785.

Rosendaal FR (2005) Venous thrombosis: the role of genes, environment, and behavior. Hematology Am Soc Hematol Educ Program 2005:1–12.

Rosendaal FR (1999) Venous thrombosis: a multicausal disease. Lancet 353:1167–1773.

Ross R (1999) Atherosclerosis--an inflammatory disease. N Engl J Med 340:115–126. doi: 10.1056/NEJM199901143400207

Rothberg JM, Hinz W, Rearick TM, et al (2011) An integrated semiconductor device enabling non-optical genome sequencing. Nature 475:348–352.

Ruggeri ZM (2001) Structure of von Willebrand factor and its function in platelet adhesion and thrombus formation. Best Pract Res Clin Haematol 14:257–279.

Sabater-Lleal M, Martinez-Perez A, Buil A, et al (2012) A genome-wide association study identifies KNG1 as a genetic determinant of plasma factor XI Level and activated partial thromboplastin time. Arterioscler Thromb Vasc Biol 32:2008–2016.

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74:5463–5467.

Schlötterer C (2004) Opinion: The evolution of molecular markers — just a matter of fashion? Nat Rev Genet 5:63–69.

Schwarz JM, Rödelsperger C, Schuelke M, Seelow D (2010) MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods 7:575–576.

Segers O, van Oerle R van, ten Cate H ten, et al (2010) Thrombin generation as an intermediate phenotype for venous thrombosis. Thromb Haemost 103:114–122.

Sennblad B, Basu S, Mazur J, et al (2017) Genome-wide association study with additional genetic and post-transcriptional analyses reveals novel regulators of plasma factor XI levels.

Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26:1135–1145.

Sidelmann JJ, Gram J, Jespersen J, Kluft C (2000) Fibrin clot formation and lysis: basic mechanisms. Semin Thromb Hemost 26:605–618.

Smith SA (2009) The cell-based model of coagulation. J Vet Emerg Crit Care (San Antonio) 19:3–10.

Soria JM, Almasy L, Souto JC, et al (2002) A quantitative-trait locus in the human factor XII gene influences both plasma factor XII levels and susceptibility to thrombotic disease. Am J Hum Genet 70:567–574.

Soria JM, Morange P-E, Vila J, et al (2014) Multilocus Genetic Risk Scores for Venous Thromboembolism Risk Assessment. J Am Heart Assoc 3:e001060.

Souto JC (2002) Search for new thrombosis-related genes through intermediate phenotypes. Genetic and household effects. Pathophysiol Haemost Thromb 32:338–340.

Souto JC, Almasy L, Borrell M, et al (2000a) Genetic susceptibility to thrombosis and its relationship to physiological risk factors: the GAIT study. Am J Hum Genet 67:1452–1459.

Souto JC, Almasy L, Borrell M, et al (2000b) Genetic determinants of hemostasis phenotypes in Spanish families. Circulation 101:1546–1551.

Stratton MR, Rahman N (2008) The emerging landscape of breast cancer susceptibility. Nat Genet 40:17–22.

Tang W, Schwienbacher C, Lopez LM, et al (2012) Genetic Associations for Activated Partial Thromboplastin Time and Prothrombin Time, their Gene Expression Profiles, and Risk of Coronary Artery Disease.

Tavtigian S V, Deffenbaugh AM, Yin L, et al (2006) Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. J Med Genet 43:295–305.

Terwilliger JD, Weiss KM (1998) Linkage disequilibrium mapping of complex disease: fantasy or reality? Curr Opin Biotechnol 578–594.

Tirado I, Mateo J, Soria JM, et al (2005) The ABO blood group genotype and factor VIII levels as independent risk factors for venous thromboembolism. Thromb Haemost 93:468–474.

Townsend N, Wilson L, Bhatnagar P, et al (2016) Cardiovascular disease in Europe: epidemiological update 2016. Eur Heart J 37:3232–3245.

Tsuji S (2010) Genetics of neurodegenerative diseases: insights from high-throughput resequencing. Hum Mol Genet 19:R65–R70.

Turcatti G, Romieu A, Fedurco M, Tairi AP (2008) A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. Nucleic Acids Res 36:e25.

Uitte de Willige S, de Visser MCH, Houwing-Duistermaat JJ, et al (2005) Genetic variation in the fibrinogen gamma gene increases the risk for deep venous thrombosis by reducing plasma fibrinogen gamma' levels. Blood 106:4176–4183.

Van Boven HH, Reitsma PH, Rosendaal FR, et al (1996) Factor V Leiden (FV R506Q) in families with inherited antithrombin deficiency. Thromb Haemost 75:417–421.

Van Nimwegen KJM, van Soest RA, Veltman JA, et al (2016) Is the $1000 Genome as Near as We Think? A Cost Analysis of Next-Generation Sequencing. Clin Chem 62:1458–1464.

Van Veen JJ, Gatt A, Makris M (2008) Thrombin generation testing in routine clinical practice: are we there yet? Br J Haematol 142:889–903.

Venselaar H, te Beek TA, Kuipers RK, et al (2010) Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. BMC Bioinformatics 11:548.

Vilalta N, Souto JC (2014) Investigación de la trombofilia venosa. Presente y futuro. Angiologia 66:190–198.

Virchow R (1856) Phlogose und thrombose im gefässsystem. Gesammelte Abhandlungen Zur Wißenschaftlischen Medizin. Frankfurt-am-Main Von Meidinger Sohn Comp 458–636.

Visscher PM, Hill WG, Wray NR (2008) Heritability in the genomics era — concepts and misconceptions. Nat Rev Genet 9:255–266.

Wallis Y, Payne S, McAnulty C, Bodmer D (2013) Practice guidelines for the evaluation of pathogenicity and the reporting of sequence variants in clinical molecular genetics.

Wang Q, Lu Q, Zhao H (2015) A review of study designs and statistical methods for genomic epidemiology studies using next generation sequencing. Front Genet 6:149.

Welter D, MacArthur J, Morales J, et al (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res 42:D1001–D1006.

White RH (2003) The Epidemiology of Venous Thromboembolism. Circulation 107:I4–I8.

Wilbur J, Shian B (2012) Diagnosis of deep venous thrombosis and pulmonary embolism. Am Fam Physician 86:913–919.

Willer CJ, Sanna S, Jackson AU, et al (2008) Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet 40:161–169.

Williams JT, Blangero J (1999) Power of variance component linkage analysis to detect quantitative trait loci. Ann Hum Genet 63:545–563.

Wu MC, Lee S, Cai T, et al (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 89:82–93.

Xiong M, Jin L (1999) Comparison of the Power and Accuracy of Biallelic and Microsatellite Markers in Population-Based Gene-Mapping Methods. Am J Hum Genet 64:629–640.

Yeo G, Burge CB (2004) Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol 11:377–394.

Zeggini E, Morris AP (2010) Analysis of Complex Disease Association Studies: A Practical Guide. Academic Press

Anexos

S1 Table. Primers used for amplification of *CIB4* loci.

PCR	PCR forward primer	Forward primer position*	PCR reverse primer	Reverse primer position*	Size (bp)
CIB4 LR1	GGGATGTTTACTCAGGAGAGCATAGG	chr2:26,865,670-26,865,695	CATGTGAAAGGAAAGTCTCATAGGCTG	chr2:26,858,044-26,858,070	7652
CIB4 LR2	GACTTGCATAAGGACACAAAGGCCTGG	chr2:26,858,213-26,858,239	GACTCTTCCATCATGTGGCCCCTTGG	chr2:26,850,160-26,850,185	8080
CIB4 LR3	CCAGACTCTCTGAATGCTACGACTGG	chr2:26,850,283-26,850,308	CTATATGTGCGGAAGATTCTACGAGGG	chr2: 26,842,280-26,842,306	8029
CIB4 LR4	TGAATCTGCTGTTGTTTCCCTGTCAGG	chr2:26,842,483-26,842,509	TGGACAGTTCTATCCATCTAAAACCTCC	chr2:26,834,456-26,834,483	8054
CIB4 LR5	CATTGAGTTCTGATTTACTAAACCATCTCC	chr2:26,834,622-26,834,651	ACAGACCCATGCTTATGCCTGTAATCC	chr2:26,826,588-26,826,614	8064
CIB4 LR6	AGTCTCCTGAAGAGCTATGACTACAGG	chr2:26,826,745-26,826,771	GTAGCTGCTGGTAGCACGAATAGTACC	chr2:26,818,839-26,818,865	7933
CIB4 LR7	CTAGGAAAACCACATTAGAAAATCTATACC	chr2:26,818,983-26,819,012	CCATACTCATTTCTACTAGAGAAACTAAGG	chr2:26,810,728-26,810,757	8285
CIB4 LR8	CCTTCACCTGCTCTGAGCCATTTCCC	chr2:26,810,950-26,810,975	TCTCTGTGCAGCTGTGAACTGTGAGG	chr2:26,803,809-26,803,834	7167

* Primer positions referred to GRCh37/hg19, UCSC Genome Browser assembly, February 2009 release (http://genome.ucsc.edu/) [1].

S2 Table. Polymorphism Associations with FVIII Levels Reported up to this Date.

Author	Gene	Polymorphism	Chr	*P*-value *	% FVIII variance**	Population studied
Morange et al. 2005 [42]	*LRP*	D2080N	12	3.0e-02	low	A subsample of the Stanislas Cohort including 100 healthy nuclear families of European origin (200 parents and 224 offsrings)
Scanavini et al. 2005 [19]	*F8*	D1241E	X	<5.0e-02	10	Cohort 1: 150 unrelated thrombotic women and 145 healthy control women Cohort 2: 283 unrelated heterozygous carries of the FV Leiden (140 VT patients and 143 asymptomatic carriers)
Marchetti et al. 2006 [43]	*LRP1*	-25 C > G	12	2.0e-02	low	200 unrelated women with DVT
Nossent et al. 2006 [20]	*F8*	D1241E (HT1) [a]	X	---	low	Combined control populations coming from LETS, SMILE and RATIO case-control studies †
Vormittag et al. 2007 [44]	*LRP1*	663 C > T	12	2.0e-02	---	152 patients with recurrent VTE and 198 healthy controls
Viel et al. 2007 [21]	*F8*	D1241E	X	<5.0e-02	10	137 unrelated healthy subjects and 398 subjects from The GAIT Project
Berger et al. 2008 [45]	*ADAMDEC1*	tgtgg/tgtgg [b]	8	4.6e-02	low	165 unrelated DVT or PE patients with high FVIII levels from the MAISTHRO (main-Isar-Thromboseregister) and 214 healthy controls
Smith et al. 2010 [46]	*ABO*	rs687289	9	<5.0e-324	10	The CHARGE Consortium (23,608 participants of European ancestry coming from 5 population-based cohorts ‡)
	VWF	rs1063856	12	3.6e-09		
	STXBP5	rs9390459	6	6.7e-10		
	SCARA5	rs9644133 rs11780263 [c]	8	4.4e-15 ---		
	STAB2	rs12229292	12	7.2e-09		
Antoni et al. 2011 [24]	*LBH*	rs6708166	2	1.30e-06	6.3	Combined analysis of three GWAS including French-Canadian families with FV Leiden (N=253), MARTHA08 § subjects (N=1,006) and MARTHA10 § subjects (N=586).
	FAM46A	rs1321761	6	9.54e-06		
	VAV2	rs12344583	9	7.92e-06		
	STAB2	rs7306642	12	2.95e-06		
	ACCN1	rs1354492	17	2.42e-06		

Author	Gene	Polymorphism	Chr	*P*-value *	% FVIII variance**	Population studied
Campos et al. 2012 [22]	*F8*	rs5945122 [d] rs6643714 [d] rs7061362 [d] rs5945258 [d] haplotype AT [e] haplotype GCTTTT [f]	X	3.9e-04 8.8e-05 3.2e-04 3.0e-04 2.0e-04 2.0e-04	---	The Artheriosclerosis Risk in Communities (ARIC) Study (10,434 healthy Americans of European (EA) or African (AA) descents)
	VWF	rs1063857 [g] rs723190 [g] rs216315 [h] rs216318 [h] rs216299 [h] rs216295 [h] rs216298 [h]	12	4.3e-05 2.5e-04 2.7e-06 2.3e-06 6.2e-05 8.0e-05 3.6e-05		
Shen et al. 2013 [23]	*F8*	CNV [i]	X	6.1e-14, OR=12.2 (males) 4.3e-10, OR=9.5 (females)	---	179 VTE patients (98 males and 81 females) and 176 healthy controls (93 males and 83 females)

Chr: chromosome; DVT: deep vein thrombosis; PE: pulmonary embolism; VT: venous thrombosis; VTE: venous thromboembolism. * *P*-value of the association with FVIII levels. ** Percentage of the variation of FVIII coagulant activity that is explained by the indicated loci. [a] The E allele of the D1241E polymorphism was associated with a 6% reduction in FVIII levels and an odds ratio (OR)=0.4 for VT risk in males carryig the HT1 haplotype. [b] Corresponds to the *ADAMDEC1* haplotype combination that is comprised of rs12674766, rs10087305, rs2291577, rs2291578 and rs3765124. [c] The rs11780263 showed a cohort-specific significance for the association with FVIII levels in ARIC (*P* = 2.0e-10) and CHS (*P* = 7.1e-06). [d] Polymorphisms associated with FVIII levels only in EA males. [e] Haplotype combination comprised of rs4898352 and rs6643714 in EA subjects and associated with FVIII levels in males only. [f] Haplotype combination comprised of rs6643622, rs5945122, rs5945258, rs7061362, rs5987077 and rs1470586 in EA subjects and associated with FVIII levels in males only. [g] Polymorphisms associated with FVIII levels only in EA males. [h] Polymorphisms associated with FVIII levels only in EA females. [i] Copy number variation of the *F8* gene associated with plasma FVIII levels and VTE. † Controls from LETS (Leiden Thrombophilia Study) consisted of 474 healthy subjects (272 women and 202 men); controls from SMILE (Study of Myocardial Infarctions Leiden) consisted of 646 short-anticoagulated men; controls from RATIO (Risk of Arterial Thrombosis in Relation to Oral Contaceptives Study) consisted of 639 healthy women that were included in the study. ‡ The 5 population-based cohorts of adults in the United States and Europe were The Atherosclerosis Risk In Communities (ARIC) Study, The Cardiovascular Health Study (CHS), The Framingham Heart Study (FHS), The Rotterdam Study (RS) and The British 1958 Birth Cohort (B58C). § Unrelated subjects with VT from The Marseille Thrombosis Association Project.

References of the Supporting Information

1. Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, Harte RA, Heitner S, Hinrichs AS, Learned K, Lee BT, Li CH, Raney BJ, Rhead B, Rosenbloom KR, Sloan CA, Speir ML, Zweig AS, Haussler D, Kuhn RM, Kent WJ. The UCSC Genome Browser database: 2014 update. Nucleic Acids Res. 2014;42:D764–70.

Anexo II: Material suplementario del Artículo 3

S1 Table. Description of the phenotypes.

Trait	Description
FVII	Plasma levels of coagulation factor VII
TF	Plasma levels of Tissue Factor (or factor III (FIII))
TFPI	Plasma levels of Tissue Factor Pathway Inhibitor
FvW	Plasma levels of von Willebrand Factor
PC	Plasma levels of protein C (activated)
PST	Plasma levels of Total Protein S
PSF	Plasma levels of Total Free Protein S
PSFT	Plasma levels of Total Functional Protein S
PK	Plasma levels of Prekalikrein
P_sel	Plasma levels of P-selectin
FXII	Plasma levels of coagulation factor XII
FXI	Plasma levels of coagulation factor XI
FIX	Plasma levels of coagulation factor IX
FVIII	Plasma levels of coagulation factor VIII
HRG	Plasma levels of Histidine-Rich Glyvoprotein
FXIIIA	Plasma levels of coagulation factor XIII activated
FXIIIS	Plasma levels of coagulation factor XIII
FV	Plasma levels of coagulation factor V
FIB	Plasma levels of Fibrinogen
FII	Plasma levels of coagulation factor II (Prothrombin)
FX	Plasma levels of coagulation factor X
AT	Plasma levels of Antithrombin
HC2	Plasma levels of Heparin Cofactor II
PLG	Plasma levels of Plasminogen
PAI	Plasma levels of Pasminogen Activator Inhibitor
tPA	Plasma levels of Plasminogen Tissue Activator
uPA	Plasma levels of urokinase-type Plasminogen Tissue Activator

S2 Table. Heritabilities ofPCA-based metaphenotypes (components 1 to 27 from the PCA model).

Metaphenotype	h2r
PCA_C1	0.38*
PCA_C2	0.5***
PCA_C3	0.34***
PCA_C4	0
PCA_C5	0.51***
PCA_C6	0.57***
PCA_C7	0.37***
PCA_C8	0.49***
PCA_C9	0.49***
PCA_C10	0.53***
PCA_C11	0.55***
PCA_C12	0.59***
PCA_C13	0.44***
PCA_C14	0.56***
PCA_C15	0.41***
PCA_C16	0.35***
PCA_C17	0,14
PCA_C18	0.43***
PCA_C19	0.25***
PCA_C20	0.62***
PCA_C21	0.34***
PCA_C22	0.51***
PCA_C23	0.25***
PCA_C24	0.43***
PCA_C25	0.48***
PCA_C26	0.2***
PCA_C27	0,03

Significant thresholds for heritability estimation: * <0.05, **<0.005, *** <0.0005

S3 Table. P-values of associations of the relevant SNPs of this study with all the original phenotypes.

trait	rs9898	rs3733159	rs1621816	rs1403694	rs17255413	rs3113727	rs2731672
FXII	0,125	0,122	0,958	0,417	0,028	0,318	$7,6x10^{-36}$
FXI	0,073	0,211	0,219	0,082	0,541	0,512	0,337
FIX	0,420	0,984	0,671	0,407	0,036	0,405	0,461
FVIII	0,172	0,617	0,226	0,790	0,531	0,267	0,150
FvWIn	0,888	0,945	0,767	0,737	0,557	0,939	0,828
FVII	0,008	0,051	0,231	0,899	0,203	0,012	0,212
FX	0,163	0,413	0,121	0,200	0,121	0,636	0,539
FV	0,848	0,560	0,541	0,732	0,919	0,690	0,044
FII	0,864	0,784	0,219	0,311	0,114	0,113	0,633
FXIIIA	0,880	0,745	0,549	0,533	0,872	0,949	0,169
FXIIIS	0,536	0,591	0,927	0,793	0,674	0,658	0,904
PCAM	0,901	0,514	0,211	0,504	0,367	0,056	0,832
PST	0,009	0,138	0,047	0,075	0,235	0,361	0,514
psfR	0,482	0,645	0,682	0,920	0,688	0,005	0,978
PSF	0,064	0,323	0,088	0,069	0,553	0,139	0,353
PSFT	0,288	0,695	0,322	0,118	0,837	0,682	0,961
TFIn	0,267	0,686	0,187	0,183	0,182	0,444	0,429
TFPI	0,129	0,369	0,396	0,830	0,879	0,365	0,342
PK	0,501	0,760	0,002	0,001	0,028	0,416	0,072
P_sel	0,600	0,550	0,252	0,600	0,316	0,319	0,048
HRG	$1,9x10^{-26}$	$3,3x10^{-13}$	$1,5x10^{-09}$	$1,1x10^{-08}$	0,417	0,058	0,660
FIB	0,453	0,776	0,800	0,974	0,629	0,064	0,914
AT	0,029	0,0797	0,282	0,540	0,047	0,009	0,243
HC2	0,447	0,395	0,046	0,504	0,004	0,150	0,967
PLG	0,728	0,292	0,452	0,733	$6,3x10^{-06}$	0,457	0,310
PAI	0,804	0,354	0,293	0,229	0,764	0,240	0,917
TPA	0,305	0,384	0,093	0,307	0,028	0,034	0,217
uPAInv	0,301	0,503	0,121	0,293	0,003	0,285	0,948

S4 Table. Loadings (weights) of the coagulation phenotypes in the obtained ICA-based metaphenotypes.

	ICA-C1	ICA-C2	ICA-C3	ICA-C4	ICA-C5	ICA-C6	ICA-C7	ICA-C8	ICA-C9	ICA-C10	ICA-C11	ICA-C12	ICA-C13	ICA-C14	ICA-C15
FVII	-0,07	-0,06	0,00	-0,08	-0,05	0,37	-0,07	0,21	-0,10	-0,01	0,30	0,25	-0,14	-0,04	-0,44
TF	0,00	-0,03	0,04	0,22	-0,17	0,28	-0,25	0,16	-0,16	-0,06	0,08	0,41	-0,04	0,13	0,41
TFPI	-0,04	-0,04	-0,02	0,10	-0,09	0,11	-0,57	-0,10	0,31	0,45	-0,28	0,03	0,28	0,01	0,14
FvW	-0,04	-0,07	0,18	0,25	0,15	-0,07	-0,01	0,08	0,26	-0,03	-0,01	0,02	0,03	0,06	-0,32
PC	-0,08	-0,03	-0,02	-0,14	0,03	0,11	0,03	-0,04	0,20	-0,15	0,22	0,32	-0,34	0,30	-0,11
PST	-0,09	0,01	-0,23	0,20	0,01	-0,14	0,07	-0,07	-0,04	0,05	0,13	$1,2x10^{-4}$	-0,11	-0,05	0,06
PSF	-0,09	0,15	0,02	0,12	0,05	-0,15	-0,03	-0,05	0,01	0,00	0,17	-0,01	-0,10	-0,05	0,10
PSFT	-0,06	0,17	-0,12	0,03	-0,05	0,18	0,06	0,05	0,04	0,05	-0,14	-0,01	0,18	0,05	-0,19
PK	-0,07	-0,06	0,03	-0,08	0,04	0,02	-0,09	0,16	-0,21	0,04	0,11	-0,08	-0,39	-0,11	0,36
P_sel	-0,03	0,02	0,06	0,12	-0,05	0,27	0,08	0,05	0,09	-0,24	-0,48	-0,29	-0,48	-0,53	-0,13
FXII	-0,04	-0,08	0,09	-0,05	0,33	0,12	0,06	0,00	-0,20	0,69	0,00	-0,15	-0,19	-0,29	0,06
FXI	-0,08	-0,05	0,05	-0,04	0,12	-0,04	0,02	0,07	-0,03	-0,07	0,07	-0,18	0,09	0,21	0,46
FIX	-0,09	-0,05	0,11	-0,03	-0,02	-0,08	0,12	-0,05	0,05	0,03	-0,04	0,09	0,11	0,18	0,06
FVIII	-0,05	-0,10	0,17	0,19	0,31	0,01	0,07	-0,04	0,27	-0,05	0,01	0,05	0,06	0,06	-0,25
HRG	-0,03	-0,03	0,04	0,01	0,17	0,12	-0,30	-0,70	-0,47	-0,34	0,01	-0,04	0,07	-0,05	-0,19
FXIIIA	-0,03	-0,03	0,02	0,02	-0,20	-0,26	-0,28	0,18	-0,22	0,03	-0,11	-0,47	-0,31	0,39	-0,44
FXIIIS	-0,06	-0,05	0,08	$-1,8x10^{-4}$	-0,18	-0,14	-0,17	0,29	-0,25	-0,07	-0,01	-0,12	0,22	0,06	-0,12
FV	-0,06	-0,04	0,05	-0,05	-0,09	0,35	0,10	0,01	0,21	-0,17	0,26	-0,62	0,18	0,19	0,31
FIB	-0,06	-0,13	0,02	-0,01	-0,10	-0,21	0,00	-0,05	0,30	-0,23	-0,13	0,05	0,15	-0,22	0,00
FII	-0,08	-0,04	-0,07	-0,22	0,04	0,00	0,01	-0,08	0,03	0,09	0,18	-0,13	0,39	-0,08	-0,25
FX	-0,09	0,05	-0,02	-0,14	-0,05	0,09	-0,06	0,13	-0,08	0,02	0,02	0,07	0,20	-0,39	-0,20
AT	-0,04	0,04	-0,15	-0,10	0,44	0,12	-0,04	0,15	-0,08	-0,16	-0,54	0,04	$2,5x10^{-4}$	0,54	0,12
HC2	-0,08	-0,03	-0,03	-0,17	-0,10	-0,11	-0,09	-0,09	0,10	-0,18	-0,24	0,06	0,01	-0,37	0,39
PLG	-0,08	-0,03	0,03	-0,20	0,04	-0,21	-0,04	0,18	-0,03	-0,05	-0,19	0,32	-0,09	-0,08	-0,05
PAI	-0,05	0,00	0,18	0,04	-0,18	-0,03	0,42	-0,13	-0,36	0,19	-0,29	0,11	0,26	0,13	0,05
tPA	-0,07	-0,04	0,17	0,11	-0,28	0,13	0,16	-0,11	-0,12	0,04	-0,15	0,11	-0,06	0,18	0,04
uPA	-0,02	0,02	-0,02	-0,19	-0,24	0,00	0,03	-0,42	0,30	0,28	-0,08	0,02	-0,45	0,33	-0,06

S5 Table. Loadings (weights) of the coagulation phenotypes in the obtained PCA-based metaphenotypes.

	PCA-C1	PCA-C2	PCA-C3	PCA-C4	PCA-C5	PCA-C6	PCA-C7	PCA-C8	PCA-C9	PCA-C10	PCA-C11	PCA-C12	PCA-C13	PCA-C14	PCA-C15	PCA-C16	PCA-C17	PCA-C18	PCA-C19	PCA-C20	PCA-C21	PCA-C22	PCA-C23	PCA-C24	PCA-C25	PCA-C26	PCA-C27
FVII	0,19	0,12	-3x10^{-3}	0,11	-0,06	0,40	-0,08	0,20	0,09	-0,27	-0,23	0,12	0,03	0,34	-0,21	-0,16	-0,26	-0,06	-0,13	0,26	0,09	0,04	-0,01	-7x10^{-17}	0	0	0
TF	-0,01	0,06	-0,05	-0,31	-0,19	0,30	-0,26	0,15	0,15	-0,07	-0,37	0,03	-0,10	-0,32	0,41	-0,35	0,06	0,09	-0,02	-0,23	-0,17	0,03	2x10^{-4}	-2x10^{-18}	-3x10^{-17}	1x10^{-15}	4x10^{-17}
TFPI	0,11	0,07	0,03	-0,14	-0,10	0,12	-0,60	-0,10	-0,28	0,25	-0,02	-0,23	-0,01	-0,11	-0,35	0,11	0,08	-0,15	-0,06	0,07	-0,08	-0,03	0,01	-2x10^{-17}	9x10^{-17}	-2x10^{-16}	1x10^{-16}
FvW	0,12	0,15	-0,26	-0,34	0,17	-0,07	-0,01	0,07	-0,24	0,01	-0,02	-0,03	-0,05	0,25	-0,02	-0,13	0,05	0,09	0,32	-0,08	0,15	-0,51	-0,03	-2x10^{-18}	5x10^{-17}	-5x10^{-17}	-2x10^{-17}
PC	0,24	0,07	0,02	0,19	0,03	0,12	0,03	-0,04	-0,18	-0,19	-0,28	0,28	-0,24	0,09	-0,07	0,27	0,05	-0,16	-0,16	-0,08	-0,33	-0,24	0,01	1x10^{-19}	5x10^{-17}	-6x10^{-17}	1x10^{-19}
PST	0,26	-0,02	0,33	-0,28	0,01	-0,15	0,07	-0,07	0,04	-0,11	-1x10^{-4}	0,09	0,04	-0,05	-0,02	-0,05	-0,02	-0,09	2x10^{-3}	0,07	0,03	0,02	-0,40	-0,17	-0,54	-0,41	-0,07
PSF	0,24	-0,31	-0,02	-0,16	0,06	-0,16	-0,04	-0,04	-0,01	-0,15	0,01	0,09	0,04	-0,08	-0,01	-0,09	-0,03	-0,11	-0,04	0,07	0,03	0,02	0,46	0,05	0,19	-0,37	0,57
PSFT	0,18	-0,35	0,17	-0,04	-0,06	0,19	0,07	0,05	-0,04	0,13	0,01	-0,15	-0,04	0,15	0,12	0,16	0,10	0,33	-0,10	-0,03	-0,05	-0,02	-0,01	-0,14	-0,34	0,44	0,41
PK	0,19	0,12	-0,04	0,11	0,05	0,02	-0,09	0,16	0,20	-0,10	0,07	0,32	0,08	-0,28	-0,31	0,02	0,24	0,27	0,01	-0,15	0,20	-0,07	0,01	-4x10^{-17}	-1x10^{-17}	9x10^{-17}	-1x10^{-16}
P_sel	0,08	-0,04	-0,09	-0,16	-0,06	0,29	0,08	0,05	-0,09	0,43	0,25	0,40	0,42	0,10	-0,07	2x10^{-3}	-0,05	-0,19	-0,09	-0,11	-0,24	0,06	-2x10^{-3}	7x10^{-17}	-2x10^{-17}	-9x10^{-17}	-5x10^{-17}
FXII	0,11	0,16	-0,13	0,06	0,37	0,12	0,07	-2x10^{-3}	0,19	4x10^{-3}	0,13	0,16	0,23	-0,04	0,38	0,06	-0,04	0,03	-0,17	0,08	0,02	-0,10	0,01	4x10^{-17}	6x10^{-18}	2x10^{-17}	8x10^{-17}
FXI	0,24	0,11	-0,07	0,05	0,13	-0,04	0,02	0,07	0,03	-0,07	0,16	-0,07	-0,17	-0,35	-0,20	0,21	-0,19	0,11	-0,05	-0,41	-0,14	0,24	-0,03	1x10^{-15}	1x10^{-16}	9x10^{-17}	-1x10^{-16}
FIX	0,25	0,09	-0,16	0,03	-0,03	-0,08	0,13	-0,04	-0,04	0,04	-0,08	-0,09	-0,14	-0,05	0,22	0,31	-0,20	-0,01	0,08	-0,04	-0,18	-0,29	4x10^{-3}	1x10^{-15}	-4x10^{-17}	3x10^{-17}	-5x10^{-17}
FVIII	0,15	0,20	-0,24	-0,26	0,35	0,01	0,07	-0,04	-0,25	-0,01	-0,05	-0,05	-0,05	0,20	-0,03	-0,09	-0,02	0,30	0,10	0,06	-0,16	0,55	0,01	6x10^{-19}	-7x10^{-18}	2x10^{-17}	-4x10^{-17}
HRG	0,08	0,06	-0,06	-0,02	0,18	0,13	-0,31	-0,69	0,44	-0,01	0,04	-0,06	0,04	0,14	0,01	0,08	0,07	0,13	-0,04	0,02	0,03	-0,07	-5x10^{-3}	9x10^{-18}	4x10^{-17}	8x10^{-18}	-7x10^{-17}
FXIIIA	0,09	0,05	-0,02	-0,03	-0,22	-0,28	-0,29	0,17	0,21	0,10	0,41	0,26	-0,31	0,34	0,16	-0,10	0,11	-0,01	0,03	-0,13	-0,18	-0,01	-5x10^{-3}	3x10^{-17}	3x10^{-17}	2x10^{-17}	3x10^{-18}
FXIIIS	0,17	0,11	-0,11	3x10^{-4}	-0,20	-0,15	-0,18	0,28	0,24	0,01	0,11	-0,18	-0,05	0,10	0,01	0,12	-0,40	0,13	-0,14	0,28	0,05	0,07	-5x10^{-4}	-1x10^{-17}	-3x10^{-17}	-5x10^{-18}	2x10^{-17}
FV	0,18	0,07	-0,07	0,07	-0,10	0,38	0,10	0,01	-0,20	-0,23	0,55	-0,15	-0,15	-0,24	0,05	-0,15	0,24	0,04	0,03	0,28	0,10	-0,12	-0,02	-2x10^{-16}	-6x10^{-18}	-6x10^{-18}	-7x10^{-18}
FIB	0,16	0,27	-0,03	0,01	-0,11	-0,22	1x10^{-3}	-0,05	-0,28	0,12	-0,05	-0,13	0,18	3x10^{-3}	0,18	-0,10	0,02	0,02	-0,64	-0,18	0,36	-0,06	0,02	3x10^{-15}	2x10^{-16}	9x10^{-17}	2x10^{-17}
FII	0,22	0,09	0,09	0,30	0,04	6x10^{-4}	0,01	-0,08	-0,03	-0,17	0,11	-0,32	0,06	0,19	0,16	-0,28	0,14	-0,38	0,04	-0,22	-0,24	0,17	0,01	1x10^{-15}	1x10^{-17}	-3x10^{-17}	7x10^{-17}
FX	0,24	-0,10	0,03	0,20	-0,06	0,09	-0,06	0,13	0,08	-0,01	-0,07	-0,17	0,31	0,16	-0,07	-0,03	-0,09	0,03	0,37	-0,44	0,30	-0,02	3x10^{-3}	1x10^{-17}	-2x10^{-16}	-1x10^{-16}	4x10^{-17}
AT	0,10	-0,08	0,21	0,13	0,49	0,13	-0,04	0,15	0,07	0,49	-0,03	-2x10^{-4}	-0,43	-0,10	0,04	-0,19	-0,14	-0,22	3x10^{-4}	0,04	0,25	-0,03	0,02	-1x10^{-17}	3x10^{-17}	9x10^{-17}	1x10^{-17}
HC2	0,23	0,05	0,04	0,24	-0,11	-0,12	-0,09	-0,09	-0,09	0,22	-0,05	-0,01	0,30	-0,30	0,17	-0,12	-0,26	0,10	0,34	0,30	-0,23	-0,07	-0,02	7x10^{-17}	-2x10^{-16}	-1x10^{-17}	4x10^{-17}
PLG	0,21	0,06	-0,04	0,27	0,05	-0,23	-0,04	0,17	0,03	0,17	-0,28	0,07	0,06	0,04	0,02	-0,01	0,59	0,11	0,04	0,27	-0,03	0,12	-0,02	-7x10^{-17}	2x10^{-16}	4x10^{-15}	1x10^{-17}
PAI	0,14	-4x10^{-3}	-0,26	-0,06	-0,20	-0,03	0,45	-0,13	0,33	0,26	-0,10	-0,21	-0,10	-0,04	-0,37	-0,36	0,03	0,01	-0,14	-2x10^{-3}	-0,19	-0,15	-0,01	-3x10^{-17}	8x10^{-17}	2x10^{-17}	2x10^{-16}
tPA	0,19	0,09	-0,24	-0,16	-0,31	0,14	0,17	-0,11	0,11	0,13	-0,10	0,05	-0,14	-0,03	0,17	0,36	0,14	-0,35	0,22	0,06	0,37	0,30	0,03	-7x10^{-17}	-2x10^{-17}	-3x10^{-17}	-8x10^{-18}
uPA	0,06	-0,03	0,03	0,26	-0,27	-2x10^{-3}	0,03	-0,41	-0,28	0,07	-0,02	0,38	-0,26	0,04	0,03	-0,26	-0,21	0,26	0,06	-0,08	0,18	0,12	-3x10^{-3}	-1x10^{-15}	1x10^{-15}	6x10^{-17}	-1x10^{-16}

S1 Table. Primers used for amplification of *KNG1* and *F11* loci.

PCR	PCR forward primer	Forward primer position[1]	PCR reverse primer	Reverse primer position[1]	Size (bp)
KNG1 Short[2]	CCATGTTTTACATCTCTCCAAGAA	chr3:186,433,630-186,433,653	GACGCTAGTTGCCTTGTTCC	chr3:186,435,160-186,435,179	1550
KNG1 LR1[3]	GCTTTGAGGGAGAAGGGTTC	chr3:186,434,347-186,434,366	CTCAGTTTACCCGTCTTGTAGAAGATGG	chr3:186,439,367-186,439,394	5048
KNG1 LR2[3]	GATGACTGCCAGTCATCTGCCATTACC	chr3:186,438,983-186,439,009	GACCAGTGACGTGAGCTTTAGTCTTCC	chr3:186,445,266-186,445,292	6310
KNG1 LR3[3]	CTCAATTATGTCCGTATAGTGCTATCAGG	chr3:186,444,510-186,444,538	CCATAATGACATCAGTCAGAGACTATTGG	chr3:186,451,791-186,451,819	7310
KNG1 LR4[3]	CTCTGTTTTCTATATGTAAAGGACTGAGG	chr3:186,450,655-186,450,683	CTCTTCATTACTTTCCTTGGAACATGTGG	chr3:186,456,934-186,456,962	6308
KNG1 LR5[3]	CCACTTAGAAAAGGAAATGGGACCAACC	chr3:186,456,680-186,456,707	CCTAGTCAACTTCTTCCAGTCACATTGG	chr3:186,463,137-186,463,164	6485
F11 LR1[3]	GCCTTCCAGATAGAACATCACC	chr4:187,185,623-187,185,644	GGAAAGTCTACACTTGGAGG	chr4:187,193,112-187,193,131	7509
F11 LR2[3]	GTGTTTATTGCCTTGATTTCC	chr4:187,192,623-187,192,643	CCTCGGTAATGTTGGGTTTG	chr4:187,197,713-187,197,732	5110
F11 LR3[3]	TCATCATGTCCTATACAGTTAGATCC	chr4:187,197,186-187,197,211	TGTTCTGATGCTGGGAGACC	chr4:187,205,575-187,205,594	8409
F11 LR4[3]	CACCATTCACGTTATCATTTGAAGGAGG	chr4:187,204,968-187,204,995	CCTGTAATCTCAGCTACTTGG	chr4:187,210,960-187,210,980	6013

[1] Primer positions referred to GRCh37/hg19, UCSC Genome Browser assembly, February 2009 release (http://genome.ucsc.edu/) [1]. [2] Primers used for Short PCR conditions. [3] Primers used for LR-PCR conditions.

S2 Table. PCR master mix for short and LR-PCR amplifications.

Short PCR Component	Short PCR Volum (µl)	LR-PCR Component	LR-PCR Volum (µl)	
			KNG1	***F11***
FastStart High Fidelity Reaction Buffer, 10X	1.88	SequalPrep 10X Reaction Buffer	2	
DMSO	0.94	DMSO	0.4	
PCR Grade Nucleotide Mix	0.38	SequalPrep 10X Enhancer B[1]	2	
FastStart High Fidelity Enzyme Blend 5 U/µl	0.19	SequalPrep Long Polymerase 5 U/µl	0.36	0.4
Primer mix (7.5 µM)	1.88	Primer mix (7.5 µM)	2	2.5
Template DNA (50 ng/µl)	1.13	Template DNA (50 ng/µl)	1	2
DNase-free water	to 18.75	DNase-free water	to 20	

[1] SequalPrep 10X Enhancer in LR-PCR was Enhancer B 1X in all PCR amplicons except for: F11 LR1, F11 LR2 and F11 LR3 (Enhancer A).

S3 Table. Thermocycling conditions for Short PCR amplifications.

Thermal cycler steps	Temperature (ºC)	Time	Cycles
Initial denaturation	95	2 min	1
Denaturation	95	30 sec	x 38
Annealing	56	30 sec	
Elongation	72	2 min	
Final elongation	72	4 min	1
Cooling	4	∞	

S4 Table. Thermocycling conditions for LR-PCR amplifications.

Thermal cycler steps	Temperature (ºC)	Time	Cycles
Initial denaturation	94	2 min	1
Denaturation	94	10 sec	
Annealing[1]	64	30 sec	x 10
Elongation[2]	68	18 min	
Denaturation	94	10 sec	
Annealing[1]	64	30 sec	x 22
Elongation[2]	68	18 min (+ 20 sec/cycle)	
Final elongation	72	5 min	1
Cooling	4	∞	

[1] Annealing temperature was 64ºC in all PCR amplicons except for: F11 LR1 (60ºC), F11 LR2 (54ºC) and F11 LR3 (54ºC).
[2] Elongation temperature in LR-PCR was 68ºC in all PCR amplicons except for F11 LR1 (60ºC).

S5 Table. All the unique biallelic variants identified in the *KNG1* and *F11* loci from the discovery sample.

Gene	Chromosome	Genome Location	Reference Allele	Alternate Allele	dbSNP v137
KNG1	3	186433681	C	T	rs60678173
KNG1	3	186433716	A	G	rs59106543
KNG1	3	186433755	CA	C	rs71708606
KNG1	3	186433783	A	C	rs62292643
KNG1	3	186433822	C	T	rs10440056
KNG1	3	186433872	G	A	rs191943692
KNG1	3	186433893	TTTTTTG	T	
KNG1	3	186433898	TG	T	rs71714968
KNG1	3	186433932	T	G	rs10439970
KNG1	3	186433961	A	G	rs35065127
KNG1	3	186434029	T	C	rs1648717
KNG1	3	186434038	G	T	rs2651642
KNG1	3	186434063	G	A	
KNG1	3	186434108	G	A	rs10440057
KNG1	3	186434180	A	G	rs80355270
KNG1	3	186434188	C	T	rs10439971
KNG1	3	186434262	T	C	rs3821815
KNG1	3	186434405	G	T	rs5029969
KNG1	3	186434417	G	GAGA	rs144977500
KNG1	3	186434475	C	A	rs146533531
KNG1	3	186434478	C	G	rs188096067
KNG1	3	186434491	T	C	rs5029970
KNG1	3	186434819	T	C	rs3806688
KNG1	3	186434887	T	C	rs5029973
KNG1	3	186435077	A	G	rs3806689
KNG1	3	186435370	G	A	rs1050274
KNG1	3	186435561	G	A	rs13072823
KNG1	3	186435583	T	C	rs13095338
KNG1	3	186436191	C	T	rs9825929
KNG1	3	186436217	G	C	rs13317089
KNG1	3	186436268	G	A	
KNG1	3	186436306	G	A	rs1523435
KNG1	3	186436398	A	G	rs1851665
KNG1	3	186436503	T	A	rs1851664
KNG1	3	186436604	T	C	rs1836860
KNG1	3	186436659	C	T	rs113373239
KNG1	3	186436787	GA	G	rs34745532
KNG1	3	186436918	C	T	rs5029975
KNG1	3	186437296	A	G	rs5029976
KNG1	3	186437300	C	T	rs5029977
KNG1	3	186437342	T	C	rs1656908
KNG1	3	186437344	A	C	
KNG1	3	186437353	TA	T	
KNG1	3	186437369	G	T	
KNG1	3	186437375	A	G	
KNG1	3	186437379	G	A	
KNG1	3	186437385	G	A	rs5029979
KNG1	3	186437386	A	C	rs182665559
KNG1	3	186437411	T	G	
KNG1	3	186437450	A	G	
KNG1	3	186437467	G	A	
KNG1	3	186437473	A	G	
KNG1	3	186437504	C	T	
KNG1	3	186437513	A	G	rs13084325
KNG1	3	186437517	CA	C	
KNG1	3	186437532	G	A	
KNG1	3	186437554	A	T	rs9869244
KNG1	3	186437661	G	C	rs2304451
KNG1	3	186437944	T	C	rs5029980
KNG1	3	186438037	T	C	rs2304452
KNG1	3	186438314	G	A	rs5029981
KNG1	3	186438346	C	T	rs2689197
KNG1	3	186438393	G	A	rs2651640
KNG1	3	186438547	G	A	
KNG1	3	186438554	A	G	rs2651641
KNG1	3	186438560	GGTT	G	rs71759015
KNG1	3	186438567	G	T	rs112087464
KNG1	3	186438572	GC	TT	
KNG1	3	186438620	T	C	
KNG1	3	186438639	C	T	
KNG1	3	186438648	G	A	rs5029985
KNG1	3	186438664	T	C	
KNG1	3	186438669	T	C	

Gene	Chromosome	Genome Location	Reference Allele	Alternate Allele	dbSNP v137
KNG1	3	186438672	A	G	rs5029986
KNG1	3	186438686	G	C	
KNG1	3	186438703	T	C	
KNG1	3	186438724	A	G	
KNG1	3	186438728	A	G	rs5029987
KNG1	3	186438729	C	T	rs5029988
KNG1	3	186438739	T	C	
KNG1	3	186438754	G	A	
KNG1	3	186438800	A	G	
KNG1	3	186438813	G	C	
KNG1	3	186438819	C	G	rs5029990
KNG1	3	186438820	T	C	
KNG1	3	186438824	T	C	rs5029991
KNG1	3	186438830	A	G	
KNG1	3	186438836	T	C	rs5029992
KNG1	3	186438868	A	G	rs1656909
KNG1	3	186438935	C	A	rs1624230
KNG1	3	186438993	A	C	rs1656910
KNG1	3	186439173	T	C	rs1621816
KNG1	3	186439295	G	A	rs5029993
KNG1	3	186439373	T	C	rs1656911
KNG1	3	186439529	C	T	rs1648716
KNG1	3	186439618	A	G	rs1648715
KNG1	3	186439723	A	C	rs1656912
KNG1	3	186439840	G	A	rs1829886
KNG1	3	186439906	CAAAAA	C	rs56170982
KNG1	3	186439998	T	C	rs1403694
KNG1	3	186440243	G	A	rs1469859
KNG1	3	186440483	T	C	rs1648714
KNG1	3	186440559	C	T	rs5029999
KNG1	3	186440910	C	CT	rs5030001
KNG1	3	186440986	G	A	rs5030002
KNG1	3	186441028	G	T	rs5030003
KNG1	3	186441252	T	C	rs1656914
KNG1	3	186441444	CT	C	
KNG1	3	186441547	CCAGCCT	C	rs113946377
KNG1	3	186441571	A	G	
KNG1	3	186441579	G	A	rs2378115
KNG1	3	186441585	G	A	
KNG1	3	186441599	C	T	
KNG1	3	186441600	G	A	
KNG1	3	186441603	C	T	
KNG1	3	186441604	G	A	rs5030008
KNG1	3	186441606	A	C	
KNG1	3	186441617	G	A	
KNG1	3	186441631	T	C	
KNG1	3	186441678	G	A	
KNG1	3	186441686	A	G	
KNG1	3	186441689	C	T	
KNG1	3	186441694	C	T	
KNG1	3	186441722	T	C	
KNG1	3	186441732	T	C	
KNG1	3	186441734	T	C	
KNG1	3	186441740	GT	G	
KNG1	3	186441763	T	C	rs5030009
KNG1	3	186441823	A	G	rs1656915
KNG1	3	186442016	A	T	rs1648713
KNG1	3	186442034	A	G	rs5030011
KNG1	3	186442209	C	T	rs1648712
KNG1	3	186442299	C	T	rs1656916
KNG1	3	186442318	T	C	rs1656917
KNG1	3	186442486	CT	C	
KNG1	3	186442544	G	C	rs5030014
KNG1	3	186442612	C	T	rs1656918
KNG1	3	186442705	A	G	rs1656919
KNG1	3	186442707	T	C	rs1648711
KNG1	3	186442747	G	C	rs1656920
KNG1	3	186442798	C	G	rs145330334
KNG1	3	186442833	G	C	rs1656921
KNG1	3	186442851	T	C	rs186482877
KNG1	3	186443018	T	C	rs1656922
KNG1	3	186443250	T	C	rs166479
KNG1	3	186443332	AT	A	
KNG1	3	186443355	A	G	rs191335700
KNG1	3	186443357	A	G	
KNG1	3	186443365	T	G	
KNG1	3	186443415	G	C	

Gene	Chromosome	Genome Location	Reference Allele	Alternate Allele	dbSNP v137
KNG1	3	186443426	T	C	
KNG1	3	186443458	T	C	
KNG1	3	186443482	C	T	rs186831050
KNG1	3	186443490	C	CG	
KNG1	3	186443493	T	C	
KNG1	3	186443499	AT	A	rs34640484
KNG1	3	186443511	A	T	
KNG1	3	186443530	G	A	
KNG1	3	186443617	C	G	rs266725
KNG1	3	186443685	AG	A	rs63499661
KNG1	3	186443707	G	GA	rs59230183
KNG1	3	186443731	AT	A	
KNG1	3	186443756	C	G	rs266724
KNG1	3	186443848	G	T	
KNG1	3	186443850	A	G	
KNG1	3	186443864	C	A	rs28496811
KNG1	3	186444063	CT	C	
KNG1	3	186444077	T	C	
KNG1	3	186444080	T	C	
KNG1	3	186444083	TG	T	
KNG1	3	186444085	T	C	rs28390219
KNG1	3	186444094	GA	AG	
KNG1	3	186444193	C	T	rs5030020
KNG1	3	186444644	G	A	rs5030023
KNG1	3	186444809	A	ATG	rs139519865
KNG1	3	186444831	T	G	rs62294376
KNG1	3	186444844	A	G	rs9882598
KNG1	3	186444893	A	G	rs2304454
KNG1	3	186444903	T	A	rs2304455
KNG1	3	186445052	T	G	rs2304456
KNG1	3	186445186	G	GA	rs3841557
KNG1	3	186445237	A	C	rs5030025
KNG1	3	186445351	T	C	rs5030026
KNG1	3	186445436	C	T	rs4686798
KNG1	3	186445455	T	G	rs28607874
KNG1	3	186445566	A	G	rs5030027
KNG1	3	186445754	C	T	rs5030028
KNG1	3	186446087	T	G	rs5030102
KNG1	3	186446483	T	A	
KNG1	3	186446538	T	C	
KNG1	3	186446550	A	C	
KNG1	3	186446561	TC	T	
KNG1	3	186446741	C	T	
KNG1	3	186446854	A	G	rs185528245
KNG1	3	186447047	A	C	rs266723
KNG1	3	186447484	T	G	
KNG1	3	186447491	T	C	
KNG1	3	186447493	G	C	rs822624
KNG1	3	186447518	A	G	
KNG1	3	186447520	A	G	
KNG1	3	186447522	C	T	
KNG1	3	186447523	G	A	
KNG1	3	186447524	C	T	
KNG1	3	186447525	G	A	rs5030035
KNG1	3	186447530	T	C	
KNG1	3	186447547	C	T	
KNG1	3	186447554	C	T	rs5030036
KNG1	3	186447555	G	A	
KNG1	3	186447564	T	C	
KNG1	3	186447582	T	C	
KNG1	3	186447587	G	A	
KNG1	3	186447610	T	C	
KNG1	3	186447619	GG	AC	
KNG1	3	186447620	G	C	
KNG1	3	186447622	C	T	
KNG1	3	186447623	G	A	
KNG1	3	186447628	G	A	
KNG1	3	186447649	T	C	
KNG1	3	186447651	A	G	
KNG1	3	186447689	A	G	
KNG1	3	186447707	T	C	
KNG1	3	186447942	GT	G	rs34650598
KNG1	3	186448159	T	C	rs5030039
KNG1	3	186448245	T	C	rs13098645
KNG1	3	186448270	T	C	
KNG1	3	186448302	G	A	
KNG1	3	186448305	T	C	

Gene	Chromosome	Genome Location	Reference Allele	Alternate Allele	dbSNP v137
KNG1	3	186448314	A	G	
KNG1	3	186448315	G	C	
KNG1	3	186448316	A	G	
KNG1	3	186448317	C	T	
KNG1	3	186448332	AC	CT	
KNG1	3	186448333	C	T	
KNG1	3	186448360	A	G	
KNG1	3	186448367	T	C	
KNG1	3	186448370	A	G	
KNG1	3	186448377	A	G	
KNG1	3	186448393	T	C	
KNG1	3	186448399	T	C	
KNG1	3	186448401	C	T	
KNG1	3	186448402	G	A	
KNG1	3	186448420	T	G	
KNG1	3	186448423	C	T	rs865880
KNG1	3	186448424	G	A	
KNG1	3	186448428	A	G	
KNG1	3	186448432	G	A	
KNG1	3	186448440	A	G	
KNG1	3	186448443	A	T	
KNG1	3	186448444	CA	C	rs10571187
KNG1	3	186448458	A	T	
KNG1	3	186448460	A	T	
KNG1	3	186448462	A	T	
KNG1	3	186448466	A	T	
KNG1	3	186448468	A	T	
KNG1	3	186448470	A	T	
KNG1	3	186448472	A	T	
KNG1	3	186448473	AT	A	
KNG1	3	186448478	T	C	
KNG1	3	186448480	T	C	
KNG1	3	186448482	T	C	
KNG1	3	186448484	T	C	
KNG1	3	186448492	T	C	
KNG1	3	186448494	T	C	
KNG1	3	186448496	T	C	
KNG1	3	186448528	T	C	rs11929295
KNG1	3	186448595	TG	T	rs5030041
KNG1	3	186448928	T	C	rs5030103
KNG1	3	186448989	C	T	rs1648722
KNG1	3	186449122	AT	GC	
KNG1	3	186449123	T	C	rs1656925
KNG1	3	186449194	A	G	rs1648697
KNG1	3	186449198	T	C	rs5030045
KNG1	3	186449267	A	G	rs5030104
KNG1	3	186449582	G	C	rs1648698
KNG1	3	186449884	C	T	rs1622922
KNG1	3	186449890	G	C	rs1648699
KNG1	3	186449916	T	C	rs1648700
KNG1	3	186450069	T	C	rs1624569
KNG1	3	186450200	A	C	
KNG1	3	186450279	T	C	
KNG1	3	186450761	G	A	rs5030047
KNG1	3	186450765	G	A	
KNG1	3	186450775	A	G	
KNG1	3	186450780	C	T	
KNG1	3	186450784	A	C	
KNG1	3	186450803	A	G	
KNG1	3	186450807	C	T	
KNG1	3	186450811	T	C	
KNG1	3	186450818	CA	TG	
KNG1	3	186450819	A	G	
KNG1	3	186450826	A	G	
KNG1	3	186450833	A	G	
KNG1	3	186450840	T	G	
KNG1	3	186450863	T	C	rs5030049
KNG1	3	186450891	C	T	
KNG1	3	186450895	T	C	rs1656926
KNG1	3	186450896	G	A	
KNG1	3	186450903	T	C	
KNG1	3	186450906	A	G	rs5030105
KNG1	3	186450907	C	T	
KNG1	3	186450908	A	G	
KNG1	3	186450915	G	A	
KNG1	3	186450929	A	G	
KNG1	3	186450950	G	A	

Gene	Chromosome	Genome Location	Reference Allele	Alternate Allele	dbSNP v137
KNG1	3	186450959	T	C	rs62294377
KNG1	3	186450967	A	G	
KNG1	3	186450989	T	C	
KNG1	3	186450991	C	T	
KNG1	3	186450992	G	A	
KNG1	3	186451009	T	TG	
KNG1	3	186451013	A	G	
KNG1	3	186451020	C	T	
KNG1	3	186451028	T	C	
KNG1	3	186451031	CTAAA	C	rs10576660
KNG1	3	186451033	A	C	rs201094402
KNG1	3	186451036	T	A	
KNG1	3	186451040	T	A	
KNG1	3	186451227	T	A	rs148418093
KNG1	3	186451236	T	C	rs4686799
KNG1	3	186451357	A	C	rs5030106
KNG1	3	186451387	C	T	rs5030107
KNG1	3	186451576	A	G	
KNG1	3	186451584	C	T	rs35477316
KNG1	3	186451585	T	C	
KNG1	3	186451587	A	G	
KNG1	3	186451608	C	T	
KNG1	3	186451617	T	C	
KNG1	3	186451618	A	G	
KNG1	3	186451625	T	C	
KNG1	3	186451626	A	G	
KNG1	3	186451652	A	G	
KNG1	3	186451672	A	G	
KNG1	3	186451682	G	A	
KNG1	3	186451699	A	G	
KNG1	3	186451710	C	T	
KNG1	3	186451715	T	C	
KNG1	3	186451741	G	A	rs4686439
KNG1	3	186452415	T	G	rs5030058
KNG1	3	186452455	C	G	rs5030059
KNG1	3	186452490	A	G	rs822366
KNG1	3	186452504	T	G	rs822620
KNG1	3	186452716	A	G	rs822364
KNG1	3	186452720	CT	C	rs5030109
KNG1	3	186452820	A	G	rs710449
KNG1	3	186452885	A	G	rs710448
KNG1	3	186452917	C	T	rs5030060
KNG1	3	186452991	C	G	rs822363
KNG1	3	186453065	A	C	rs822362
KNG1	3	186453129	A	G	
KNG1	3	186453134	T	C	
KNG1	3	186453144	T	C	
KNG1	3	186453178	A	G	
KNG1	3	186453217	A	G	
KNG1	3	186453232	A	C	
KNG1	3	186453233	A	C	
KNG1	3	186453236	A	G	
KNG1	3	186453259	C	T	
KNG1	3	186453263	T	C	
KNG1	3	186453267	T	C	
KNG1	3	186453275	T	C	
KNG1	3	186453276	G	A	
KNG1	3	186453277	G	C	
KNG1	3	186453278	G	A	
KNG1	3	186453279	T	C	
KNG1	3	186453284	A	G	
KNG1	3	186453299	C	T	
KNG1	3	186453322	A	G	
KNG1	3	186453326	A	G	
KNG1	3	186453330	T	C	
KNG1	3	186453339	G	A	
KNG1	3	186453347	CA	C	
KNG1	3	186453355	A	C	
KNG1	3	186453356	A	G	
KNG1	3	186453357	A	T	
KNG1	3	186453362	A	G	
KNG1	3	186453363	T	C	
KNG1	3	186453373	C	A	
KNG1	3	186453385	T	C	
KNG1	3	186453418	A	T	rs4686800
KNG1	3	186453430	GAATT	G	rs138610068
KNG1	3	186453524	G	A	rs112323692

Gene	Chromosome	Genome Location	Reference Allele	Alternate Allele	dbSNP v137
KNG1	3	186453692	A	C	rs5030110
KNG1	3	186454180	A	C	rs5030062
KNG1	3	186454294	A	T	rs5030063
KNG1	3	186454309	A	G	rs148892840
KNG1	3	186454447	T	C	rs5030064
KNG1	3	186454577	T	C	rs5030065
KNG1	3	186454652	CA	C	
KNG1	3	186454827	G	A	rs5030068
KNG1	3	186455394	T	G	rs181235963
KNG1	3	186455414	T	C	
KNG1	3	186455424	T	C	
KNG1	3	186455435	A	T	
KNG1	3	186455478	C	T	rs5030111
KNG1	3	186455546	T	C	rs5030072
KNG1	3	186455567	T	C	
KNG1	3	186455569	G	A	rs5030112
KNG1	3	186455573	C	T	
KNG1	3	186455674	T	C	rs5030113
KNG1	3	186455787	CA	TG	
KNG1	3	186455788	A	G	rs5030073
KNG1	3	186456035	G	A	rs5030074
KNG1	3	186456207	TA	T	rs5030114
KNG1	3	186456358	A	G	
KNG1	3	186456362	A	G	
KNG1	3	186456363	T	C	
KNG1	3	186456373	T	C	
KNG1	3	186456402	G	A	
KNG1	3	186456407	A	G	
KNG1	3	186456410	T	C	
KNG1	3	186456426	G	A	
KNG1	3	186456433	C	T	
KNG1	3	186456434	G	A	
KNG1	3	186456465	A	G	
KNG1	3	186456492	A	G	
KNG1	3	186456495	C	T	
KNG1	3	186456496	G	A	
KNG1	3	186456500	A	T	
KNG1	3	186456503	T	C	
KNG1	3	186456509	A	C	
KNG1	3	186456528	A	G	
KNG1	3	186456549	T	C	
KNG1	3	186456557	T	C	
KNG1	3	186456559	C	T	
KNG1	3	186456560	A	G	
KNG1	3	186456567	G	A	
KNG1	3	186456575	TG	CA	
KNG1	3	186456576	G	A	
KNG1	3	186456591	T	C	
KNG1	3	186456596	C	T	rs35025634
KNG1	3	186456622	C	T	
KNG1	3	186456623	G	A	
KNG1	3	186456630	C	T	
KNG1	3	186456631	A	G	
KNG1	3	186456635	CA	C	
KNG1	3	186456648	T	A	
KNG1	3	186456652	G	A	
KNG1	3	186457332	C	T	rs5030115
KNG1	3	186457585	T	A	
KNG1	3	186457644	T	C	
KNG1	3	186457679	G	A	
KNG1	3	186457689	C	A	
KNG1	3	186457733	G	A	
KNG1	3	186457752	C	G	
KNG1	3	186457798	C	T	
KNG1	3	186457825	C	T	rs184637126
KNG1	3	186457832	T	C	
KNG1	3	186457838	CA	C	
KNG1	3	186457915	G	A	
KNG1	3	186458322	C	T	rs3856930
KNG1	3	186458764	GT	G	
KNG1	3	186458821	C	T	
KNG1	3	186458822	G	A	
KNG1	3	186458825	T	C	
KNG1	3	186458839	G	A	
KNG1	3	186458843	C	T	
KNG1	3	186458844	G	A	
KNG1	3	186458846	C	G	

Gene	Chromosome	Genome Location	Reference Allele	Alternate Allele	dbSNP v137
KNG1	3	186458852	A	G	
KNG1	3	186458869	A	G	
KNG1	3	186458871	G	C	
KNG1	3	186458875	T	C	
KNG1	3	186458886	A	G	
KNG1	3	186458907	T	C	
KNG1	3	186458910	G	A	rs5030081
KNG1	3	186458928	A	G	
KNG1	3	186458929	T	C	
KNG1	3	186458931	T	C	
KNG1	3	186458939	T	C	
KNG1	3	186458944	A	G	
KNG1	3	186458949	A	G	rs5030082
KNG1	3	186458971	T	C	
KNG1	3	186458981	G	C	
KNG1	3	186459003	A	G	
KNG1	3	186459018	A	G	
KNG1	3	186459025	T	C	
KNG1	3	186459028	T	C	
KNG1	3	186459030	G	A	rs142599822
KNG1	3	186459034	T	C	
KNG1	3	186459037	T	C	
KNG1	3	186459052	A	G	
KNG1	3	186459062	T	C	
KNG1	3	186459183	C	T	rs5030083
KNG1	3	186459227	A	G	rs698078
KNG1	3	186459475	C	G	rs5030084
KNG1	3	186459646	G	A	rs5030085
KNG1	3	186459775	C	T	rs5030086
KNG1	3	186459927	T	C	rs710446
KNG1	3	186460110	G	C	rs5030087
KNG1	3	186460222	A	G	rs5030088
KNG1	3	186460877	T	C	rs5030091
KNG1	3	186461091	G	A	rs5030093
KNG1	3	186461158	T	C	rs5030094
KNG1	3	186461181	T	C	rs2062632
KNG1	3	186461216	G	A	rs266760
KNG1	3	186461349	G	C	rs5030095
KNG1	3	186461524	C	T	rs76438938
KNG1	3	186462156	T	C	rs2651639
KNG1	3	186462332	G	T	rs266761
KNG1	3	186462745	G	C	rs16861082
KNG1	3	186462817	TA	T	rs5855112
KNG1	3	186462843	C	T	rs266762
KNG1	3	186463081	C	A	rs905056
KNG1	3	186463084	A	C	rs905057
F11	4	187186111	A	G	rs3756009
F11	4	187186201	C	G	rs4253393
F11	4	187186356	C	G	rs4253811
F11	4	187186372	G	A	rs4253394
F11	4	187186818	TA	T	rs11290631, rs60426781
F11	4	187186879	C	T	rs4253396
F11	4	187187005	G	T	rs3822056
F11	4	187187135	C	G	rs3733403
F11	4	187187495	C	T	
F11	4	187187569	G	A	rs925451
F11	4	187187829	CCA	C	rs4253397
F11	4	187188061	T	C	rs4253398
F11	4	187188094	T	G	rs4253399
F11	4	187188141	G	GAT	rs35709976, rs150319849
F11	4	187188152	A	C	rs3822057
F11	4	187188519	A	C	
F11	4	187189294	T	A	rs4241823
F11	4	187189326	A	G	rs4253403
F11	4	187189805	G	A	rs190524001
F11	4	187190285	T	A	rs925452
F11	4	187190810	A	G	rs4253405
F11	4	187191392	G	T	rs4253406
F11	4	187191787	G	A	rs4241824
F11	4	187191803	GCA	G	rs10535096
F11	4	187191859	T	A	
F11	4	187192481	T	C	rs2036914
F11	4	187193308	A	T	rs4253407
F11	4	187193632	T	C	
F11	4	187193704	C	T	
F11	4	187193858	G	A	rs4253408
F11	4	187194613	CT	C	

Gene	Chromosome	Genome Location	Reference Allele	Alternate Allele	dbSNP v137
F11	4	187194685	C	T	rs4253409
F11	4	187195373	C	T	rs5973
F11	4	187195551	T	A	rs1593
F11	4	187195610	T	C	rs4253410
F11	4	187195796	T	C	rs4253411
F11	4	187196383	A	G	
F11	4	187196442	C	T	
F11	4	187196445	C	G	
F11	4	187196455	A	G	
F11	4	187196460	T	C	
F11	4	187196510	G	A	rs2055916
F11	4	187196511	C	T	
F11	4	187196524	T	C	
F11	4	187196541	T	C	
F11	4	187196553	A	G	
F11	4	187196556	A	G	
F11	4	187196580	C	T	rs4253413
F11	4	187196610	T	C	
F11	4	187196620	A	G	
F11	4	187196635	T	C	
F11	4	187196643	C	T	
F11	4	187196652	T	TA	rs35685802
F11	4	187196853	T	C	rs4253414
F11	4	187197156	A	G	
F11	4	187197736	CT	C	rs33965536
F11	4	187197778	C	T	
F11	4	187197796	G	A	
F11	4	187197807	T	C	
F11	4	187197825	T	C	
F11	4	187197838	C	T	
F11	4	187197843	C	T	
F11	4	187197866	G	A	rs111363884
F11	4	187197878	C	T	
F11	4	187197891	A	C	
F11	4	187197903	G	GC	
F11	4	187197943	A	T	rs4253415
F11	4	187197963	A	G	
F11	4	187197974	T	C	
F11	4	187197978	G	A	
F11	4	187197989	T	C	
F11	4	187197994	C	T	rs4253416
F11	4	187198012	A	G	
F11	4	187198026	G	A	
F11	4	187198424	G	A	rs76605739
F11	4	187198551	T	C	rs188644705
F11	4	187199005	T	C	rs4253417
F11	4	187199110	C	A	rs4253843
F11	4	187199497	A	G	rs4253418
F11	4	187199888	A	G	rs4253419
F11	4	187200260	T	C	
F11	4	187200264	T	C	
F11	4	187200312	T	C	
F11	4	187200314	A	G	
F11	4	187200331	T	A	
F11	4	187200332	A	T	
F11	4	187200334	G	A	
F11	4	187200341	T	G	
F11	4	187200386	G	A	
F11	4	187200392	C	T	
F11	4	187200393	A	G	
F11	4	187200396	T	C	
F11	4	187200405	A	G	
F11	4	187200406	T	G	
F11	4	187200408	T	C	
F11	4	187200429	A	G	
F11	4	187200430	A	G	
F11	4	187200451	A	G	
F11	4	187200460	C	T	
F11	4	187200468	G	A	rs139338150
F11	4	187200485	A	G	
F11	4	187200490	C	T	
F11	4	187200491	G	A	
F11	4	187200492	C	T	
F11	4	187200493	G	A	
F11	4	187200494	A	C	
F11	4	187200514	T	C	
F11	4	187200517	CA	C	

Gene	Chromosome	Genome Location	Reference Allele	Alternate Allele	dbSNP v137
F11	4	187200523	A	C	
F11	4	187200524	A	G	
F11	4	187200530	A	C	
F11	4	187200536	CA	C	rs35203561
F11	4	187200550	A	T	rs56810541
F11	4	187200776	G	A	rs149743449
F11	4	187201001	A	G	rs2241817
F11	4	187201211	A	G	rs5974
F11	4	187201454	G	A	rs281875257
F11	4	187201908	T	A	
F11	4	187202010	CT	C	rs33985758
F11	4	187202125	C	A	rs4253846
F11	4	187202242	C	T	rs141119295
F11	4	187202260	T	C	
F11	4	187202276	T	G	
F11	4	187202794	G	A	rs4253849
F11	4	187202977	G	A	rs3775306
F11	4	187203065	A	G	
F11	4	187203738	C	T	rs76665916
F11	4	187204059	G	C	rs188961841
F11	4	187204069	G	T	
F11	4	187204406	C	CGT	rs58285796, rs10661556
F11	4	187204447	C	T	rs12498667
F11	4	187204525	T	A	rs12503530
F11	4	187204937	A	G	rs4253421
F11	4	187205002	C	G	rs4253422
F11	4	187205033	A	G	rs4253423
F11	4	187205054	C	T	
F11	4	187205301	T	C	rs5970
F11	4	187205426	G	A	rs116667976
F11	4	187205712	T	A	rs4253424
F11	4	187205777	A	ATGTG	rs56310168, rs34783500
F11	4	187205929	T	C	rs4253425
F11	4	187206180	G	A	rs3822058
F11	4	187206249	C	A	rs3756011
F11	4	187206272	G	A	rs190654183
F11	4	187206511	C	T	
F11	4	187207354	C	T	rs2289251
F11	4	187207381	C	T	rs2289252
F11	4	187207535	G	T	rs2289253
F11	4	187208027	G	A	
F11	4	187208141	G	T	
F11	4	187208150	T	C	
F11	4	187208154	C	G	
F11	4	187208158	C	T	rs184914024
F11	4	187208159	G	A	
F11	4	187208166	C	T	
F11	4	187208167	G	A	
F11	4	187208189	A	G	
F11	4	187208190	C	T	
F11	4	187208191	G	A	
F11	4	187208197	G	A	
F11	4	187208206	TT	CA	
F11	4	187208207	T	A	
F11	4	187208214	A	G	
F11	4	187208230	T	C	
F11	4	187208252	C	T	
F11	4	187208265	T	C	
F11	4	187208268	T	C	
F11	4	187208272	T	C	
F11	4	187208273	A	G	
F11	4	187208275	C	G	
F11	4	187208276	CA	C	
F11	4	187208285	C	T	
F11	4	187208289	A	C	
F11	4	187208290	A	G	
F11	4	187208295	AT	A	
F11	4	187208318	C	T	
F11	4	187208319	G	A	rs111736461
F11	4	187208328	T	C	
F11	4	187208330	A	G	
F11	4	187208335	A	G	
F11	4	187208350	A	G	
F11	4	187208374	T	C	
F11	4	187208385	T	C	
F11	4	187208398	A	G	
F11	4	187208415	C	A	

Gene	Chromosome	Genome Location	Reference Allele	Alternate Allele	dbSNP v137
F11	4	187208475	C	T	
F11	4	187208504	A	G	
F11	4	187208507	A	G	
F11	4	187208508	T	C	
F11	4	187208512	T	C	
F11	4	187208520	T	C	
F11	4	187208536	A	G	
F11	4	187208541	TG	CA	
F11	4	187208542	G	A	
F11	4	187208566	T	C	
F11	4	187208599	A	G	
F11	4	187208606	T	C	
F11	4	187208608	T	G	
F11	4	187208610	T	C	
F11	4	187208630	C	T	
F11	4	187208645	A	G	
F11	4	187208648	T	G	
F11	4	187208652	C	T	rs71640037
F11	4	187208653	G	A	
F11	4	187208661	T	C	
F11	4	187208662	T	C	
F11	4	187208669	T	C	
F11	4	187208676	A	T	rs71640038
F11	4	187208686	T	C	
F11	4	187208690	A	T	rs138207591
F11	4	187208692	T	C	rs71640039
F11	4	187208693	G	A	
F11	4	187208695	A	G	
F11	4	187208716	C	T	
F11	4	187208723	A	G	
F11	4	187208741	T	TAA	
F11	4	187208745	AAT	A	
F11	4	187208747	T	A	rs186001900
F11	4	187208968	C	T	rs5975
F11	4	187209067	G	T	
F11	4	187209082	A	G	
F11	4	187209089	C	T	
F11	4	187209097	T	C	
F11	4	187209122	G	A	
F11	4	187209128	A	G	
F11	4	187209140	G	C	
F11	4	187209144	T	C	
F11	4	187209152	A	G	
F11	4	187209159	C	A	
F11	4	187209186	A	G	
F11	4	187209196	C	T	
F11	4	187209210	T	C	
F11	4	187209225	G	A	rs4253427
F11	4	187209227	G	A	rs4253428
F11	4	187209229	CT	C	
F11	4	187209229	C	CT	
F11	4	187209240	T	G	
F11	4	187209243	G	T	
F11	4	187209256	C	T	rs189805890
F11	4	187209273	G	A	
F11	4	187209302	A	G	
F11	4	187209308	T	C	
F11	4	187209312	T	C	
F11	4	187209357	C	T	
F11	4	187209545	T	C	rs186657082
F11	4	187209559	A	G	rs5966
F11	4	187209702	G	T	rs5971
F11	4	187209729	G	A	rs5976
F11	4	187210033	A	G	rs4253429
F11	4	187210064	G	C	rs4253430
F11	4	187210090	G	A	rs4253865
F11	4	187210172	A	G	rs139159299
F11	4	187210247	A	T	rs1062547
F11	4	187210318	TA	T	rs147642874
F11	4	187210334	G	C	
F11	4	187210620	G	A	rs4253431
F11	4	187210836	AT	A	rs57738793, rs67843441,
F11	4	187210870	C	T	rs4253433

References of the Supporting Information

1. Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, Harte RA, Heitner S, Hinrichs AS, Learned K, Lee BT, Li CH, Raney BJ, Rhead B, Rosenbloom KR, Sloan CA, Speir ML, Zweig AS, Haussler D, Kuhn RM, Kent WJ. The UCSC Genome Browser database: 2014 update. Nucleic Acids Res. 2014;42:D764–70.