



UNIVERSITÀ DEGLI STUDI DI TRIESTE

XXVII CICLO DEL DOTTORATO DI RICERCA IN
SCIENZE DELLA RIPRODUZIONE E DELLO SVILUPPO
Indirizzo genetico-molecolare

DEVELOPMENT OF A NEW DIAGNOSTIC ALGORITHM FOR THE STUDY OF DISEASES CHARACTERIZED BY HIGH GENETIC HETEROGENEITY

Elaborazione di un nuovo algoritmo diagnostico per lo studio
di patologie ad elevata eterogeneità genetica

Settore scientifico-disciplinare: MED03

Dottoranda
Elena Nicchia
Elena Nicchia

Coordinatore
Prof. Giuliana Decorti

Supervisore di tesi
Prof. Anna Savoia

ANNO ACCADEMICO 2013/2014

TABLE OF CONTENTS

RIASSUNTO	1
ABSTRACT	3
INTRODUCTION	5
ION TORRENT TECHNOLOGY	5
FANCONI ANEMIA	7
• Clinical features of FA	7
• Genetic of FA	7
• FA pathway	8
• FA diagnosis	9
INHERITED THROMBOCYTOPENIA	10
• Clinical features	10
• Genetic of ITs	10
• Biogenesis of platelets.	11
• ITs diagnosis	12
AIM OF THE STUDY	13
MATERIAL AND METHODS	14
Biological Samples	14
• Fanconi Anemia Samples	14
• Inherited thrombocytopenia samples	14
IPGM Sequencing	15
• Libraries preparation	15
• Enrichment of libraries (Ion One Touch 2 System)	15
• Ion PGM sequencing	16
• Ion PGM sequencing data analyses	16
• Validation of Ion PGM sequencing data on FA DNA samples (Validation I)	17
• Characterization of novel FA alleles (Validation II)	17
• Characterization of novel IT alleles	17
• Detection of copy number variations in FA samples	17
• Detection of copy number variations in IT samples	18
SANGER SEQUENCING	19

SNP ARRAY	19
RESULTS AND DISCUSSION	20
VALIDATION OF ION TORRENT TECHNOLOGY ON FA SAMPLES	20
• Quality control: high coverage for 95% of amplicons	20
• Validation I: high sensitivity and specificity of target NGS sequencing	21
• Validation II: characterization of novel FA alleles	22
• Identification of FA copy number variations:	24
correct assignment of gender	24
• Identification of FA copy number variations: deletions of the entire <i>FANCA</i> and <i>FANCD2</i> genes	25
• Identification of FA copy number variations: characterization of large intragenic deletions of <i>FANCA</i>	26
DEVELOPMENT OF A NEW DIAGNOSTIC PROCESS FOR THE STUDY OF DISEASES WITH HIGH GENETIC HETEROGENEITY	28
APPLICATION OF THE NOVEL DIAGNOSTIC PROCESS TO THE DIAGNOSIS OF INERITHED THROMBOCYTOPENIA	30
• First step: quality control.	30
• Second step: characterization of novel IT alleles.	31
• Third step: CNVs analysis	33
CONCLUSIONS	36
TABLES	37
BIBLIOGRAPHY	62

RIASSUNTO

Le tecnologie di Next Generation Sequencing (NGS) consentono di analizzare contemporaneamente più geni e più campioni in un'unica corsa di sequenziamento. In questo modo potrebbe essere possibile ridurre i tempi e i costi di analisi di tutte quelle patologie caratterizzate da elevata eterogeneità genetica e fenotipica, la cui caratterizzazione risulta essere spesso complessa e dispendiosa.

Al fine di elaborare un nuovo algoritmo diagnostico che consenta la rapida elaborazione di una diagnosi molecolare di tali patologie, abbiamo deciso di validare una tra le più innovative tecnologie NGS attualmente in commercio, la metodica Ion Torrent, su due differenti malattie, entrambe caratterizzate da eterogeneità genetica. l'anemia di Fanconi (FA) e le piastrinopenie ereditarie (IT).

Siccome la FA è una patologia meglio caratterizzata rispetto alle IT, durante la prima fase di questo lavoro di tesi abbiamo analizzato 30 campioni (25 dei quali già precedentemente analizzati con sequenziamento Sanger), di cui 2 wild type e 28 affetti. In seguito all'esclusione dalla nostra analisi di 2 campioni FA a causa di una bassa qualità di sequenziamento, abbiamo determinato la sensibilità (95%) e la specificità (100%) della nuova metodica confrontando i dati di sequenziamento Ion Torrent e quelli Sanger a nostra disposizione. Inoltre, utilizzando i dati di copertura della sequenza, abbiamo messo a punto un'analisi statistica volta all'identificazione delle Copy Number Variation (CNV), confermando le delezioni a carico del gene *FANCA* presenti in 5 pazienti. Abbiamo quindi caratterizzato 25 dei 26 pazienti analizzati, identificando inoltre 2 casi con mutazioni nei rari gruppi di complementazione *FANCF* e *FANCL* e 10 mutazioni in loci differenti dai geni causativi. Poiché non escludiamo la possibilità che un nuovo gene possa essere coinvolto nella patologia, riteniamo che l'unico paziente ancora privo di diagnosi molecolare possa essere un buon candidato per lo studio dell'esoma. Infine, avvalendoci dei buoni risultati ottenuti, abbiamo elaborato un nuovo processo diagnostico con il quale identificare in modo semplice e rapido sia le mutazioni sia le CNV a carico dei 16 geni coinvolti nella FA.

Nella seconda parte del nostro studio, abbiamo verificato se l'applicazione di tale algoritmo possa essere estesa anche ad altre patologie ad elevata eterogeneità genetica. Per questo motivo abbiamo analizzato 21 campioni affetti da piastrinopenie ereditarie, già precedentemente analizzati mediante sequenziamento Sanger. Grazie all'algoritmo proposto abbiamo potuto selezionare tra le 2225 varianti identificate le 75 (56 differenti) che sono risultate essere potenzialmente patogenetiche in base alla loro frequenza nella popolazione

(MAF<0.01), alla loro presenza nei database di mutazione e all'analisi bioinformatica di patogenicità. Trenta (27 differenti) di queste varianti sono state confermate mediante sequenziamento Sanger, di cui in particolare 14 (12 differenti) presenti in geni diversi da quelli causativi. Alla luce di questo dato si rendono necessari studi funzionali su tali varianti al fine di comprendere i meccanismi molecolari alla base delle piastrinopenie ereditarie. Infine, utilizzando l'algoritmo proposto, è stato possibile confermare la diagnosi molecolare in 17 dei 21 pazienti IT, compresi i 2 affetti da trombocitopenia con assenza del radio (TAR) e portatori di una delezione sul cromosoma 1q21.1. I restanti 4 alleli mutati non sono stati identificati a causa di una bassa copertura di sequenziamento.

In conclusione, in base ai dati raccolti sui campioni affetti da FA e IT, possiamo affermare che la tecnologia di sequenziamento Ion Torrent e l'algoritmo diagnostico da noi proposto sono degli strumenti utili per ottenere una diagnosi molecolare completa, veloce ed economica.

ABSTRACT

Next Generation Sequencing (NGS) technologies, such the Ion Torrent platform, could allow to simplify the diagnostic process of diseases characterized by an high genetic and phenotypic heterogeneity, because of the possibility to sequence simultaneously more genes and more patients in a single sequencing run. In order to develop a new diagnostic algorithm for rapid molecular diagnosis of these disorders, we have applied the Ion Torrent technology on two different genetically heterogeneous diseases, Fanconi anemia (FA) and inherited thrombocytopenias (IT). Since FA is a disorder better characterized than ITs, we first validated the Ion torrent technology on 30 samples (2 wild type and 28 FA), 25 of which were already analyzed with Sanger sequencing. Because of their low sequencing quality, we have excluded from this type of analysis 2 of the 28 FA samples. Then, comparing Ion Torrent and Sanger sequencing data, we have evaluated the sensitivity (95%) and the specificity (100%) of Ion Torrent technology. Moreover, in order to detect copy number variations (CNVs) in FA genes, we have improved a statistical analysis based on coverage sequencing data, confirming the presence of large intragenic deletions on *FANCA* in 5 patients. In summary we have characterized 25 of the 26 FA patients analyzed, identifying also 4 mutant alleles in the rare complementation group *FANCL* and *FANCF* and 10 mutations in loci different from genes causing the disease. Since we cannot exclude that new genes are involved in FA, the only patient without any mutation identified is suitable for whole exome analysis. Taking advantage from these good sequencing data, we have developed a diagnostic algorithm that combines the identification of both point mutations and CNVs. In order to verify if this new diagnostic process could be applied also to other genetically heterogeneous diseases, we have analyzed 21 IT patients, already characterized by Sanger sequencing. Among the 2225 variants identified by Ion torrent technology, using this new approach, we have select those (N=75, 56 different) potentially pathogenetic because of their frequency (MAF<0.01), or of their presence in IT mutation database o because of bioinformatics analysis. Thirty of these variants were confirmed by Sanger sequencing, 14 (12 different) of which localized in loci different from the gene causing the disease. It would be interesting to carry out functional studies on these additional variants to unravel the molecular basis of ITs. In summary we were able to characterized 17 of the 21 IT patients, including 2 patients with deletions in *RBM8A* (Thrombocytopenia and Absent Radii syndrome, TAR). The remaining 4 mutant alleles were not detected because of a low sequencing coverage.

In conclusion, according to our data, we can consider the Ion Torrent technology and in particular the diagnostic algorithm proposed in our study, as a feasible approaches for the study of diseases characterized by high genetic and phenotypic heterogeneity.

INTRODUCTION

The high genetic heterogeneity that characterizes human diseases has important implications in gene discovery and in development of molecular diagnostic protocols (McClellan and King, 2010). Because of the high cost of Sanger sequencing, up to now a lot of different methodologies, such as genome-wide association studies or SNP-array analysis, were performed to discover unknown loci associated with the disease. However the development of Next Generation Sequencing technologies (NGS) has changed the comprehensiveness of human genetic analysis and significantly reduced the costs associated with sequencing (Johnsen et al, 2013).

Before sequencing, all currently available platforms require different steps of DNA sample processing, which consists of generation of a target library (Buermans and Dunnen, 2014). Among the different sequencing platforms, one of the most innovative is the Ion Torrent technology because is unique in detecting the slight change in pH that takes place during the DNA synthesis (Rothberg et al, 2011), avoiding the need of optical system and reducing the time and the cost of sequencing (Loman et al., 2012).

ION TORRENT TECHNOLOGY

Ion Torrent technology is based on chips with a highly dense microwell array, below which there is a layer of Ion-Sensitive Field-Effect Transistor (ISFET), sensor elements that are able to convert a chemical signal (pH changes) into a digital one (Rothberg et al, 2011). Each microwell acts as an individual DNA polymerization reaction chamber, containing a bead with million copies of a single DNA fragment to sequence. The sequencing chemistry is a simple process: each fragment is single-stranded and bound with primer and DNA polymerase. Sequencing reaction begins when one of the four nucleotide (dNTP) flows into the well. To distinguish the different dNTPs, the single bases are added in a predefined flow order and when one of it is incorporated into the nascent strand, a proton (H^+) is released. The release of H^+ is detected as a change in the pH within the sensor wells, that is converted to a voltage change signal. This electric input is then digitized by off-chip electronics. Finally, a signal-processing software converts the full series of raw-data of a single well into a sequencing read (Figure 1A). Since the technology is based on pH change detection, Ion Torrent platform is the only commercially available chemistry that uses entirely natural nucleotides. Moreover due to the lack of the time consuming imaging, a sequencing run can be completed within 4 hours (Rothberg et al, 2011; Meriman et al., 2012; Buermans and Dunnen, 2014).

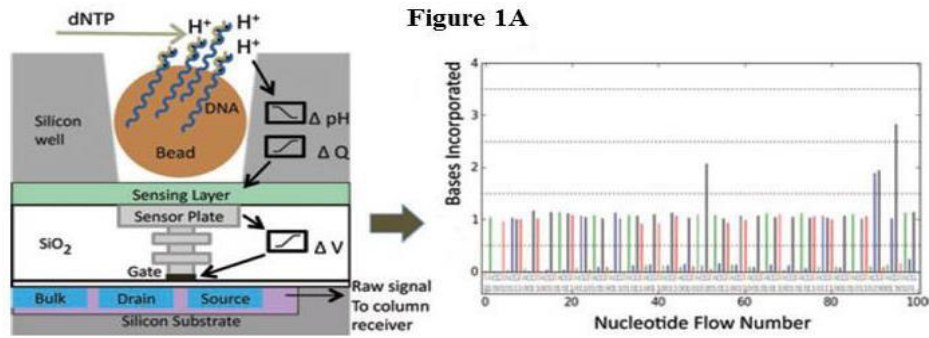


Figure 1A

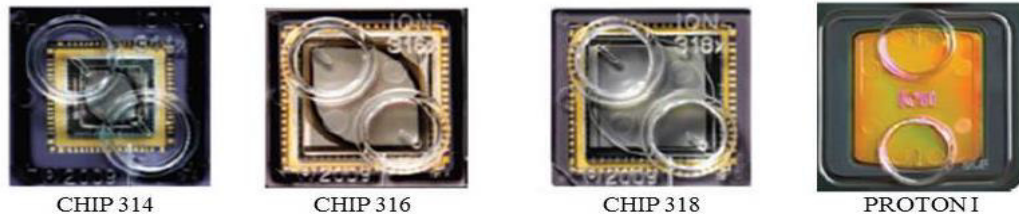


Figure 1B

Figure 1: Ion Torrent sequencing technology. A) A simplified drawing of a well, a bead containing DNA template and the underlying sensor and electronics. Protons (H⁺) are released when dNTPs are incorporated on the growing DNA strands, changing the pH of the well (ΔpH). The ΔpH is converted in a voltage change (ΔV) that is digitized by the off chip electronics. The full digital signs were then processing in a sequencing read. B) Here are shown the four chip commercially available and the chip name (Meriman et al, 2012).

Nowadays, two different Ion Torrent sequencing systems are available, the Ion Personal Genome Machine (IPGM) and the Ion Proton sequencer. The two platforms use chips that differ each other for the level of throughput. Indeed, for 200bp chemistry, the IPGM chips show a capacity ranging from 50Mb of chip 314 to 1Gb of chip 318, allowing the sequencing of small target genomic region. Instead, Proton I has an output of 10Gb for exome, transcriptome and larger genome sequencing (Figure 1B). The sequencing read length is usually of 200bp, even if it is now possible to sequence fragments of 400bp with a good accuracy. It has been shown that for a 200bp chemistry the error rate for substitutions is ~0.1%. However, this percentage increases in the homopolymeric regions, being of 3.5% in 5-mer homo-polymer (Buermans and Dunnen, 2014).

Therefore, the Ion Torrent technology is feasible for molecular diagnosis of rare genetically heterogeneous diseases characterized by relatively small targets, such as Fanconi anemia (FA) and inherited thrombocytopenias (IT), both subject of our research.

FANCONI ANEMIA

Fanconi Anemia (FA) is a rare genetic bone marrow failure disorder with an incidence of 1/350000 borns. The heterozygote frequency is estimated at 1/350 in Europe and the United States, even if founder effects have been demonstrated in different ethnic groups, such as Ashkenazi Jews and Afrikaner. In these populations the carrier frequency is of 1/89 and 1/83 borns, respectively (Tischkowitz and Hodgson, 2003).

• Clinical features of FA

FA is characterized by a high phenotypic heterogeneity. Indeed in the first decade of life FA patients often, but not always, show a combination of various congenital abnormalities, such as short stature; typical facies and microphthalmia; thumb and radius malformations; skin hyperpigmentation; neuronal, cardiac, renal and genitourinary malformations. (Tischkowitz and Hodgson, 2003). FA phenotype can overlap with that of other diseases, like VACTERL syndrome, Thrombocytopenia and Absent Radii syndrome (TAR), Seckel syndrome and Diamond-Blackfan syndrome (Auerbach, 2009). Moreover, it is possible that individuals of the same family express different phenotype, even if they have the same gene mutations.

Notably, approximately one-third of FA patients do not manifest major congenital malformations. In these patients, FA diagnosis is generally made only after the manifestation of hematologic dysfunctions, which generally occurs at a median age of 7 years. FA patients have also an increased risk of develop hematological and solid tumors in early age (Tischkowitz and Hodgson, 2003).

• Genetic of FA

At present, 16 genes (FANC genes) are known to be involved in the disease (*FANCA*, *FANCB*, *FANCC*, *FANCD1/BRCA2*, *FANCD2*; *FANCE*; *FANCF*; *FANCG*; *FANCI*; *FANCIJ/BRIP1*, *FANCL*, *FANCM*, *FANCN/PALB2*, *FANCO/RAD51C*, *FANCP/SLX4*, *FANCO/ERCC4*). Mutations in these genes are inherited in autosomal manner except for those in *FANCB* gene, which is located on chromosome X. The most frequent genes involved in this pathology are *FANCA*, *FANCG* and *FANCC*, which together account for almost 90% of the cases. Furthermore, there is a wide spectrum of private mutations, including large intragenic deletions that characterized almost 20% of the *FANCA* alleles (Fanconi Database, The Rockefeller University, <http://www.rockefeller.edu/fanconi/>; Morgan et al, 1999).

- **FA pathway**

FANCD2 and FANCI proteins are involved in a common DNA repair signaling pathway, which closely cooperates with other DNA repair proteins for resolving DNA interstrand cross-links during replication (Moldovan and D'Andrea, 2009). This pathway is composed of three different complexes. Complex I (FANCA, FANCB, FANCC, FANCE, FANCF, FANCG, FANCL and FANCM, also known as the FA core complex) functions as a nuclear E3 ubiquitin ligase. During the DNA repair process, the FA core complex, together with the FA associated proteins FAAP24 and FAAP100, promotes the monoubiquitination of the FANCD2/FANCI heterodimer (complex II). The FANCD2 ubiquitination is indispensable for the DNA repair process and the activation of FA pathway. On the contrary FANCI ubiquitination is not essential, although it may enhance DNA repair process. Mono-ubiquitinated complex II interacts with the third complex, composed by FANCD1/BRCA2, FANCF/BRIP1, FANCG/PALB2, FANCL/RAD51C, FANCM/SLX4 proteins, promoting the DNA repair process via homologous recombination (Figure 2)

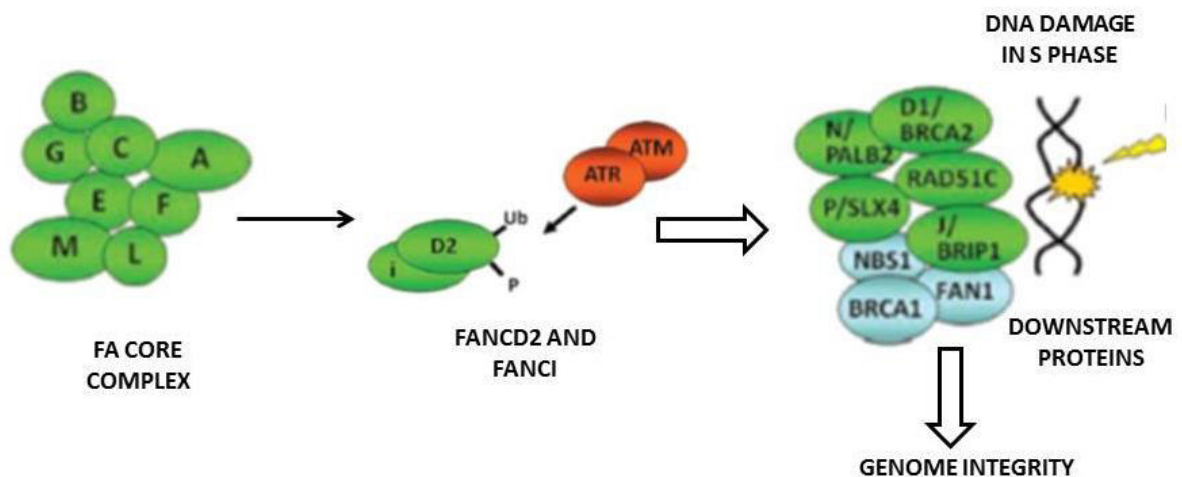


Figure 2: FA pathway. A schematic diagram showing the three FA complexes and their role in the DNA repair process. FA proteins are indicated in green. (Adapted from Soulier, 2011)

- **FA diagnosis**

The genetic heterogeneity together with the numerous mutations identified in the FA genes makes the molecular diagnosis of FA complex and time consuming. As indicated in figure 3, it is a tiered process starting from clinical suspicion that is confirmed at the cellular level testing sensitivity of patient's cells to DNA interstrand cross-linking (ICL) agents, such as diepoxybutane (DEB test). If the DEB test is positive, identification of mutations is fundamental for a correct management of patients in terms of genetic counseling, carrier testing, and prenatal diagnosis (Auerbach, 1993). An inconclusive DEB test might occur in patients with hematopoietic mosaicism arising from spontaneous genetic events, such as back mutations, leading to reversion of the FA cellular phenotype (Waisfisz et al., 1999; Gross, 2002). Without any knowledge on candidate genes, molecular genetic testing is usually carried out starting from *FANCA*, which is mutated in 60-80% of FA families (Castella, 2011; De Rocco, 2014). The analysis consists of Sanger sequencing of 43 coding exons combined with MLPA (Multiple Ligation-dependent Probe Amplification). If no *FANCA* mutation is identified, the screening is extended to the other FA genes. Alternatively, the candidate gene can be identified by complementation analysis or restricted to the components of the FA pathway playing a role up or down of FANCD2 on the basis of its monoubiquitination status (Kottemann et al, 2013; Bogliolo et al, 2013). Therefore, FA diagnosis would greatly benefit from application of NGS, which could allow to avoid the use of complementation and western blot analyses, reducing both time and cost of the process.

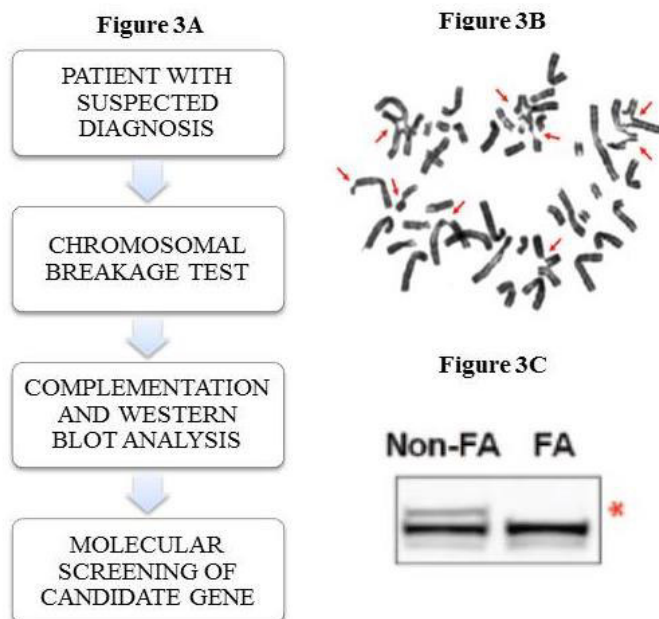


Figure 3: FA diagnosis. A) Schematic flowchart of FA diagnosis. B) FA cells are hypersensitive to ICL agents that cause chromosomal aberrations (red arrows) that are detected by the DEB test. C) Western Blot analysis reveals FANCD2 monoubiquitination defects, indicating in which FA complex mutations have occurred.

INHERITED THROMBOCYTOPENIA

Inherited Thrombocytopenias (IT) are an heterogeneous group of different diseases, all characterized by a platelet counts lower than $150 \times 10^9/L$. ITs are considered exceedingly rare, but their prevalence is unknown because no population-based study has been performed. However, a recent work carried-out by the group of Balduini in 2012 has shown that, at least in Italy the most frequent disorders are Bernard Soulier Syndrome (BSS) and MYH9-related diseases (MYH9-RD).

- **Clinical features**

The most important feature of ITs is increased bleeding tendency, which consists of mucocutaneous haemorrhages, nose bleeds, menorrhagia and gastrointestinal bleeding. Patients with platelet counts lower than $20-30 \times 10^9/L$ suffer from spontaneous often life-threatening haemorrhages since birth; in most cases the degree of thrombocytopenia is mild and bleeding is occasional and related to trauma or surgery. However, some patients, such those affected by BSS, show a bleeding tendency disproportionate to the degree of thrombocytopenia because of defects associated with platelet function (Balduini et al, 2012). Furthermore, approximately half of the ITs are syndromic disorders, associated with additional defects that could affect nearly all organs and system. For example patients with Wiskott-Aldrich syndrome (WAS) are at risk of infections and those with TAR or Congenital Thrombocytopenia with Radio-Ulnar synostosis (CTRUS) have skeletal defects. In other cases, such as in MYH9-RD, affected individuals can develop glomerulonephritis, deafness and cataracts during life (Balduini et al, 2013).

- **Genetic of ITs**

Knowledge about the genetic causes of ITs has greatly improved in the past decade and to date 22 are the genes involved in these disorders (*ABCG5, ABCG8, ACTN1, ANKRD26, CYCS, FLII, FLNA, GATA1, GPIBA, GPIBB, GP9, HOXA11, ITGA2B, ITGB3, MHY9, MPL, NBEAL2, RBM8A, RUNX1, TUBB1, VWF, WAS*). The pattern of inheritance depends on the type of disorders and often is not easy to identify because of *de novo* mutations. Finally, transmission in some disorders, as BSS, may be both dominant and recessive, causing a more or less severe phenotype (Balduini et al, 2012; Balduini and Savoia, 2012; Kunishima et al, 2013).

- **Biogenesis of platelets.**

The IT genes encode for proteins that are involved in the production of platelets from megakaryocytes (MK). These are myeloid cells that reside primarily in the bone marrow. In early development, the megakaryocytogenesis occurs within the yolk sac and fetal liver and it consists of hematopoietic stem cells (HSCs) differentiation and maturation into MKs (Patel et al, 2005). The differentiation of MKs in bone marrow is mediated by thrombopoietin (THPO), which binds to the c-MPL receptor on the surface of the cell and activates a series of signaling pathways. At the end of maturation, MKs result in polyploid cells with multilobed nucleus and DNA content up to 128N. Moreover, during maturation, MKs increase in size, become full of platelet specific granules, expand their cytoplasmic content of cytoskeletal proteins and develop a highly tortuous invaginated membrane system (IMS), a membrane reservoir for platelets' production. During this phase, MKs migrate from the osteoblastic to the vascular niche of the bone marrow, where they start extending proplatelets that protrude into the vascular lumen. Finally, platelets are released into the blood stream as preplatelets for their conversion into platelets in the vascular vessels (Machlus and Italiano, 2013).

The complex process of platelets' biogenesis involves different transcriptional factors, such as GATA1, RUNX1 and FLI1, and several cytoskeletal proteins, such as ACTN1, TUBB1, FLNA and MYH9. For most of the IT genes, their functional role remains to be clarified.

- **ITs diagnosis**

The recognition of genetic origin of ITs is often hampered by several difficulties. First of all, the clinical recognition of ITs is often delayed because an overlapping between ITs and acquired forms of thrombocytopenia. Thus, to distinguish these forms, it is important to carefully analyze the family medical history of probands. Moreover, ITs could be often misdiagnosed because of ambiguous platelet count that results both from the combination of ethnic-, age- and sex variables and from errors in count measurement by cell counters, especially for patients with very large platelets (Balduini et al., 2013).

Once an IT is diagnosed, to avoid the problem of variable expressivity of clinical features, the diagnostic process usually take advantage of the algorithm proposed by the Italian Platelet Study Group in 2003 (Figure 4; Balduini et al, 2003). The first step consists of discrimination of the syndromic forms from the nonsyndromic ones. The latter are then classified according to the platelet size. Finally, to find the potentially candidate gene, molecular tests, such as blood film and bone marrow evaluation, ristocetin response analysis and immunofluorescence test, are performed.

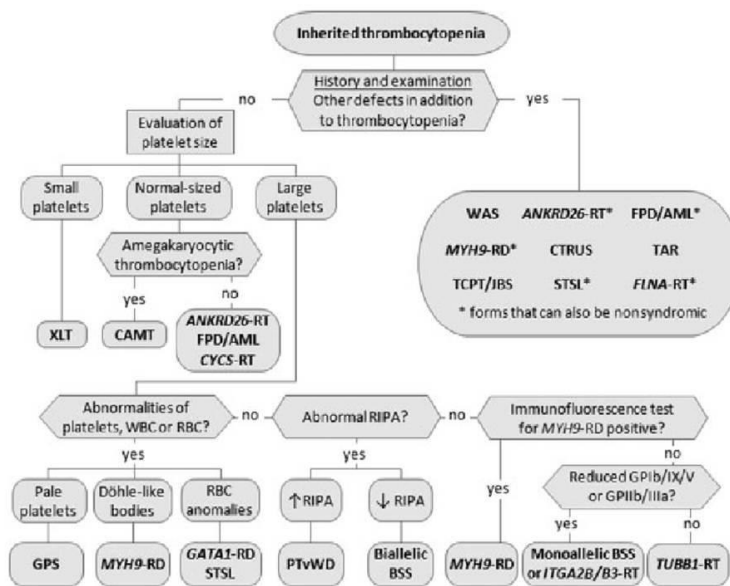


Figure 4: Diagnostic algorithm for ITs: medical history and physical examination are sufficient in most cases for suspecting syndromic forms. For the nonsyndromic forms, further investigations of the platelets' size and laboratory tests are necessary (Balduini et al, 2013).

Despite of the application of diagnostic algorithm, about the 50% of the IT patients remain without a molecular confirm of clinical diagnosis (Balduini et al, 2012). For this reason, ITs are a great challenge for the application of the NGS technologies.

AIM OF THE STUDY

Thanks to the development of NGS technologies, such the Ion Torrent platform, is now possible to sequence simultaneously more genes and more patients in a single sequencing run. This possibility could simplify the diagnostic process of diseases characterized by an high genetic and phenotypical heterogeneity. For this reason, we have applied the Ion Torrent technology on two different genetically heterogeneous diseases, Fanconi anemia (FA) and inherited thrombocytopenias (IT), in order to develop a new diagnostic algorithm for rapid molecular diagnosis.

MATERIAL AND METHODS

BIOLOGICAL SAMPLES

- **Fanconi Anemia Samples**

Twenty-eight DNA samples from FA probands with positive DEB test and two from healthy controls (14 males and 16 females) were analyzed in this study (Table 1). DNA samples were isolated from peripheral blood (n=25) or from lymphoblastoid cell lines (n=5) using a High Pure PCR Template Preparation Kit (Roche, Indianapolis, USA). In 25 samples, at least one of the most frequently mutated genes (*FANCA*, *FANCC* or *FANCG*) was previously analyzed by Sanger sequencing (De Rocco et al, 2014). Of the 28 affected individuals, 15 had both FA alleles characterized. In three cases, including two with hematopoietic mosaicism (Waisfisz et al, 1999), only one FA allele was identified. In this cohort, 23 FA alleles were still missing. To find all potential Copy Number Variants on *FANCA* gene, all samples were previously tested using the Salsa MLPA (multiplex ligation-dependent probe amplification) kit with probe mix P031/P032 according to the manufacture protocol (MRC-Holland, Amsterdam, the Netherlands). Reactions were run on ABI PRISM 3130 Genetic Analyzer (Applied Biosystems, Foster City, USA), and analyzed by Gene Mapper version 3.2 (Applied Biosystems, Foster City, USA) and MRC-Holland Coffalyzer software. (MRC-Holland, Netherlands)

- **Inherited thrombocytopenia samples**

In this study we have analyzed twenty-one DNA samples (12 males and 9 females) with different diagnostic suspicion of Inherited Thrombocytopenia (Table 2). All of these DNA samples were isolated from peripheral blood using a High Pure PCR Template Preparation Kit (Roche Indianapolis, USA). In 20 samples at least one of the IT genes was previously analyzed with Sanger sequencing and all the subjects received a molecular confirm of the diagnosis. Only one sample was never screened before.

All the subjects or their legal guardians gave written informed consent to the investigation, according to the Declaration of Helsinki.

IPGM SEQUENCING

The IPGM sequencing was carried out according to manufacturer's protocols (Life Technologies). Briefly, to sequence the coding exons and their flanking regions of the 16 FA genes and of the 22 IT genes, we designed two different panel of primers using the Ion Ampliseq Designer software (Life Technologies, <https://www.ampliseq.com/browse.action>). Both panel were divided in two pools consisting in 348 and 345 amplicons each FA pools and in 327 and 323 amplicons in each IT pools.

- **Libraries preparation**

To construct the libraries we made two multiplex PCR, one for each primer pool, using Ion Ampliseq Library kit 2.0 LV, as indicated by manufacturer's protocols. Then to distinguish different sample on the same sequencing chip, each library was partially digest and ligated to Ion Xpress Barcode Adapters.

Libraries were cleaned using Agencourt AMPure XP reagent according to manufacturer protocols. Each libraries were then quantified with real-time PCR using Ion Library Quantitation Kit in accordance with protocol. Real time PCR was carried out on a 7900 HT System

- **Enrichment of libraries (Ion One Touch 2 System)**

To pool together all the libraries in a single sample, we have calculated template dilution factor for a final concentration of 10pM per library. Then an enrichment reaction of the template was carried out using Ion One Touch 2 system (Ion PGM template OT2 200 kit), which consist of two different instruments, the Ion OneTouch Instrument and the Ion OneTouch ES. In the first one all the amplicons are bound to the Ion Sphere Particles (ISPs), magnetic beads on which the emulsion-PCR reactions takes place. Primers used in this step are the same for all the amplicons because complementary to the adapter sequence. The reaction products are ISPs presenting millions of copies of a single amplicon on their surface.

The selection of the template positive ISPs is performed on the Ion OneTouch ES.

The percentage Template ISPs is determined using the Qubit 2.0 fluorometer. As indicate in the manufacturer's protocol, the optimal amount of library corresponds to the library dilution point that gives Percent Templated ISPs between 10–30%.

- **Ion PGM sequencing**

Libraries combined in a single samples and enriched with Ion One Touch 2 system were sequenced using the Ion PGM sequencer (Ion PGM sequencing 200 kit v.2), according to manufacturer's protocol.

The number of combined libraries that can be accommodated in a single sequencing run depends on the size of the chip, the ability to reliably quantify and combine barcoded libraries, and the coverage required. In order to assure high coverage we loaded into 316 chips 12 FA libraries (6 DNA samples) or 14 IT libraries (7 DNA samples) into 316 chips.

- **Ion PGM sequencing data analyses**

Sequencing data were analyzed using Ion Torrent Suite software (v 4.0). Using the plug-in Coverage Analysis (TSCA v 4.0), we evaluated the quality of data sequencing. Data were aligned with hg19 human genomic sequence using the plug-in Variant Caller (TSVC v.4.0). Functional annotations of all the sequence variants were performed using the w-annovar software (<http://wannovar.usc.edu/>). Libraries alignment was visualized using Integrative Genomics Viewer (IGV) software (Thorvaldsdóttir et al, 2013).

Moreover the pathogenic effects of all those variants with MAF<1% was evaluated using four pathogenicity prediction programs, such as PolyPhen-2 (<http://genetics.bwh.harvard.edu/pph2/>), Mutation Taster (<http://www.mutationtaster.org/>), Mutation Assessor (<http://mutationassessor.org/>), and SIFT (<http://sift.jcvi.org>). Since the different software use different prediction scales, we converted each output assigning a value from 0 to 2. Specifically, we assigned values 0, 1, and 2 to the PolyPhen-2 prediction of "benign", "possibly damaging", and "probably damaging", respectively. Values of 0 and 2 were attributed to "polymorphism" and "disease causing" output of Mutation Taster, respectively. To the Mutation Assessor predictions of "neutral", "low", and "medium" we assigned values of 0, 1, and 2, respectively. Finally, when the SIFT output was "tolerated" the value was 0 and for "not tolerated" was 2. The pathogenicity score was obtained by summing the values attributed to the output from the single software. Scores ≥ 5 were considered pathogenic. The in silico analyses for detection of potential splicing mutations were carried out using Human Splicing Finder Version 2.4.1 (<http://www.umd.be/HSF/>).

- **Validation of Ion PGM sequencing data on FA DNA samples (Validation I)**

To assess the reliability of Ion PGM sequencing technology, we evaluated the sensitivity [$TP/(TP + FN)$] and the specificity [$TN/(TN + FP)$], the positive [$PPV=TP/(TP+FP)$] and the negative [$NPV=TN/(TN+FN)$] predictive values of the technology, comparing Ion PGM sequencing data with Sanger data on *FANCA*, *FANCC* and *FANCG* genes at our disposal on the 25 FA samples previously analyzed (Table 1).

- **Characterization of novel FA alleles (Validation II)**

All those exons that present variants with $MAF < 1\%$ were analyzed by Sanger sequencing.

- **Characterization of novel IT alleles**

Among the variants called by TSCV on the 21 IT samples, we selected those with $MAF < 0.01$ and considered pathogenetic either because present in one of the four public mutation database (HGMD, LOVD, OMIM, Clinical Variation), or because having a pathogenicity score ≥ 5 . All these variant were validated with Sanger sequencing.

- **Detection of copy number variations in FA samples**

Statistical analyses for detection of copy number variations (CNV) were performed on coverage data supplied by TSCA. These analyses implied two normalizations for each sample. The first was an intrasample normalization, which was calculated by dividing the total number of reads of each amplicon by the total reads average obtained from the same library. Since almost 20% of the *FANCA* alleles are large intragenic deletions (Morgan et al, 1999) and *FANCB* is localized on chromosome X, we excluded the *FANCA* and *FANCB* amplicons from the calculation of the total reads of each library. The second was an intersample normalization, which was determined by dividing the intrasample normalisation of each amplicon by the average of all the intrasample normalisation of this amplicon from control samples.

As control samples in the intersample normalization of *FANCA* and *FANCB* amplicons, we used data from 21 samples without *FANCA* deletions (from MLPA analysis) and from 15 females, respectively. For the intrasample normalization of amplicons of the other FA genes, we use all the samples ($n=28$) except for P3 and P4, which had the lowest uniformity amplicon coverage. All those amplicons that did not achieve the threshold of 30X in more than three samples were excluded from the analysis. We considered intersample normalisation ratios lower than 0.7 indicative of heterozygous deletions and ratios higher than 1.3 suggestive of

duplications. Graphs and statistical analyses were performed using the software Minitab® Statistical Software. Median and interquartile range (IQR) were used to describe non-normal distribution data, whose comparison was performed by the Mann-Whitney test.

- **Detection of copy number variations in IT samples**

Also for IT samples the same statistical analysis based on two normalization steps of coverage sequencing data was improved to detect CNV. In this case since *WAS*, *GATA1* and *FLNA* genes are localized on chromosome X and large deletions on *RBM8A* gene are responsible of TAR syndrome (Albers et al, 2012), we excluded the *WAS*, *GATA1*, *FLNA* and *RBM8A* amplicons from the calculation of the total reads of each library in the first normalization step. Since 4 samples were excluded from control samples of CNVs because of low uniformity, data from 16 samples without *RBM8A* deletions and from 10 males were used as control samples in the intersample normalization of *RBM8A* and *WAS*, *GATA1* and *FLNA* amplicons, 17 samples were used for the intrasample normalization of amplicons of the other IT genes. All those amplicons that did not achieve the threshold of 30X in more than three samples were excluded from the analysis. The same parameters used to indicate the presence of CNVs in FA samples were used for IT samples.

Validation analysis of CNVs was carried out with SNP array, except for *FANCA* gene, which is analyzed with MLPA analysis, and for all those genes localized on chromosome X (*FANCB*, *WAS*, *GATA1* and *FLNA*).

SANGER SEQUENCING

Bidirectional Sanger sequencing was performed on FA and IT exons carrying variants with MAF<1%. PCR reaction was performed using Kapa2G Fast Hot Start mastermix (KapaBiosystem) as indicated by manufacturer's protocol. Quality control of PCR products was performed with electrophoresis on agarose gel (1% Tris Borate EDTA Buffer).

Following sequencing steps were performed by IGA technology services (Udine, Italy).

Sequencing data analysis was done using DNASTAR Lasergene Seqman Software.

SNP ARRAY

SNP array analysis was performed using the Human OmniExpress-12 Bead Chip (Illumina Inc., San Diego, CA) according to Illumina's Infinium HD Assay protocol. Normalization of raw image intensity data, genotype clustering and individual sample genotype calls were performed using Illumina's GenomeStudio software v2011.1 (CNV partition 3.2.0). The CNV calls were determined with generalized genotyping methods implemented in the Penn CNV program.

RESULTS AND DISCUSSION

Since FA is a high genetic heterogeneity disease that is better characterized than ITs, we first validated the IPGM technology on 2 wild type and 28 FA samples. Of note, among the 56 mutant alleles, 33 were previously characterized by Sanger sequencing.

VALIDATION OF ION TORRENT TECHNOLOGY ON FA SAMPLES

- **Quality control: high coverage for 95% of amplicons**

The 16 FA genes represent a target of 74.2 kb split into 693 amplicons ranging in size from 125 to 225bp. The amplicons cover 96% of the coding exon sequences with approximately 30bp of their flanking intronic sequences. The regions excluded by the design (4%) are variably dispersed through all the genes, varying from 0.03% (*FANCO*) to 10% (*FANCE*) of the targeted sequences. All samples satisfied the quality control parameters except for P3, which was excluded from further analysis because the uniformity of amplicon coverage was below 50% (Table 3).

Coverage analysis found that 35 of the 693 amplicons did not reach the threshold of 30X in at least one of the 29 samples (Table 4). Five amplicons were not covered in at least 10 samples, suggesting that “constitutive” features of DNA could be responsible for the low sequencing efficiency. It is reasonable to hypothesize that repetitive sequences or primer self-annealing interfere with amplification rate. For instance amplicon AMPL544050257 (exon 1 of *FANCA*), which is a GC rich region (78%), did not reach the threshold of 30X in 15 samples. For amplicons AMPL1197148282 (exon 4 of *FANCB*), AMPL404630568 and AMPL388340296 (exons 11 and 20 of *FANCD1*), and AMPL433543078 (exon 17 of *FANCI*), the coverage was less than 30X in 12, 10, 21, 15 samples, respectively, more likely because of self-annealing or intra-strand loop formation of primers, as revealed by inspection of primer sequences. For the remaining 30 amplicons the coverage was less than 30X in 1-7 samples, indicating the presence of “occasional” interfering factors (Table 4). The two *FANCA* amplicons of exon 15 were not sequenced in P4 because of two large heterozygous deletions encompassing this exon (Table 1). In P28, a compound heterozygote for c.523-?_2981+?del and c.548G>A, amplicon AMPL544193477 was not covered because it was hemizygous on one allele and not amplified from the other for the presence of c.548G>A in one primer complementary sequence. Inspection of the other uncovered amplicons did not provide us with any insights into the

potential mechanisms of their low coverage. It is likely that allelic drop-out for the presence of rare variants occurred.

- **Validation I: high sensitivity and specificity of target NGS sequencing**

After sequencing quality evaluation, we determined whether the variants were true (TPs) or false positives (FPs) (Figure 5). Of the 2,005 annotations (approximately 70 per sample), we selected those (n=173; 48 different) that were located in the genomic regions of *FANCA*, *FANCC*, and *FANCG*, the genes we previously screened using Sanger sequencing (208,001nt) (De Rocco et al, 2014). Comparing data from the two technologies, we found that all the 173 variants were TPs because also detected by Sanger sequencing. Among these were 23 known mutations of *FANCA* (Table 1). Of note, no FPs was annotated. In the same regions, however, the NGS sequencing did not call 8 variants (false negative - FN; 6 different), one of which was the heterozygous c.3761_3762dup mutation of *FANCA* (P11). This duplication was not called because of allelic drop-out as it hit the complementary primer sequences of amplicons covering exon 37. Inspection of the other few FNs revealed that they were mainly localized at the 5' and 3' ends of amplicons, where the coverage drops reducing sequencing reliability. Considering these data, it will be useful to carry a careful visual inspection of these low covered regions with IGV software to evaluate the presence of potentially pathogenic variants that are not called by the TSVC plugin.

Using these data and considering the number of true negative (TN; N=207,820), we calculated the probability that the IPGM sequencing detects mutant or wild type nucleotides. At least for the *FANCA*, *FANCC*, and *FANCG* regions included in this validation (validation I; Figure 5), sensitivity $[TP/(TP + FN)]$ and specificity $[TN/(TN + FP)]$ are 95.58% and 100%, respectively. Considering that the prevalence of the mutant nucleotide is 0.09%, we estimated the positive $[PPV=TP/(TP+FP)]$ and the negative $[NPV=TN/(TN+FN)]$ predictive values as 100% and 99.99%, respectively, indicating that IPGM is a suitable technology for the screening of the FA genes.

- **Validation II: characterization of novel FA alleles**

In addition to 173 TPs (validation I), the analysis called another 1,832 variants (Figure 5), some of which were expected to be the mutations (23 alleles) not characterized yet in our cohort (Table 1). The $MAF < 0.01$ filter allowed us to select 171 putative pathogenetic variants. Screening the relative genomic regions by Sanger sequencing (36,299nt; validation II; Figure 5), we could validate another 36 annotations that were previously excluded because of their $MAF > 0.01$. Of these 207 variants, 83 (67 different) were TPs and 124 (19 different) FPs. No FN variant was missed.

Among the 67 different TPs, we searched for known FA mutations and/or deleterious variants, such as nonsense and frameshift mutations, putative alternative splicing variants and amino acid substitutions with high pathogenetic prediction score (see Materials and Methods). This allowed us to characterize 15 mutant alleles of *FANCA*, *FANCC*, *FANCF* and *FANCL* in eight (P21-P24, P26, P27, P29 and P30) of the ten samples without any FA mutations (Table 1). In particular, in sample P30 we identified three different missense mutations in *FANCL*, whose pathogenetic effect has been demonstrated by complementation analysis (De Rocco and Hanneberg, unpublished data). Of interest, one pathogenetic variant of *FANCL* (p.Pro17Arg) was also detected on one allele of control sample W2 (Table 1). The remaining TPs were heterozygous variants, most of which reported as SNPs in databases. They mainly hit intronic regions outside the consensus splice site sequences, though five could create cryptic splice sites (Table 5). Further analyses on patients' RNA samples should be performed to confirm their pathogenetic effect.

Others TPs were synonymous or benign missense variants, as determined by bioinformatics tools. However, nine, all in genes other than those causing the disease, were predicted to be pathogenetic: seven missense mutations of *FANCD1*, *FANCD2*, *FANCI*, *FANCM*, and *FANCN*, one frameshift (p.Thr367Asnfs*13) one nonsense (p.Glu417*) mutations of *FANCL* and *FANCM*, respectively (Table 1, Table 5).

Of the 19 different FPs, four were called in more than 22 patients, suggesting that they could be filtered in a standardized diagnostic workflow (Table 6). As expected, FPs were mainly localized in homopolymeric regions (Bragg et al, 2013).

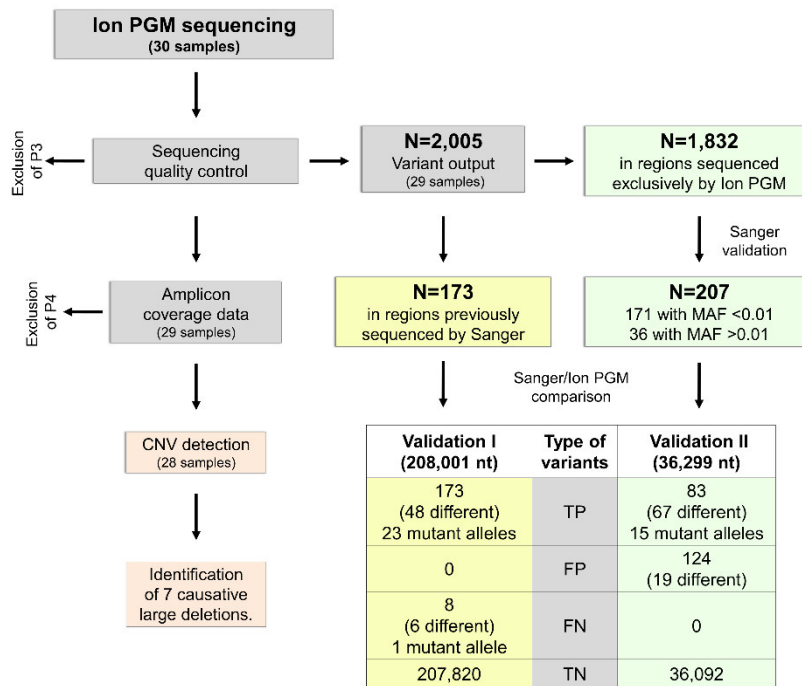


Figure 5: Schematic representation of the study on FA samples. The IPGM sequencing output underwent a quality control process leading to exclude P3 from all the further analyses and P4 from the CNV detection. Among the 2,005 variants called by IPGM, 173 were selected for validation I, as they were localized in regions (208,001nt) that were previously analyzed by Sanger sequencing in at least one of the most frequent FA gene (*FANCA*, *FANCC* and *FANCG*). Results of validation I were used to calculate sensitivity and specificity (95,58% and 100%, respectively). In validation II, variants with MAF<0.01 (N=171) were filtered from of the 1,832 variants in regions analyzed exclusively by IPGM for Sanger sequencing, analysis that included another 36 variants with MAF>0.01. The number of TP (true positive), FP (false positive), FN (false negative), and TN (true positive) variants from validations I and II are also indicated. Of the 83 TP, 15 (two on the same *FANCL* allele) were pathogenetic variants. Amplicon coverage data were also analyzed to detect CNVs (N=7)

- **Identification of FA copy number variations:
correct assignment of gender**

To determine whether the IPGM data were suitable for copy number variation (CNVs) detection, we analyzed *FANCB*, the only FA gene located on chromosome X. Consistent with our knowledge on gender, all males had median and IQR below 0.7, which is indicative of hemizyosity (Figure 6). In females, whereas the median always ranged between 0.7 and 1.3, the box plots was just above the threshold of 1.3 in three cases (P17, P22, and P23). In P4, the attribution of gender was instead ambiguous as the IQR ranged from 0.27 to 1.31 (Figure 6). Although we cannot exclude a partial deletion of *FANCB*, P4 was excluded from the CNV analysis also considering its relatively low uniformity index (86.9%; Table 3). The differences observed between males and females were not due to random variations of the *FANCB* coverage in almost all cases, as indicated by the Mann-Whitney tests' results (Table 7).

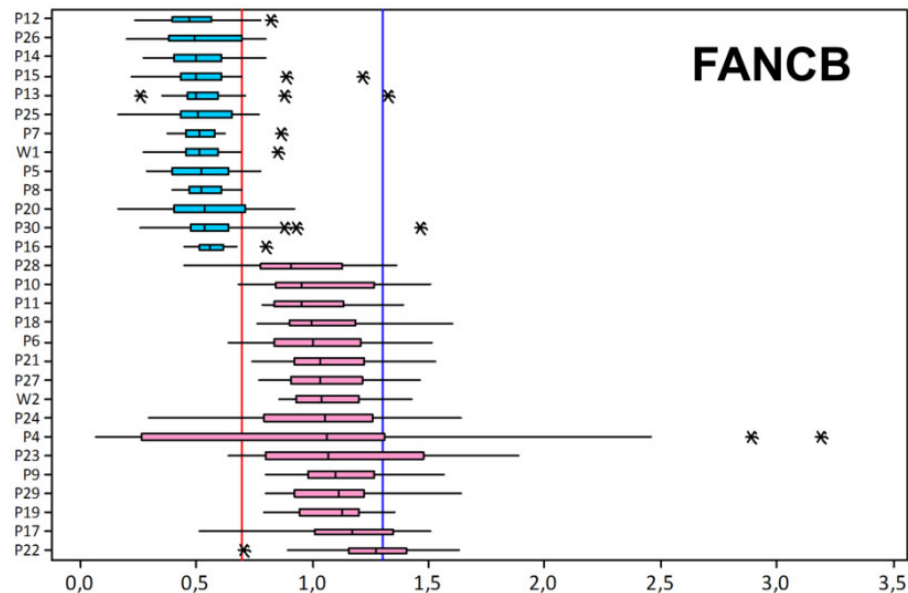


Figure 6: CNV analysis of the *FANCB* amplicons. The values were represented by medians of intersample normalization ratio and box plots reporting the interquartile range (IQR). A median below 0.7 or between 0.7 and 1.3 is indicative of a single or double copy of the gene. In the *FANCB* box plot, the median is always below 0.7 and between 0.7 and 1.3 in all the males (blue) and females (pink) respectively.

- **Identification of FA copy number variations: deletions of the entire *FANCA* and *FANCD2* genes**

After gender attribution, we determined whether other FA genes were affected by CNVs. Using the same criteria as above, we found that the median and IQR of the intersample normalization ratios in *FANCA* and *FANCD2* was significant lower than expected in samples P15 and P9, respectively (Figure 7A and 7B). The hemizygous condition of the two genes was confirmed by MLPA and SNP array, respectively (Figure 7A and 7B). Of note, since P9 was a compound heterozygous with two mutant alleles of *FANCA* (Table 1), it would be interesting to explore whether loss-of-function of *FANCA* together with haploinsufficiency of *FANCD2* makes the phenotype worse.

FIGURE 7A

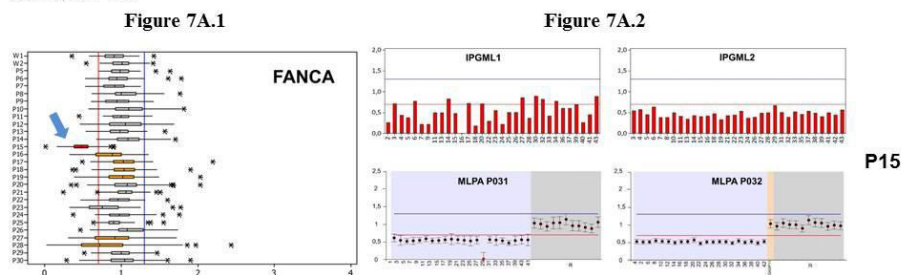


FIGURE 7B

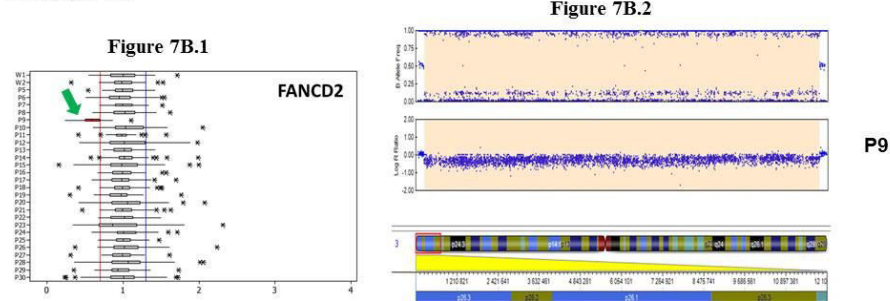


Figure 7: Detection of large deletion of the entire *FANCA* and *FANCD2* genes. **A)** In the *FANCA* box plot (figure 7A.1) the median of amplicon coverage is lower than 0.7 in the sample P15 (blue arrow), suggesting the presence of a deletion of the entire *FANCA* gene. This data was confirmed by MLPA analysis (figure 7A.2). **B)** The median of amplicon coverage of *FANCD2* is in the normal range for all the samples except for P9 (green arrow, figure 7B.1). This data is indicative of a large deletion of the entire *FANCD2*, which is confirmed by SNP array. In figure 7B.2, the plot for the B allele frequency shows a heterozygous deletion of 11.6 Mb (235,748-11,880,816) on chromosome 3p26.3-p25.2 containing the *FANCD2* gene. In this patient the plot has 10% heterozygous (AB) SNP calls, as shown by the additional allele frequency, indicating a mosaicism of approximately 90%. For probes that are normal copy number, the signal intensity ratio of the subject versus controls is expected to be 1, and $\log_2 R$ ratio should be approximately 0.0 ($\log_2 1 = 0$). In the other plot loss of copy number results in a negative \log_2 ratio of approximately -0.5 .

- **Identification of FA copy number variations: characterization of large intragenic deletions of *FANCA***

In trying to detect partial CNVs of *FANCA*, we first used the IGV software and found that in samples with known intragenic deletions the coverage of the deleted exons was reduced respect to that of the flanking undeleted amplicons. Then, a statistical evaluation revealed that the first quartile of the coverage distribution were below the threshold of 0.7 in P16 (IQR=0.67-1.0), P23 (IQR=0.58-0.95), P27 (IQR=0.66-1.09), and P28 (IQR=0.48-1.01) (Figure 8).

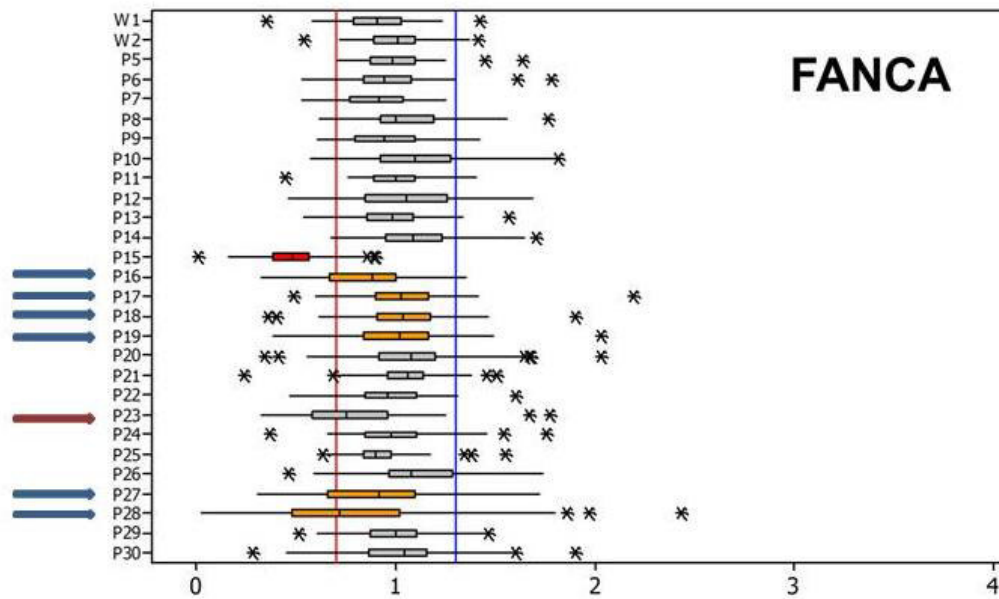
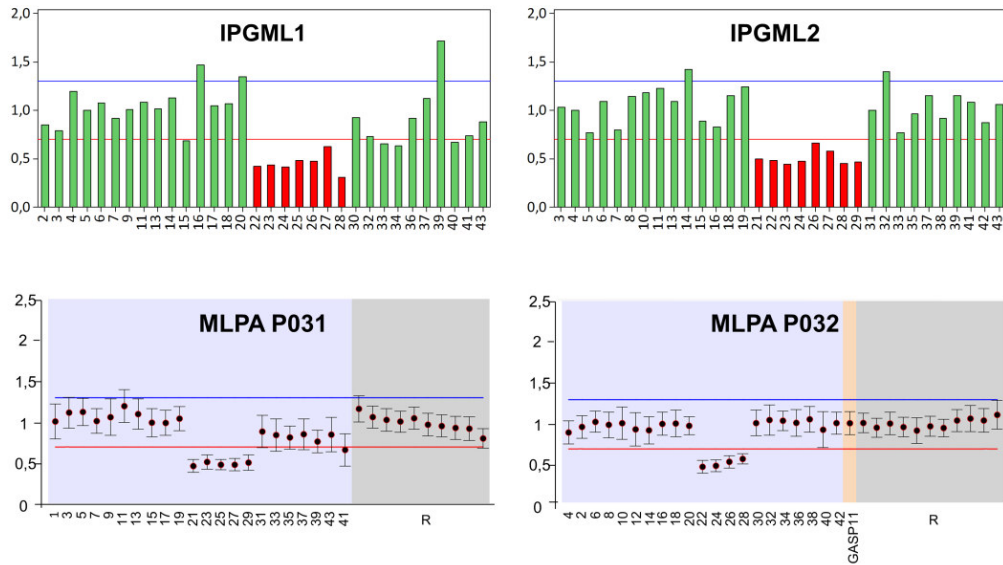


Figure 8: Large intragenic deletions in *FANCA* gene. In patient P16, P27 and P28 (blue arrows) the 1st quartile of the coverage distribution was below 0.7 (P16 =0.67, P27=0.66 and P28=0.48) suggesting potential intragenic deletions of *FANCA*. This data were confirmed with MLPA analysis. Also in P23 (red arrow) the 1st quartile of the coverage distribution was below 0.7, being of 0.58, but MLPA analysis didn't show any CNVs in this patient. The asterisks indicate the outliers.

Considering that smaller the deletion more the intersample normalization ratios of amplicons coverage resides within the normal range, we plotted the intersample normalization ratio of each amplicon. In P16, P17, P18, and P19 we confirmed the deletions we previously detected by MLPA (Table 1) (De Rocco et al, 2014). In P27 and P28 we identified two novel deletions encompassing exons 21-29 and exons 6-30, respectively, which were confirmed by MLPA analysis (Figure 9). Despite its first quartile was below 0.7, amplicon plots and MLPA analysis did not reveal any CNV in P23.

P27



P28

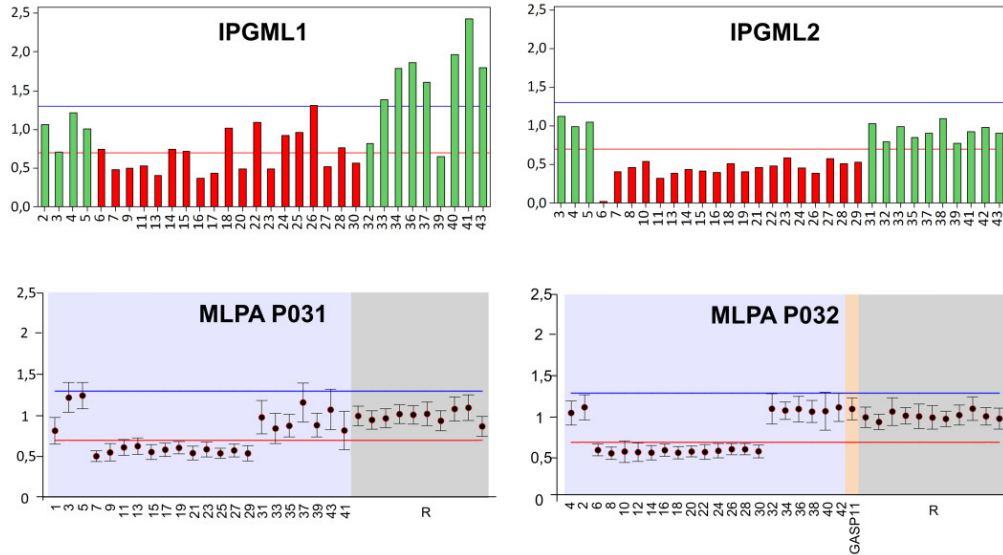


Figure 9: Detection of large intragenic deletions in FANCA. IPGML and MLPA analysis of two novel deletions of exons 21-29 (P27) and exons 6-30 (P28). Amplicons from the two IPGML libraries (IPGML1 and IPGML2) are reported in graphs showing hemizygous amplicons in red. MLPA output of two probes mix (MLPAP031 and MLPAP032) from the Coffalalyzer.net software, showing *FANCA* exons and reference loci (R) values. In both IPGML and MLPA analysis, the intersample normalization of deleted adjacent exons is under the threshold of 0.7 (red line)

DEVELOPMENT OF A NEW DIAGNOSTIC PROCESS FOR THE STUDY OF DISEASES WITH HIGH GENETIC HETEROGENEITY

After exclusion of two samples whose sequencing quality was not satisfactory, the IPGM technology allowed us to identify 44 of the 52 mutant alleles, including large intragenic deletions of *FANCA*, expected in our cohort of 26 FA DNA samples. In 19, both mutant alleles were characterized including those in relatively rare – at least in our population – complementation groups, such as *FANCC*, *FANCF*, and *FANCL*. In six cases (P11, P13, P19, P20, P28 and P29), only one heterozygous mutation was detected whereas no variant was found in P25. Whereas the second allele in P11 and P28 were FNs, the other six missed alleles could lie within uncovered amplicons. Thus, Sanger sequencing of these regions identified the second mutant *FANCA* allele in patient P29 (c.50dup), that is located in exon 1 of *FANCA*, the only amplicon not covered in the gene (Table 4). In the remaining patients, one (P13, P19, and P20) or two (P25) pathogenetic alleles remained uncharacterized. Considering that P13 and P19 were referred as potential hematopoietic mosaics and that a mosaic condition could not be excluded in P20, back mutations could have interfered with the detection of the second mutation if present only in a small percentage of reads. Finally, since in P25 we did not find any mutation, it is likely that a novel FA gene is responsible for the disease in this patient.

Taken together these data indicate that IPGM is a feasible platform for identifying mutations in the FA genes. For this reason, we have developed a diagnostic process for rapid and efficient molecular diagnosis (Figure 10). Briefly, without any previous complementation or protein analyses positive DEB test samples are analyzed with IPGM technology. After a quality control of sequencing, both CNVs and point mutations are searched. Detection of CNVs is based on normalization analyses. Results ranging from 0.7 and 1.3 are indicative of normal data, those out from this range are suggestive of presence of CNVs, which are eventually confirmed or with MLPA or with SNP array. For identification of point mutations, we consider the variant caller list, in which all the variants with $MAF < 0.01$ will be selected for further filtering. The variants enlisted in FA database or predicted to be pathogenetic will be confirmed by Sanger sequencing. Targeting all the FA genes, we cannot exclude identification of mutations in more than one gene. However, the recessive model of inheritance or complementation analysis will help defining correctly the gene responsible for the disease. Applying this algorithm to our FA patient cohort, we would have confirmed only 47 (1.2 per sample) variants by Sanger sequencing. Of these, 38 hit the disease-causing genes and 9 were

potential mutant alleles in the other FA genes, suggesting that this flow-chart is useful to further reduce time and costs of the analyses.

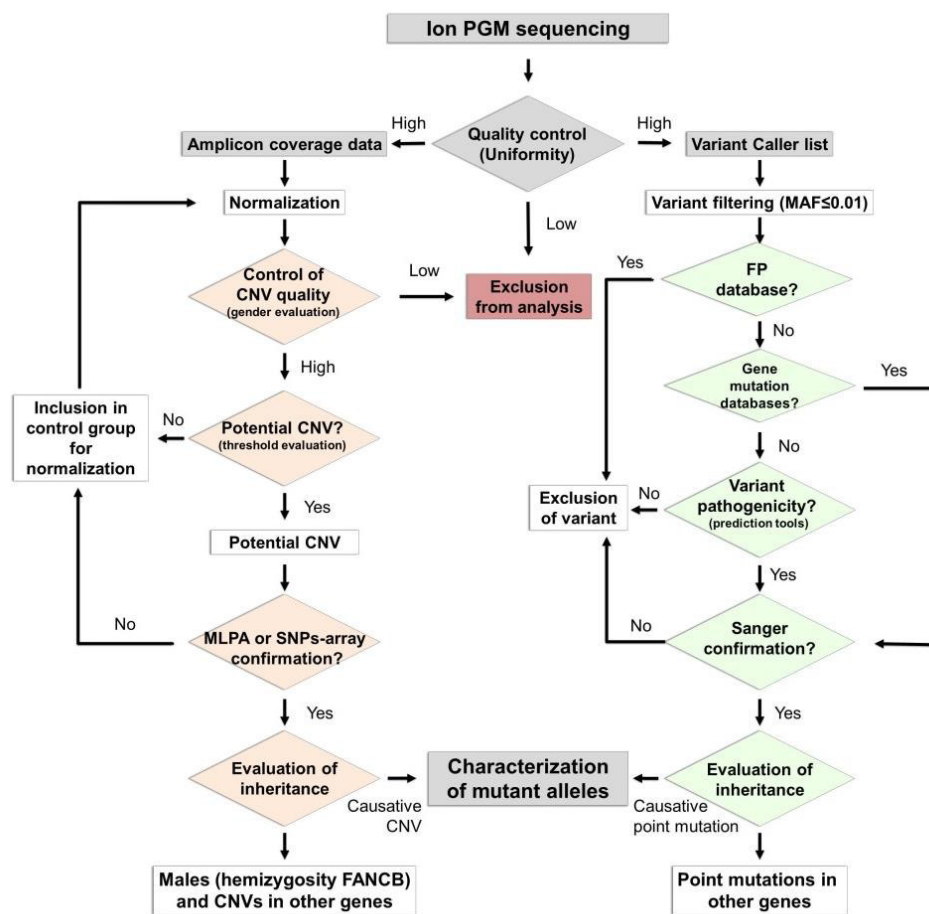


Figure 10: Proposal for a diagnostic algorithm using NGS in FA. Flow-chart showing the different steps for IPGM data processing. Samples with high uniformity of amplicon coverage are analyzed to detect CNVs and point mutations. CNV detection implies two amplicon coverage normalizations and CNV quality control consisting of correct gender assignment based on hemizygosity FANCB in males. Normalization ratios of other FA genes allow identification of potential CNVs (ratios <0.7 and >1.3 are indicative of deletions and duplications, respectively) that can then be confirmed by MLPA or SNP arrays. Simultaneously, variants from the Variant Caller list are filtered out if $MAF > 0.01$ or if included in the FA false positive (FP) database. Mutations enlisted in FA database or resulting pathogenetic using prediction tools are selected for Sanger sequencing. The recessive model of inheritance will help defining correctly the gene responsible for the disease.

In conclusion the new diagnostic algorithm proposed for FA in this study combines the detection of CNVs with the identification of point mutations in all the FA genes. This allows us to provide rapidly patients and their families with molecular diagnosis, which is crucial for a correct management of patients in terms of genetic counseling, carrier testing and prenatal diagnosis. Therefore, we hypothesized that the application of this new diagnostic algorithm could be also extend to the study of other diseases with high genetic heterogeneity, as inherited thrombocytopenias (IT), which are another field of interest of our laboratory.

APPLICATION OF THE NOVEL DIAGNOSTIC PROCESS TO THE DIAGNOSIS OF INERITHED THROMBOCYTOPENIA

To verify whether the IPGM technology could be applied for molecular diagnosis of ITs, we analyzed 21 patients with different forms of IT. In all, mutations (27 mutant alleles) were previously identified according to the diagnostic algorithm for ITs (Balduini et al, 2003).

- **First step: quality control.**

The coding exons of the 22 IT genes and approximately 30bp of the intronic flanking regions, as well as the 5' UTR regions of the *RBM8A* and *ANKRD26* genes, results in 122.85kb in size separated into 650 amplicons, which cover the 95% of the target. The remaining 5% consists of regions variably dispersed through all the genes varying from 0.07% (*FLNA*) to 29.93% (*RBM8A*). In particular *RBM8A* and *ANKRD26* are the lowest covered genes (coverage of 70.07% and 76.62% respectively) because of the presence of repetitive sequences in the 3'UTR.

The sequencing quality was lower than that reached for FA samples (Tables 3 and 8). In particular, the percentage of amplicon coverage uniformity ($87.2\% \pm 5.8$), as well as the percentage of reads on target ($77.9\% \pm 8.8$), was lower in IT than in FA samples. DNA sample quality, presence of highly repetitive regions and pseudogenes (*ANKRD26* and *TUBB1*) are likely to have interfered with sequencing efficiency. Despite this, we carried out the CNV study as the coverage uniformity was homogeneous among all IT samples, except for the IT8 (72.97%), IT10 (82.57%), IT16 (83.8%) and IT18 (72.48%) samples that were excluded from CNVs control samples.

The low efficiency of IT primer panel is confirmed also by the coverage analysis: 56 of the 650 amplicons did not achieve the threshold of 30X (Table 9). In particular, 27 were not covered in at least 10 samples, indicating that “constitutive” features of DNA, such as primer self-annealing and intra-strand loop formation or presence of GC rich regions (77%) and repetitive sequences, could have interfered with amplification reactions (Table 9).

- **Second step: characterization of novel IT alleles.**

After sequencing quality control, we analyzed the 2,225 variants identified in the 21 IT samples (on average 105 variants per sample). Among these, we selected the variants (N=116; 70 different) with MAF<0.01 (Figure 11, Table 10). Since there is no a comprehensive IT mutational database, we took into consideration the LOVD, HGMD, ClinVar, and OMIM databases, where 14 (13 different) potentially pathogenic variants resulted to be enlisted (Figure 13, Table 10).

The remaining 102 (57 different) variants were divided into these categories:

- 22 synonym (12 different);
- 35 missense (19 different);
- 1 frameshift;
- 1 nonsense;
- 43 intronic (24 different);

In order to recognize the remaining pathogenic alleles, first of all we looked at deleterious alterations, such as the nonsense mutation c.2187C>A p.(Tyr729*) in *NBEAL2* or the frameshift deletion c.2613del p.(Leu872Cysfs*38) in *ITGA2B*. Then, we evaluated the pathogenicity of missense and intronic variants with the same software used for FA analysis and identified 61 (27 different) variants with a high pathogenetic prediction score. These mutation together with those enlisted in the IT mutation database variants (N=75; 40 different) were validated by Sanger sequencing (Figure 11, Table 10 indicated by red arrow). Of these, 30 (27 different) were confirmed: 16 mutations and 14 in genes other than those causing the disease. (Figure 11, Table 10 in blue). As described for FA, we found FPs (N=45; 13 different) mainly located into homopolymeric regions that were included into a FP database we have been generating for ITs to be used as an additional filter.

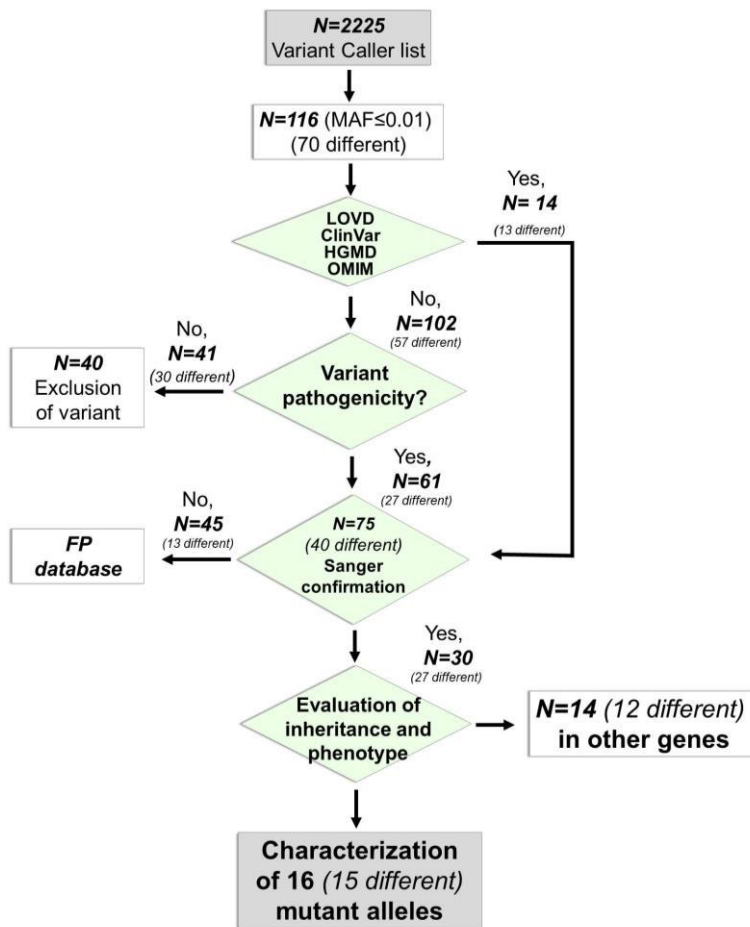


Figure 11: Application of the IT diagnostic algorithm to search point mutations. Of the 2,225 variants called by the Variant Caller in 21 IT patients, 75 were potentially pathogenic on the basis of MAF, presence in mutation databases and pathogenicity. Of these variants, 30 (16 IT mutant alleles and 14 in genes different than those causing the disease) were confirmed by Sanger sequencing. The remaining 45 FP variants were used to create a FP database to be used as an additional filter.

Six known mutant alleles were not called by IPGM. An accurate visual inspection of IGV alignment revealed that the homozygous c.392A>C mutation of *GPIBB* was not covered by primer design, whereas the heterozygous c.967+5_2del variant of *RUNX1* was located within the complementary primer sequence of exon 7. Another heterozygous mutation (c.287C>T) of *MYH9* is instead localized in a region with low sequencing reliability, being at the 3' end of AMPL1254778695. Finally, the homozygous c.382G>T nucleotide substitution in *MPL* is localized in AMPL41082843, an amplicon that was covered less than 30X in more than 3 patients (Table 10). Since the remaining 2 alleles are large intragenic deletions of *RBM8A* in IT18 and IT19 patients (Table 2), the CNV analysis could allow us to detect them.

- **Third step: CNVs analysis**

a. *Identification of IT CNVs: Correct Assignment of Gender*

In order to detect CNVs in IT samples, we first determined the correct assignment of gender, analyzing *WAS*, *GATA1* and *FLNA*, the three IT genes localized on chromosome X. As indicate in figure 13, the gender attribution was correct in all patients. Indeed, the hemizyosity condition of these genes is indicated by median values below 0.7 in males and values ranging from 0.7 to 1.3 in females. The only exception is IT16, whose median value is above the threshold for *FLNA*. As for FA, we performed the Mann-Whitney test to confirm that the differences observed between males and females were not due to random variations of the *FLNA*, *WAS* and *GATA1* coverage in almost all cases (Table 11)

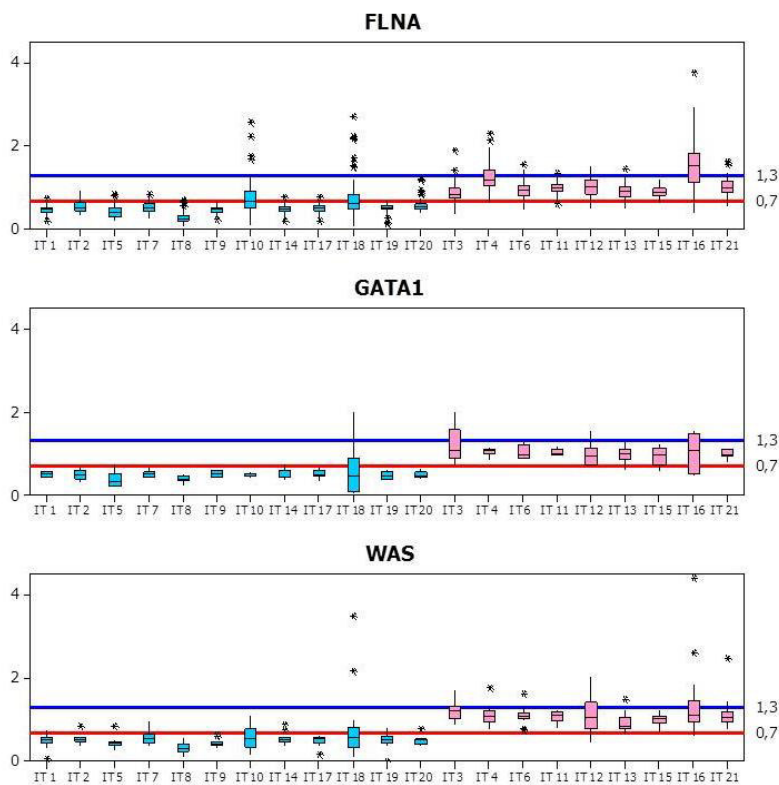


Figure 13: CNV analysis of the *FLNA*, *GATA1* and *WAS* amplicons. Values are represented by medians of intersample normalization ratio and box plots reporting the interquartile range (IQR). A median below 0.7 or between 0.7 and 1.3 is indicative of a single or double copy of the gene. In the *FLNA*, *GATA1* and *WAS* the median is always below 0.7 and between 0.7 and 1.3 in all the males (blue) and females (pink), respectively, except for the *FLNA* amplicons which are above the threshold in samples IT16.

b. Identification of IT Copy Number Variations: Deletions of RBM8A

TAR patients are compound heterozygotes carrying one null allele (mainly microdeletions on chromosome 1q21.1) and one of the two low-frequency noncoding SNPs (rs139428292 and rs201779890) in the 5'UTR region or intron 1 of *RBM8A* on the other allele (Albers et al, 2012). Since the two TAR individuals (IT18 and IT19) included in this study were apparently homozygous for the c.-21G>A SNP (rs139428292), we used the IPGM output data to confirm the hemizygous condition of *RBM8A*. In both cases, the median and the IQR values were below the threshold of 0.7 (Fig. 14).

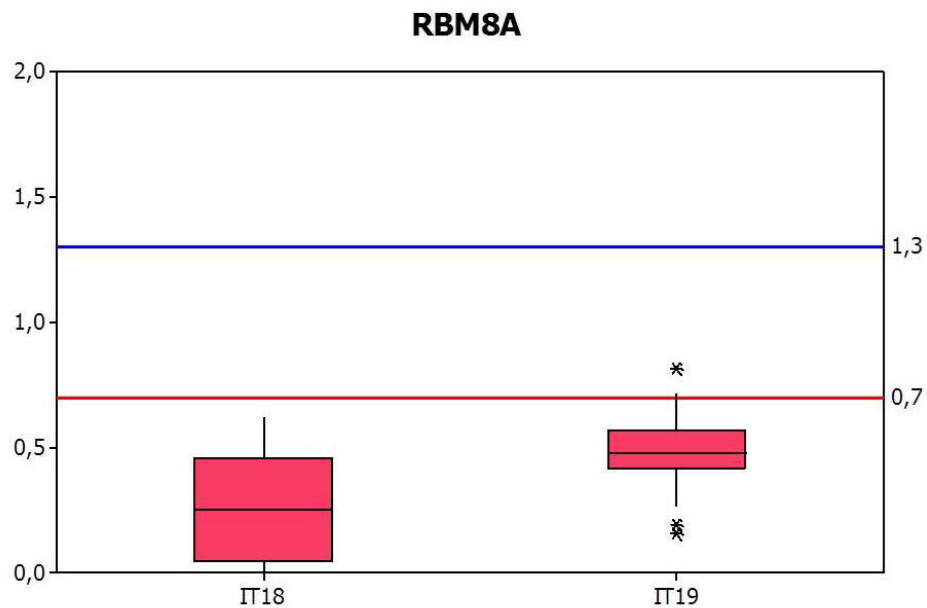


Figure 14: CNV analysis of *RBM8A* gene. In the *RBM8A* box plot, the median of coverage amplicons is below 0.7 in TAR patients IT18 and IT19, confirming the hemizygous condition of *RBM8A* likely due to a large deletion on chromosome 1q12.

c. Identification of Copy Number Variations in the other IT genes.

The statistical analysis implemented to detect CNVs did not show any structural variants in the other IT genes, except for *ANKRD26*, which resulted to be duplicated in IT5 and IT8 and deleted in IT10, IT16, IT18 (Figure 15). As indicated in the database of Genomic Structural Variations (dbVar, <http://www.ncbi.nlm.nih.gov/dbvar>), several CNVs have been reported within the chromosomal region containing *ANKRD26*. Some are also associated with diseases other than ITs. However, since their pathogenic effect is not clear, they should be first confirmed by complementary techniques, such as SNP array, and then considered in genotype/phenotype correlation studies to understand whether the *ANKRD26* CNVs make the patients' phenotype worse.

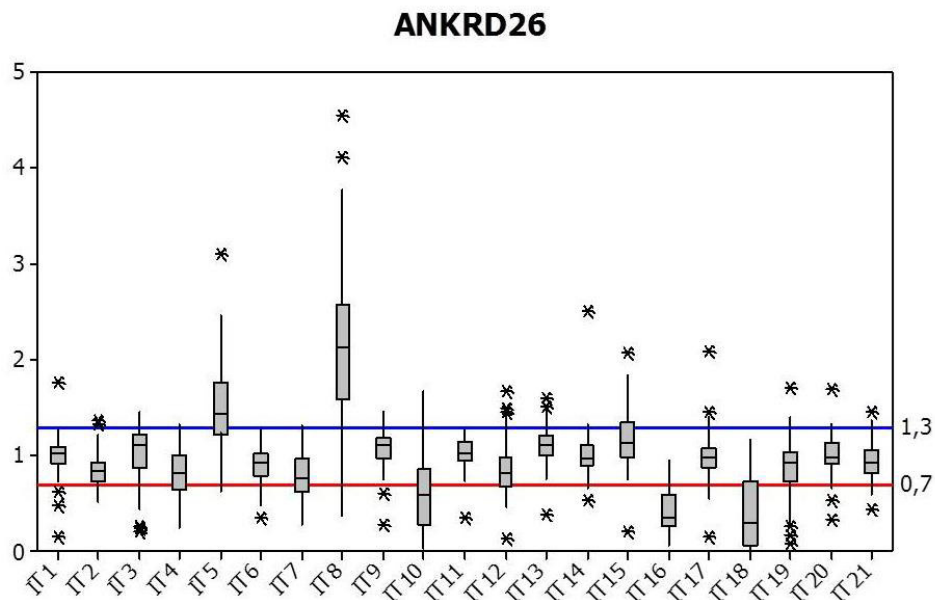


Figure 15: CNVs analysis of *ANKRD26* gene in the IT samples analyzed in this study. Box plots report median, IQR and outliers (asterisks). For all the samples, median and IQR range between 0.7 and 1.3 except for IT5 and IT8, in which a median above 1.3 is indicative of duplications, and for IT10, IT16 and IT18, in which a median lower than 0.7 is indicative of deletions..

CONCLUSIONS

In this work, we tested whether the IPGM sequencing technology could be applied to the study of diseases with high genetic heterogeneity. First of all, we validated the technology on 28 FA patients and 2 wild type samples. After exclusion of 2 samples due to low sequencing quality, we found that at least for *FANCA*, *FANCC* and *FANCG* genes, the sensitivity and the specificity were high, being of 95.58% and 100%, respectively. Moreover, we identified mutations, including large deletions of *FANCA*, in 25 of the 26 FA patients. Since we cannot exclude that new genes are involved in the disease, the only patient without any mutation identified is suitable for whole exome analysis.

Taking advantage from these data we have developed a diagnostic algorithm, that combine the study of point mutations with the CNVs analysis. In order to evaluate if this diagnostic process is feasible also for the study of the other genetically heterogeneous diseases, we have applied it also to ITs.

Despite the lower efficiency of IT primer panel, sequencing 21 IT samples we identified 2,225 variants. Of these, those (N=75) enlisted in the IT mutational databases or predicted to be pathogenic were confirmed by Sanger sequencing. Mutations, including deletions of *RBM8A* in TAR, were characterized in 17 patients.. The mutations missed in the remaining 4 patients were located or in regions with low or no coverage, as c.392A>C in *GP1BB*. To increase the mutational detection rate in IT, we will design a second panel of primers for amplification of uncovered regions, as well as of the coding regions of genes (*ETV6*, *GFI1B*, and *PRAKCG*) that have recently been cloned (Zhang, 2015; Stevenson et al, 2014; Manchev et al, 2014). As in FA, potential heterozygous pathogenic mutations were identified in loci different than the disease-causing gene. Since FA is a recessive disease, mutations in more than one gene could affect potential genotype/phenotype correlation but not the correct attribution of the complementation group. On the contrary, ITs are mainly dominant diseases. Identification of variants in more than one gene would prevent us from identifying the molecular basis of the thrombocytopenia unless genetic and functional studies are carried out. Therefore, we regard the diagnostic protocol we have developed as a useful tool for the routine molecular diagnosis of FA and ITs. It allows us to achieve a rapid and reliable search without any previous investigations, such complementation and/or protein analyses in FA and a series of molecular analyses in IT (Balduini, 2003), aimed at identifying candidate genes for mutational screening.

TABLES

TABLE 1: FEATURES OF FA DNA SAMPLES INCLUDED IN THE STUDY.

DNA sample ID	Gender	DNA sample source [†]	FANC genes analyzed by Sanger sequencing	Mutant gene	Mutations identified by Sanger sequencing [‡]	Mutations identified by Ion PGM TM [‡]	Reference
W1	M	PBC	A, G	Wild type	none	None	This study
W2	F	PBC	A, C, G	Wild type	none	[c.50C>G, p.(Pro17Arg) in <i>FANCL</i>]	This study
P3	M	PBC	A	<i>FANCA</i>	c.3558dup, p.(Arg1187Glufs*28)	Excluded from all analyses	De Rocco et al, 2014 (FA29)
					c.1360-?_2778+?del p.(Ala454_His926del) (del ex15_28)		
P4	F	PBC	A	<i>FANCA</i>	c.1360-?_1470+?del p.(Ala454_Gln490del) (del ex15)	Excluded from CNV analysis	De Rocco et al, 2014 (FA40)
					c.894-?_3348+?del p.(Trp298*) (del ex11_33)		
P5	M	PBC	A	<i>FANCA</i>	c.3788_3790del, p.(Phe1263del)	Confirmed	De Rocco et al, 2014 (FA1)
					c.3788_3790del, p.(Phe1263del)		
P6	F	PBC	A	<i>FANCA</i>	c.3239G>A, p.(Arg1080Gln)	Confirmed	De Rocco et al, 2014 (FA10)
					c.3971C>T, p.(Pro1324Leu)		
P7	M	PBC	A	<i>FANCA</i>	c.457C>T, p.(Gln153*)	Confirmed [c.1466T>C, p.(Ile489Thr) in <i>FANCL</i> and c.1249G>T, p.(Glu417*) in <i>FANCM</i>]	De Rocco et al, 2014 (FA13)
					c.709+1G>A		

P8	M	PBC	A	FANCA	c.3788_3790del, p.(Phe1263del)	Confirmed [c.1096_1099dup, p.(Thr367Asnfs*13) in <i>FANCL</i>]	De Rocco et al, 2014 (FA21)
					c.3971C>T, p.(Pro1324Leu)		
P9	F	PBC	A	FANCA	c.2638C>T, p.(Arg880*)	Confirmed [c.-78-?_5024+?del (del ex1_44) in <i>FANCD2</i>]	De Rocco et al, 2014 (FA37)
					c.3164G>A, p.(Arg1055Gln)		
P10	F	LCL	A	FANCA	c.3638_3639del, p.(Pro1213Argfs*64)	Confirmed	De Rocco et al, 2014 (FA42)
					c.3971C>T, p.(Pro1324Leu)		
P11	F	LCL	A	FANCA	c.3761_3762dup, p.(Glu1255Argfs*12)	False negative [#]	De Rocco et al, 2014 (FA43)
					c.2574C>G, p.(Ser858Arg)	Confirmed [c.2204G>A, p.(Arg735Gln) in <i>FANCD2</i>]	
P12	M	PBC	A	FANCA	c.2290C>T, p.(Arg764Trp)	Confirmed	De Rocco et al, 2014 (FA44)
					c.4029T>G, p.(His1343Gln)		
P13 [§]	M	PBC	A	FANCA	c.1450G>C, p.(Glu484Gln)	Confirmed	This study
					not found	not found	
P14	M	PBC	A, G	FANCA	c.1115_1118del, p.(Val372Alafs*42)	Confirmed	De Rocco et al, 2014 (FA85)
					c.1126C>T, p.(Gln376*)		
P15	M	PBC	A	FANCA	c.2812_2830dup, p.(Asp944Glyfs*5)	Confirmed	De Rocco et al, 2014 (FA20)
					c.-42-?_5481+?del (del ex1_43)		
P16	M	PBC	A	FANCA	c.1850_1859del, p.(Leu617Profs*20)	Confirmed	De Rocco et al, 2014 (FA67)
					c.1827-?_2778+?del p.(Arg609Serfs*2) (del ex21_28)		

P17	F	PBC	A	FANCA	c.457C>G, p.(Gln153Glu)	Confirmed [c.1538G>A, p.(Arg513Gln) in FANCG and c.2816T>G, p.(Leu939Trp) in FANCM]	De Rocco et al, 2014 (FA65)
					c.3490C>T, p.(Pro1164Ser)		
					c.1471-?_1626+?del p.(Val491_Glu542del) (del ex16_17)		
P18	F	PBC	A, C, G	FANCA	c.893+5G>A p.(Phe879Valfs*19)	Confirmed [c.1151C>T, p.(Ser384Phe) in FANCD1]	De Rocco et al, 2014 (FA28)
					c.190-?_283+?del p.(Val64Alafs*43) (del ex3)		
P19 ^l	F	PBC	A, G	FANCA	c.1360-?_1826+?del p.(Ala454Serfs*3) (del ex15_20)	Confirmed	De Rocco et al, 2014 (FA61)
					not found		
P20 ^{ll}	M	PBC	A, C, G	FANCA	c.4258G>T, p.(Glu1420*)	Confirmed [c.874C>G, p.(Pro292Ala) and c.5848T>G, p.(Leu1950Val) in FANCM]	De Rocco et al, 2014 (FA41)
					not found		
P21	F	PBC	A, G	FANCC	not found	c.67del, p.(Asp231Ilefs*23) c.67del, p.(Asp231Ilefs*23)	De Rocco et al, 2014 (FA88)
P22	F	LCL	A, G	FANCC	not found	c.37C>T, p.(Gln13*) c.692_694del, p.(Lys231del)	De Rocco et al, 2014 (FA89)
P23	F	PBC	A, G	FANCC	not found	c.37C>T, p.(Gln13*) c.1069C>T, p.(Gln357*) [c.9875C>T, p.(Pro3292Leu) in FANCD1]	This study

P24	F	LCL	G	FANCF	not found	c.484_485del p.(Leu162Aspfs*103)	This study
						c.484_485del p.(Leu162Aspfs*103) [c.1538G>A, p.(Arg513Gln) in FANCG]	
P25	M	PBC	A, G	not found	not found	not found	This study
P26	M	PBC	none	FANCA	not analyzed	c.3788_3790del, p.(Phe1263del)	De Rocco et al, 2014 (FA58)
						c.826+3del, p.(250_251insGlyAlaPhe MetThrArgCysGlyPheLeu)	
P27	F	PBC	none	FANCA	not analyzed	c.1776+7A>G, p.(Ile573Serfs*12)	De Rocco et al, 2014 (FA66)
						c.1827-?_2852+?del, p.(Ala610_Arg951del) (del ex21_29)	
P28	F	PBC	none	FANCA	not analyzed	c.548G>A, p.(Trp183*) [^]	De Rocco et al, 2014 (FA49)
						c.523-?_2981+?del p.(Ser175Leufs*5) (del ex6_30)	
P29	F	PBC	none	FANCA	not analyzed	c.3660del, p.(Asn1221Thrfs*26)	De Rocco et al, 2014 (FA47)
						c.50dup, p.(Arg18Profs*19) ^{††}	
P30	M	LCL	none	FANCL	not analyzed	c.50C>G, p.(Pro17Arg)	This study
						c.676C>T, p.(Arg226Cys)	
						c.1021T>A, p.(Trp341Arg)	

[†]PBC, peripheral blood cells; LCL, lymphoblast cell line.

[‡] Nucleotide A of the ATG translation initiation start site of the *FANCA*, *FANCC*, *FANCD1*, *FANCD2*, *FANCF*, *FANCG*, *FANCI*, *FANCL*, *FANCM* and *FANCN* cDNAs from GenBank sequences NM_000135.2, NM_000136.2, NM_000059.3, NM_001018115.1, NM_022725.3, NM_004629.1, NM_032043.2, NM_018062.3, NM_020937.2 and NM_024675.3, respectively is indicated as nucleotide +1. In square brackets are potential pathogenetic heterozygous variants identified in genes different from that causing the disease.

[§] Sib of patient FA70 in De Rocco et al. (2014). As FA 70, who was found to be a, P19 was enrolled as an FA patient with a potential hematopoietic mosaicism. For this reason, the second mutant allele could be missed.

[‖] Patient with potential hematopoietic mosaicism because of revertant lymphoblastoid cell line. Complementation analysis carried out in peripheral blood T lymphocytes assigned this patient to *FANCA* genetic group but only one heterozygous mutation was identified. Due to the mosaic suspicion, the second mutant allele could be missed.

[¶] Potential hematopoietic mosaicism status was not ascertained in this patient, whose peripheral blood cells were the only biological sample available.

[#] Mutation localized in a region where the reverse and forward primers of two adjacent amplicons aligned, reducing sequence efficiency of this mutant allele. Mutation was not called by the TSVC because seen only 47 of 262 reads.

^{††} Identified by Sanger sequencing of AMPL544050257 (exon 1), the only amplicon uncovered in *FANCA*.

[^] Variant not identified because of allelic dropout in exon 15 of *FANCA*.

TABLE 2: FEATURES OF IT DNA SAMPLES INCLUDED IN THE STUDY.

DNA sample ID	Gender	DNA sample source [†]	IT genes analyzed by Sanger sequencing	Mutant gene	Mutations identified by Sanger sequencing [‡]	Mutations identified by Ion PGM ^{TM†}	Reference
IT1	M	PBC	ACTN1	ACTN1	c.751G>A, p.(Gly251Arg)	Confirmed	Bottega et al., 2014
IT2	M	PBC	ACTN1	ACTN1	c.2210C>A, p.(Thr737Asn)	Confirmed	Bottega et al., 2014
IT3	F	PBC	ANKRD26	<i>ANKRD26</i>	c.-127A>T, p.(=)	Confirmed	Pippucci et al., 2011
IT4	F	PBC	ANKRD26	<i>ANKRD26</i>	c.-134G>A, p.(=)	Confirmed [c.1212-7T>A in <i>ABCG8</i> and c.2473+53C>A in <i>NBEAL2</i>]	Pippucci et al., 2011
IT5	M	PBC	C-MPL	C-MPL	c.3239G>A, p.(Arg1080Gln) c.3971C>T, p.(Pro1324Leu)	Confirmed [CNV in <i>ANKRD26</i> ; c.5143G>A in <i>MYH9</i> and c.270+95delT in <i>RUNX1</i>]	This study
IT6	F	PBC	C-MPL	<i>C-MPL</i>	c.382G>T, p.(Asp128Tyr)	False Negative [§]	Savoia et al., 2007
IT7	M	PBC	CYCS, ANKRD26	CYCS	c.154G>A, p.(Ala52Thr)	Confirmed	This study
IT8	M	PBC	GATA1	GATA1	c.647G>A, p.(Arg216Gln)	Confirmed [CNV in <i>ANKRD26</i> ; c.220+71G>A and c.7548+66C>G in <i>WVF</i>] Excluded from CNVs control samples ^{††}	This study
IT9	M	PBC	GP1BA	<i>GP1BA</i>	c.104del, p.(Lys35Argfs*4)	Confirmed [c.67+32C>A in <i>RBM8A</i>]	This study

IT10	M	PBC	GP1BA, GP1BB, GPIX	GP1BB	c.47T>C, p.(Leu16Pro)	Confirmed [CNV in ANKRD26 and c.2473+53C>A in <i>NBEAL2</i>] Excluded from CNVs control samples ^{††}	This study
IT11	F	PBC	GP1BA, GP1BB, GPIX	GP1BB	c.392A>C, p.(Tyr131Ser)	False negative ^l	This study
IT12	F	PBC	ITGB3	<i>ITGB3</i>	c.2134+1G>C	Confirmed [c.1898+33C>T in <i>NBEAL2</i>]	This study
IT13	F	PBC	MYH9	MYH9	c.287C>T, p.(Ser96Leu)	False negative ^{ll}	This study
IT14	M	PBC	MYH9	MYH9	c.2114G>A, p.(Arg705His)	Confirmed [c.1385+89A>C in <i>ACTN1</i> and c.5026A>G, p.(Lys1676Glu) in <i>MYH9</i>]	Verver et al, 2014
IT15	F	PBC	MYH9	MYH9	c.2114G>A, p.(Arg705His)	Confirmed	Verver et al, 2014
IT16	F	PBC	MYH9	MYH9	c.5521G>A, p.(Glu1841Lys)	Confirmed [CNV in ANKRD26] Excluded from CNVs control samples ^{††}	This study
IT17	M	PBC	ANKRD26, NBEAL2	<i>NBEAL2</i>	c.2187C>A, p.(Tyr729*)	Confirmed	Bottega et al, 2013
IT18	M	PBC	RBM8A	<i>RBM8A</i>	deletion in chromosome 1q21.1 c.-21G>A (rs139428292, in homozygosis)	Confirmed [CNV in ANKRD26 and c.2054C>T, p.(Pro685Leu) in <i>MYH9</i>] Excluded from CNVs control samples ^{††}	This study
IT19	M	PBC	RBM8A	<i>RBM8A</i>	deletion in chromosome 1q21.1 c.-21G>A (rs139428292, in homozygosis)	Confirmed	This study

IT20	M	PBC	ANKRD26, RUNX1	<i>RUNX1</i>	c.967+2_5del	False negative [#] [c.2473+53C>A in <i>NBEAL2</i>]	This study
IT21	F	PBC	MYH9	<i>MYH9</i>	c.5797C>T, p.(Arg1933*)	Confirmed [c.6798+11G>A in <i>VWF</i>]	This study

[†]PBC, peripheral blood cells

[‡]Nucleotide A of the ATG translation initiation start site of the *ACTN1*, *ANKRD26*, *c-MPL*, *CYCS*, *GATA1*, *GP1BA*, *GP1BB*, *ITGB3*, *MYH9*, *NBEAL2*, *RBM8A*, *RUNX1* cDNAs from GenBank sequences NM_001102.3, NM_014915.2, NM_005373.2, NM_018947.5, NM_002049.3, NM_000173.5, NM_000407.4, NM_000212.2, NM_002473.4, NM_015175.2, NM_005105.3, NM_001001890.2, respectively is indicated as nucleotide +1. In square brackets are potential pathogenetic heterozygous variants identified in genes different from that causing the disease.

[§] Mutation localized in amplicon AMPL 410828843 not covered by IPGM sequencing.

^{||} Mutation localized in a region of *GP1BB* gene not covered by IPGM primer design

[¶] Mutation localized at the end of the amplicon AMPL1254778695, region where the coverage drops reducing sequencing reliability. Mutation was not called by the TSVC because seen only 56 of 173 reads.

[#] Mutation localized in a region where the reverse and forward primers of two adjacent amplicons aligned, reducing sequence efficiency of this mutant allele. Mutation was not called by the TSVC because seen only

^{††} These samples were excluded from CNVs control samples because of their low coverage uniformity (Table 15)

TABLE 3:QUALITY CONTROL DATA OF 30 FA SAMPLES ACCORDING TO IPGM ANALYSIS.

Sample	N. of bases per sample*	Bases with $Q \geq 20$ (%) [†]	Mapped reads [‡]	Reads on target (%) [§]	Average reads per amplicon	Uniformity of amplicon coverage (%) [¶]	Amplicons reading end-to-end (%) [#]
W1	108,116,974	89.6	781,876	97.4	1,099	96.8	71.4
W2	75,262,526	80.9	603,475	95.0	828	97.3	29.7
P3	35,785,078	82.7	267,284	97.4	375	49.2	18.8
P4	119,723,202	82.7	895,474	97.6	1,261	86.9	36.1
P5	70,204,308	85.2	521,101	96.1	722	96.5	52.7
P6	46,370,665	87.9	332,419	97.0	465	95.2	66.4
P7	91,330,404	83.2	680,112	97.1	953	96.5	38.1
P8	114,403,621	85.5	856,185	94.9	1,172	96.8	51.1
P9	78,309,348	90.1	564,085	97.5	793	97.9	77.1
P10	39,666,030	83.0	304,766	95.7	421	95.4	42.9
P11	50,060,326	80.6	403,472	93.6	546	97.3	29.4
P12	49,900,000	89.6	370,530	95.1	509	96.3	75.5
P13	96,878,877	80.8	784,389	95.5	1,082	94.2	29.3
P14	64,792,636	84.6	490,221	93.4	660	96.5	45.6
P15	170,110,091	90.1	1,226,673	98.5	1,743	93.5	72.6
P16	61,705,451	86.5	468,489	92.1	623	97.7	68.4
P17	81,994,349	84.8	603,842	96.8	843	96.8	52.0
P18	80,418,943	84.7	607,565	94.4	827	97.8	44.6
P19	47,553,650	80.3	384,555	94.4	525	96.8	26.7
P20	23,598,575	87.7	170,190	97.1	238	96.1	67.8
P21	84,534,746	84.8	634,677	91.2	835	97.3	48.3
P22	14,437,622	82.7	112,325	95.8	155	97.6	40.7
P23	127,535,800	85.7	914,955	97.3	1,284	96.3	59.6
P24	30,683,292	84.9	225,630	96.6	314	96.7	54.3
P25	57,285,000	84.8	423,405	96.0	587	96.7	53.4
P26	74,338,167	85.0	560,816	95.7	774	96.0	46.9
P27	46,719,940	80.8	377,211	95.9	523	97.3	28.4
P28	56,726,182	87.9	407,532	96.7	967	95.5	74.3
P29	78,923,835	81.2	631,032	95.6	873	98.0	30.7
P30	43,969,366	85.0	334,231	94.0	453	96.8	45.2
Min	14,437,622	80.3	112,325	91.2	155	49.2	18.8
Max	170,110,091	90.1	1,226,673	98.5	1,743	98.0	77.1
Mean	68,751,834	84.8	518,053	95.7	748	94.7	49.3
St.dev	32,789,486	2.9	240,79	16.8	350	8.8	16.8
Median	64,792,636	84.8	490,221	95.9	748	96.7	47.6
1 st quartile	46,719,940	82.7	370,530	94.9	512	96.0	36.6
3 rd quartile	81,994,349	86.3	631,032	97.0	933	97.3	64.7

*Total number of filtered and trimmed reads independent of length post filtering bases per barcode.

†The percentage of reads that have a predicted quality score of Q20 or better. A Q20 score is the predicted quality of a Phred-like score of 20 or better, or one error in 100 bp.

#Total number of reads mapped to the reference.

§The percentage of reads mapped to any targeted region relative to all reads mapped to the reference.

‡The average number of reads assigned to amplicons. The average number of reads assigned to amplicons.

¶ The percentage of bases in all targeted regions covered by at least 0.2x the average base read depth.

#The percentage of all amplicons that were considered to have a sufficient proportion of assigned reads (70%) that covered the whole amplicon target from 'end-to-end'. To allow for error the effective ends of the amplicon region for read alignment are within 2 bases of the actual ends of the region.

TABLE 4:AMPLICONS NOT COVERED MORE THAN 30X IN FA SAMPLES

Gene	Exon	ID amplicons	N. of samples with coverage <30X
<i>FANCA</i>	1	AMPL544050257 ^a	15 (including P29 ^a)
	6	AMPL544193477	1 (P28) ^b
	7	AMPL2671673445	2
	12	AMPL544284467	5
	15	AMPL544674964	1 (P4) ^c
		AMPL544681421	1 (P4) ^c
	16	AMPL544342332	1
<i>FANCB</i>	3	AMPL544417387	4
	4	AMPL1197148282 ^d	12
		AMPL544382746	1
<i>FANCC</i>	4	AMPL592054836	5
<i>FANCD1</i>	8	AMPL388297720	1
	11	AMPL388360704	1
		AMPL404630568 ^d	10
		AMPL658814386	2
	16	AMPL624416753	6
20	AMPL388340296 ^d	21	
<i>FANCD2</i>	17	AMPL434371586	4
	26	AMPL442020299	1
<i>FANCE</i>	10	AMPL544603911	1
<i>FANCG</i>	6	AMPL543852913	5
<i>FANCI</i>	11	AMPL1190859878	1
	32	AMPL544478748	2
<i>FANCI</i>	8	AMPL416725112	7
	11	AMPL414710255	6
	15	AMPL434475581	1
	17	AMPL433543078 ^d	15
<i>FANCL</i>	1	AMPL543889772	1
<i>FANCM</i>	5	AMPL1156281934	3
	6	AMPL544176673	2
	14	AMPL2595994868	1
	18	AMPL1104380167	1
	19	AMPL1103979554	6
	19	AMPL1157014550	1
<i>FANCP</i>	14	AMPL544488815	2

^aGC-rich region; in P29 Sanger sequence revealed a pathogenetic c.50dup mutation;

^bIn P28, amplicon was deleted on one allele and not amplified on the other for allelic drop-out (c.548G>A present on one primer complimentary sequence);

^cP4 had two large genomic deletion of *FANCA* encompassing exon 15;

^dSelf-annealing or intra-strand loop formation were likely to be responsible for low amplicon coverage.

TABLE 5: TRUE POSITIVE (TP) VARIANTS IDENTIFIED ON FA SAMPLES IN VALIDATION II.

Number	Gene	TP variants [†]	rs number (MAF in 1000 Genome database)	Pathogenicity	N. of samples with TP (Sample with pathogenetic variant)
1	<i>FANCA</i> (NM_000135.2)	c.548G>A p.(Trp183*)	nr	nonsense	1 (P28)
2		c.826+3del p.(250_251insGlyAla PheMetThrArgCysGlyPheLeu)	nr	splicing (in frame) [§]	1 (P26)
3		c.1776+7A>G p.(Ile573Serfs*12)	nr	splicing (frameshift) [§]	1 (P27)
4		c.3660del p.(Asn1221Thrfs*26)	nr	small deletion (frameshift)	1 (P29)
5		c.3788_3790del p.(Phe1263del)	nr	small deletion (in frame)	1 (P26)
6		c.4260+29T>C	rs1800359 (0.18)	nd	1
7		c.4332T>G p.(=)	rs149531696 (nr)	synonym	1
8	<i>FANCB</i> (NM_001018113.1)	c.147A>G p.(=)	nr	synonym (donor splice site - score 68) [†]	1
9	<i>FANCC</i> (NM_000136.2)	c.37C>T p.(Gln13*)	rs121917784 (nr)	nonsense	2 (P22, P23)
10		c.67del p.(Asp23Ilefs*23) homozygous	rs104886459 (nr)	small deletion (frameshift)	1 (P21)
11		c.554G>A p.(Arg185Gln)	rs370346767 (nr)	missense (1) [†]	1
12		c.692_694del p.(Lys231del)	rs3831244 (nr)	small deletion (in frame)	1 (P22)
13		c.816C>T p.(=)	rs55719336 (<0.01)	synonym	1
14		c.1069C>T p.(Gln357*)	nr	nonsense	1 (P23)
15	<i>FANCD1</i> (NM_000059.3)	c.800G>A p.(Gly267Glu)	rs80359036 (nr)	missense (0) [†]	1
16		c.1114A>C p.(Asn372His)	rs144848 (0.24)	missense (2) [†]	1
17		c.1151C>T p.(Ser384Phe)	rs41293475 (nr)	missense (5) [†]	1 (P18)
18		c.1938C>T p.(=)	rs28897711 (<0.01)	synonym	1
19		c.4258G>T p.(Asp1420Tyr)	rs28897727 (<0.01)	missense (2) [†]	1
20		c.8755-66T>C	rs4942486 (0.47)	nd	1
21		c.8953+98T>C	rs81002901 (<0.01)	nd	1
22		c.9875C>T p.(Pro3292Leu)	rs56121817 (<0.01)	missense (8) [†]	1 (P23)

23	<i>FANCD2</i> (NM_001018115.1)	c.65-18A>C	nr	nd	1
24		c.784-27del	nr	nd (acceptor splice site - score 68) [†]	1
25		c.1634A>G p.(Asn545Ser)	rs145522204 (<0.01)	missense (2) [†]	2
26		c.2021+31C>T	rs3864015 (nr)	nd	11
27		c.2103G>T p.(=)	rs139033444 (nr)	synonym	1
28		c.2204G>A p.(Arg735Gln)	nr	missense (8) [†]	1 (P11)
29		c.2269+15C>T	nr	nd	1
30		c.2606-40A>T	rs36075953 (<0.01)	nd	1
31		c.2977-39C>T	rs45622438 (<0.01)	nd	1
32		c.3336-74G>A	rs34197804 (<0.01)	nd	1
33		<i>FANCE</i> (NM_021922.2)	c.1572G>A p.(=)	rs115195341 (<0.01)	synonym
34	<i>FANCF</i> (NM_022725.3)	c.484_485del p.(Leu162Aspfs*103) homozygous	nr	small deletion (frameshift)	1 (P24)
35		c.557C>T p.(Ala186Val)	rs113910234 (<0.01)	missense (0) [†]	1
36	<i>FANCG</i> (NM_004629.1)	c.1538G>A p.(Arg513Gln)	rs17885240 (<0.01)	missense (0) [†]	2 (P17, P24)
37		c.1636+7A>G	rs587118 (0.25)	nd	1
38	<i>FANCI</i> (NM_001113378.1)	c.755+55G>A	rs183255195 (<0.01)	nd (donor splice site - score 66) [†]	1
39		c.976-125 T>C	nr	nd	1
40		c.1963G>A p.(Gly655Arg)	rs138026584 (nr)	missense (3) [†]	2
41		c.1992+10T>C	nr	nd	1
42		c.2203A>G p.(Ile735Val)	nr	missense (0) [†]	1
43		c.2225G>C p.(Cys742Ser)	rs2283432 (0.29)	missense (1) [†]	1
44		c.2292-120T>C	rs11855524 (0.07)	nd	1
45		c.2367G>T p.(=)	rs11857960 (0.08)	synonym	1
46	<i>FANCI</i> (NM_032043.2)	c.1466T>C p.(Ile489Thr)	nr	missense (7) [†]	1 (P7)
47		c.2286T>C p.(=)	rs61754141 (<0.01)	synonym	1
48		c.2906-31A>G	nr	nd (acceptor splice site - score 71) [†]	1
49	<i>FANCL</i> (NM_018062.3)	c.50C>G p.(Pro17Arg)	nr	missense (8) [†]	2 (W2, P30)
50		c.676C>T p.(Arg226Cys)	nr	missense (8) [†]	1 (P30)
51		c.1021T>A p.(Trp341Arg)	nr	missense (8) [†]	1 (P30)
52		c.1096_1099dup p.(Thr367Asnfs*13)	nr	smallinsertion (frameshift)	1 (P8)

53	<i>FANCM</i> (NM_020937.2)	c.171G>C p.(Leu57Phe)	rs142007602 (<0.01)	missense (4) [†]	1
54		c.874C>G p.(Pro292Ala)	rs142747831 (<0.01)	missense (6) [†]	1 (P20)
55		c.1249G>T p.(Glu417*)	nr	nonsense	1 (P7)
56		c.1397-58T>G	rs11157433 (0.24)	nd (acceptor splice site - score 77) [†]	2
57		c.1397-16_1397-14del	nr	nd	2
58		c.5848T>G p.(Leu1950Val)	rs146436929 (<0.01)	missense (6) [†]	1 (P20)
59		<i>FANCN</i> (NM_024675.3)	c.2586+58C>T	rs249954 (0.34)	nd
60	c.2586+81C>T		rs114710547 (<0.01)	nd	1
61	c.2587-38C>G		rs180177119 (nr)	nd	1
62	c.2816T>G p.(Leu939Trp)		rs45478192 (<0.01)	missense (8) [†]	1 (P17)
63	<i>FANCO</i> (NM_058216.2)	c.790G>A p.(Gly264Ser)	rs147241704 (nr)	missense (3) [†]	1
64	<i>FANCP</i> (NM_032444.2)	c.590T>C p.(Val197Ala)	rs147826749 (<0.01)	missense (0) [†]	1
65		c.753G>A p.(=)	rs8061528 (0.25)	synonym	1
66		c.4580C>T p.(Pro1527Leu)	rs149362820 (nr)	missense (0) [†]	1
67		c.5501A>G p.(Asn1834Ser)	rs111738042 (<0.01)	missense (0) [†]	1
Total TP variants					83

[†]TP (true positive) variants are all heterozygous except for mutation #34. Nucleotide A of the ATG translation initiation start site of the cDNAs from reference sequence is indicated as nucleotide +1. In gray, pathogenic variants of the disease-causing gene. In darker gray, heterozygous potential pathogenic variants in a gene different from that causing the disease.

[‡]Effect of the missense variations was evaluated using four pathogenicity prediction programs, Mutation Taster, Mutation Assessor and SIFT. and of intronic variants using Human Splicing Finder Version 2.4.1 (see Materials and Methods).

[§]Confirmed by RT-PCR by De Rocco et al (2014)

TABLE 6:FALSE POSITIVE (FP) VARIANTS IDENTIFIED ON FA SAMPLES IN VALIDATION II

Number	Gene	FP variants [†]	rs number (MAF in 1000 Genome Database)	Status	N. of samples with FP	Sequence contest of FP [‡]
1	<i>FANCC</i> (NM_000136.2)	c.1297del	not reported	homozygous	6	GCCCCCGTGATGGG
2	<i>FANCD2</i> (NM_001018115.1)	c.696-19C>T	not reported	heterozygous	1	CTTTTTTCTTTTTCT
3		c.1134+39T>G	not reported	heterozygous	1	CAGACTTAAAAGTA
4		c.1170C>T	rs1122887807 (not reported)	heterozygous	25	TAGCACCAATACTCAGACAAA
5		c.1179T>C	rs72492998 (not reported)	heterozygous	25	TAGCACCAATACTCAGACAAA
6		c.1214A>G	rs73126218 (not reported)	heterozygous	25	CTAAGAAATAAGAT
7		c.2715+96A>T	not reported	heterozygous	1	GAAAAAAAAAATTAG
8		<i>FANCE</i> (NM_021922.2)	c.970-26T>G	rs192832780 (0.0009)	heterozygous	1
9	c.1113+23_1113+24insC		not reported (0.0018)	heterozygous	1	TGGGAGGTACTC(C)AGAGTGCCAAG
10	<i>FANCI</i> (NM_001113378.1)	c.3537+41C>A	not reported	heterozygous	1	TTCTACCCCAGT
11		c.3947G>A	rs138461165 (0.0005)	heterozygous	1	ATGGGGGACAGAA
12	<i>FANCI</i> (NM_001113378.1)	c.1543G>T	not reported	heterozygous	2	GCAAGAGAAGTA
13		c.2575+8C>A	not reported	heterozygous	22	GATCTCAGCTGGG
14	<i>FANCL</i> (NM_018062.3)	c.375-26A>T	not reported	heterozygous	3	GATCATTTTTTATTC
15		c.375-25T>A	not reported	heterozygous	4	GATCATTTTTTATTC
16	<i>FANCM</i> (NM_020937.2)	c.1397-16_1397-14del	not reported	homozygous	1	TTAAAGTTTTTATATATATATATAG
17	<i>FANCP</i> (NM_032444.2)	c.3661del	not reported	homozygous/ heterozygous	2	GAGGGGGCGCTGCC
18		c.4259C>A	n not reported r	heterozygous	1	GACAGTGACCCCCCAATTCCAATTGAC
19		c.4261A>C	not reported	heterozygous	1	GACAGTGACCCCCCAATTCCAATTGAC
Total FP annotations					124	

[†] Nucleotide A of the ATG translation initiation start site of the cDNAs from reference gene sequence is indicated as nucleotide +1.

[‡] In bold are the nucleotide changes.

TABLE 7: MANN-WHITNEY TEST BETWEEN FA MALES AND FA FEMALES

Sample	Gender	<i>P</i> (<i>FANCB</i> sample / <i>FANCB</i> ♀ control)	<i>P</i> (<i>FANCB</i> sample / <i>FANCB</i> ♂ control)
All males	♂	<.0001	1
All females	♀	1	<.0001
P12	♂	<.0001	.0926
P26	♂	<.0001	.9992
P14	♂	<.0001	.3974
P15	♂	<.0001	.8992
P13	♂	<.0001	1
P25	♂	<.0001	.9237
P7	♂	<.0001	.7416
W1	♂	<.0001	.746
P5	♂	<.0001	.7212
P8	♂	<.0001	.6867
P20	♂	<.0001	.8686
P30	♂	<.0001	.3058
P16	♂	<.0001	.0477
P28	♀	.016	<.0001
P10	♀	.1654	<.0001
P11	♀	.0902	<.0001
P18	♀	.3932	<.0001
P6	♀	.2542	<.0001
P21	♀	.8961	<.0001
P27	♀	.9754	<.0001
W2	♀	.9187	<.0001
P24	♀	.7117	<.0001
P4	♀	.5231	.0011
P23	♀	.4632	<.0001
P9	♀	.1809	<.0001
P29	♀	.7117	<.0001
P19	♀	.7958	<.0001
P17	♀	.0775	<.0001
P22	♀	.0001	<.0001

To exclude that the difference observed between males and females was due to random variations of the *FANCB* coverage, the Mann-Whitney test (distribution of the data was non-normal) was performed. Samples were split in two control groups (males and females) against which we compared all the intersample normalisation ratios of *FANCB* amplicons of each patient. The difference between median of all patients and the median of controls of the opposite gender was statistically significant, while the difference between median of all patients and the median of true gender was not significant, except in samples P16, P22, and P28. Male P16 has a significantly higher median (median=0.56; IQR=0.52-0.62) than the median of male population (median=0.52; IQR=0.44-0.62). Females P22 and P28 had a significantly higher (median=1.28; IQR=1.16-1.41) and lower (median=0.91; IQR=0.78-1.13) median than that of the female population (median=1.05; IQR=0.90-1.24) (Figure 6).

Table 8: QUALITY CONTROL DATA OF 21 IT SAMPLES ACCORDING TO IPGM ANALYSIS.

Sample	N. of bases per sample*	Bases with $Q \geq 20$ (%) [†]	Mapped reads [‡]	Reads on target (%) [§]	Average reads per amplicon [¶]	Uniformity of amplicon coverage (%) ^{¶¶}	Amplicons reading end-to-end (%) [#]
IT1	48,614,966	76.88	380,196	76.50	447.5	80.46	10.92
IT2	110,190,830	84.26	713,456	78.1	857.3	90.62	31.54
IT3	161,154,396	75.96	1,302,367	84.34	1,690	89.38	8.15
IT4	40,935,105	78.11	309,174	78.53	373.5	84.8	12
IT5	42,371,873	83.85	276,771	71.48	304.4	91.38	32.46
IT6	22,019,831	88.25	149,242	51.02	117.1	91.63	48.92
IT7	171,185,275	83.87	1,100,647	78.51	1,329	87.85	31.69
IT8	62,379,030	85.66	419,581	79.67	514.3	72.97	32.77
IT9	66,773,457	76.50	531,704	80.71	660.2	92	8.15
IT10	51,723,151	75.98	412,712	87.14	553.3	82.57	8.31
IT11	139,139,962	84.02	899,608	73.4	1,016	91.85	33.85
IT12	41,683,159	88.23	248,823	74.14	283.8	91.92	48.15
IT13	54,731,660	76.13	435,859	85.53	573.5	90.46	7.85
IT14	63,099,005	77.34	490,446	65.63	495	91	11.38
IT15	175,304,921	77.52	1,330,528	82.28	1,685	90	11
IT16	19,007,723	84.06	121,246	88.85	165.7	83.8	33.69
IT17	84,513,720	77.82	644,389	84.35	836.2	89.38	11.08
IT18	40,613,036	84.24	262,707	79.02	319.4	72.48	33.08
IT19	76,364,384	84.71	489,140	87.16	655.9	87.69	32.15
IT20	76,286,614	76.16	606,098	82.78	771.9	87.54	8.77
IT21	54,890,008	84.08	357,874	66.72	367	92.32	31
Min	19,007,723	75.96	121,246	51.02	117.1	72.48	7.85
Max	175,304,921	88.25	1,330,528	88.85	1,690	92.32	49
Mean	76,332,481	81.13	546,789	77.90	667.45	87.23	23.21
St. dev	47,447,684	4.38	347,263	8.84	445.55	5.86	13.96
Median	62,379,030	84	435,859	79	553	89	31
1 st quartile	42,371,873	76.88	309,174	74.14	367.3	84.80	10.92
3 rd quartile	84,513,720	84.24	644,389	84.34	836.2	91.38	32.77

*Total number of filtered and trimmed reads independent of length post filtering bases per barcode.

[†]The percentage of reads that have a predicted quality score of Q20 or better. A Q20 score is the predicted quality of a Phred-like score of 20 or better, or one error in 100 bp.

[‡]Total number of reads mapped to the reference.

[§]The percentage of reads mapped to any targeted region relative to all reads mapped to the reference.

[¶]The average number of reads assigned to amplicons. The average number of reads assigned to amplicons.

^{¶¶}The percentage of bases in all targeted regions covered by at least 0.2x the average base read depth.

[#]The percentage of all amplicons that were considered to have a sufficient proportion of assigned reads (70%) that covered the whole amplicon target from 'end-to-end'. To allow for error the effective ends of the amplicon region for read alignment are within 2 bases of the actual ends of the region.

Table 9: AMPLICONS NOT COVERED MORE THAN 30X IN IT SAMPLES

Gene	Exon	ID amplicons	N. of samples with coverage <30X
ABCG5	3	AMPL5119879043	4
	4	AMPL1852930860*	20
	7	AMPL5119952114‡	12
ACTN1	1	AMPL476202878†	21
	12	AMPL469584238‡	12
ANKRD26	1	AMPL5120483301	4
	3	AMPL680614872	4
	24	AMPL3819594782	3
	24	AMPL3819794179	8
	24 (i)	AMPL4037050642	5
	28	AMPL3817684916	5
	29 (i)	AMPL4037001163	4
FLNA	1	AMPL1528530537	16
	6/7	AMPL1525570810	4
	24	AMPL4041039780*	21
	39	AMPL1524945354	8
	41	AMPL1528095906	4
	47	AMPL891610539	3
GATA1	1	AMPL547618158‡	21
	2	AMPL5119257697	3
	5	AMPL5119891818*	21
	5	AMPL547636375*	21
GP1BA	1	AMPL2608384287‡	21
	1	AMPL5120151981	10
GP1BB	1	AMPL5120107016*	21
	1	AMPL5120137731*	10
	1	AMPL5120138512	21
GP9	1	AMPL5102829037*	21
	1	AMPL5120134312*	20
HOXA11	1	AMPL1617532402*	19
ITGA2B	5	AMPL2606141536	4
	20	AMPL5120255688	6
	21	AMPL3949052660	4
	27	AMPL2605653688‡	13
MPL	1	AMPL3877386620	4
	3	AMPL410828843*	21
	7	AMPL405671894	3
	9	AMPL5119412286	9

MYH9	1	AMPL5120268995 [‡]	15
	5i	AMPL763436336	3
	23	AMPL762419900	7
	24	AMPL1801239905	3
	27	AMPL5119178104	5
	29	AMPL763222728	7
	29	AMPL763233917*	21
NBEAL2	1	AMPL3762892822	5
	3	AMPL5119619893	4
	16	AMPL5119576873	4
	26	AMPL2607417535	4
	33	AMPL2579769787	3
	36	AMPL2608128223*	21
	53/54	AMPL2585118671*	20
	54	AMPL5102380983*	16
RUNX1	5'utr	AMPL736666055*	18
	6	AMPL5116682733*	18
WAS	10	AMPL5119468679*	21

* Amplicons covered less than 30X because of potential self-annealing or intra-strand loop formation of primers.

[†] Amplicons characterized by the presence of a GC rich region (AMPL476202878, 77%) and repetitive sequences (AMPL5120138512), which interfere with the amplification efficiency.

[‡] Amplification efficiency is influenced by the Temperature of Annealing that results higher than used for the amplification reaction used for the library preparation

For the remaining amplicons the efficiency of amplification is worse than that of the other amplicons, thus when the overall efficiency of the libraries is low, the amplicon coverage become <30X








Table 10: IT RARE VARIANTS (MAF<0.01) CONSIDERED IN THE STUDY.

No	Gene	variant analyzed [†]	rs number (MAF in 1000 Genome databse)	Pathogenicity	N. of samples with variant	Confirmed by Sanger sequencing
1	ABCG5 (NM_022436.2)	c.1878T>C p.(=)	nr	synonym (0)	1	not analyzed
2	ABCG8 (NM_022437.2)	c.1924G>A p.(Ala642Thr)	rs113005049 (nr)	missense (1)	1	not analyzed
3		c.1212-7T>A	nr	splicing	1 (IT4)	yes
4		c.1941C>G p.(=)	rs147991100 (nr)	synonym (2)	1	not analyzed
5		c.1201A>T p.(Thr401Ser)	rs144200355 (0.0006)	missense (4)	1	not analyzed
6	ACTN1 (NM_001102.3)	c.751G>A p.(Gly251Arg)	nr	missense (7)	1 (known mutation in IT1)	yes
7		c.1385+89A>C	rs566944399 (<0.01)	generation of new potential acceptor splice site: score 67.04	1 (IT14)	yes
8		c.2210C>A p.(Thr737Asn)	nr	missense (7)	1 (known mutation in IT2)	yes
9		c.2133+44del	nr	deletion (controlla patogenicità)	6	no
10	ANKRD26 (NM_014915.2)	c.-127A>T	nr	present in public mutation database [§]	1 (known mutation in IT3)	yes
11		c.-134G>A	nr	present in public mutation database [§]	1 (known mutation in IT4)	yes
12		c.1269+7T>A	rs201014646 (nr)	no differences between wt and mut	1	not analyzed
13		c.1461+26G>C	nr	no differences between wt and mut	1	not analyzed
14		c.4086-80A>G	nr	no differences between wt and mut	1	not analyzed
15	CYCS (NM_018947.5)	c.154G>A p.(Ala52Thr)	rs573366932 (<0.01)	missense (7)	1 (known mutation in IT7)	yes



16	FLNA (NM_001110556.1)	c.355A>G p.(Ile119Val)	nr	missense (5)	3	no	←
17		c.1567+45delC	nr	small deletion	3	no	←
18		c.3980-34C>G	nr	no differences between wt and mut	1	not analyzed	
19		c.6735C>G p.(=)	nr	synonym (2)	8	not analyzed	
20		c.7686C>T p.(=)	rs76337075 (nr)	synonym (2)	1	not analyzed	
21		c.7157-35_37del	nr	deletion	7	no	←
22	GATA1 (NM_002049.3)	c.647G>A p.(Arg216Gln)	rs104894809 (nr)	missense present in public mutation database [‡]	1 (known mutation in IT8)	yes	←
23	GP1BA (NM_000173.5)	c.92T>C p.(Val31Ala)	rs201827537 (nr)	missense (4)	1	not analyzed	
24		c.104del p.(Lys35Argfs*4)	nr	present in public database [¤]	1(known mutation in IT9)	yes	←
25		c.206C>T p.(Pro69Leu)	rs138825640 (0.0008)	missense (1)	1	not analyzed	
26		c.1806T>G p.(=)	nr	synonym (2)	4	not analyzed	
27	GP1BB (NM_000407.4)	c.47T>C p.(Leu16Pro)	nr	missense present in public mutation database [¤]	1 (known mutation in IT10)	yes	←
28	GP9 (NM_000174.3)	c.236C>T p.(Thr79Ile)	rs200640594 (<0.01)	missense (0)	1	not analyzed	
29	HOXA11 (NM_005523.5)	c.455A>G p.(Glu152Gly)	nr	missense (4)	1	not analyzed	
30	ITGA2B (NM_000419)	c.998G>T p.(Gly333Val)	nr	missense (9)	2	no	←
31		c.2602G>A p.(Val868Met)	rs74988902 (<0.01)	missense (4)	1	not analyzed	
32		c.2613del p.(Leu872Cysfs*38)	nr	frameshift	1	no	←
33	ITGB3 (NM_000212.2)	c.2134+1G>C	rs398122373 (nr)	present in public mutation database [‡]	1 (known mutation in IT12)	yes	←
34		c.2184G>A p.(=)	rs376459842 (nr)	synonym (2)	1	not analyzed	
35	MPL (NM_005373.2)	c.235_236del p.(Leu79Glufs*84)	nr	present in public mutation database [*]	1 (known mutation in IT6)	yes	←
36		c.1653delG p.(Lys553Argfs*77)	nr	present in public mutation database [#]	1 (known mutation in IT6)	yes	←

37	MYH9 (NM_002473.4)	c.230C>G p.(Pro77Arg)	nr	missense (9)	1	no	←
38		c.705+11C>T	rs201738304 (<0.01)	no differences between wt and mut	1	not analyzed	
39		c.1083C>T p.(=)	rs56001030 (<0.01)	synonym (0)	1	not analyzed	
40		c.2054C>T p.(Pro685Leu)	nr	missense (7)	1 (IT18)	yes	←
41		c.2114G>A p.(Arg705His)	rs80338828 (nr)	missense present in public mutation database [‡]	2 (known mutation in IT14 and IT15)	yes	←
42		c.3837+24C>T	nr	no differences between wt and mut	2	not analyzed	
43		c.4270G>A p.(Asp1424Asn)	rs80338831 (nr)	present in public mutation database [¶]	1	no	←
44		c.4359G>A p.(=)	rs202127454 (<0.01)	synonym (2)	1	not analyzed	
45		c.5026A>G p.(Lys1676Glu)	rs138158369 (nr)	missense (7)	1 (IT14)	yes	←
46		c.5143G>A p.(Gly1715Ser)	rs148109368 (<0.01)	missense (5)	1 (IT5)	yes	←
47		c.5521G>A p.(Glu1841Lys)	rs80338834 (nr)	missense present in public mutation database [‡]	1 (known mutation in IT16)	yes	←
48		c.5393A>C p.(Glu1798Ala)	nr	missense (8)	14	no	←
49	c.5797C>T p.(Arg1933*)	rs80338835 (nr)	nonsense present in public mutation database [‡]	1 (new mutation found in IT21)	yes	←	

50	NBEAL2 (NM_015175.2)	c.1898+33C>T	rs201762779 (<0.01)	generation of new potential donor splice site: score 77.42	1 (IT12)	yes	
51		c.2187C>A p.(Tyr729*)	nr	nonsense	1 (known mutation in IT17)	yes	
52		c.2473+53C>A	nr	generation of new potential acceptor splice site: score 78.24/ potential loss of acceptor splice site: score 63.75	3 (IT10, IT4, IT20)	yes	
53		c.2556+33G>A	rs201426707 (0.0004)	no differences between wt and mut	1	not analyzed	
54		c.4530G>C p.(Glu1510Asp)	nr	missense (3)	1	not analyzed	
55		c. 8027+5G>C	nr	potential loss of acceptor splice site, score:39.47/potential loss of donor site, score:74.8	3	no	
56		8027+6C>G	nr	generation of new potential acceptor site, score:72.58	2	no	
57	RBM8A (NM_005105.3)	c.67+32G>C	nr	present in public mutation database ¹	1 (IT9)	yes	
58	RUNX1 (NM_001001890. 2)	c.270+95delT	rs568042691 (<0.01)	generation of a new potential acceptor site, score:71.91/ potential loss of acceptor site, score: 41.97/ generation of new potential donor site, score: 66.48	1 (IT5)	yes	
59		c.725-48A>T	nr	no differences between wt and mut	1	not analyzed	
60	TUBB1 (NM_030773)	c.111C>T p.(=)	rs150453159 (<0.01)	synonym (2)	1	not analyzed	

61	VWF (NM_000552.3)	c.220+57del	nr	small deletion	1	no	←
62		c.220+71G>A	rs531709650 (<0.01)	potential loss of acceptor splice site, score:39.1	1 (IT8)	yes	←
63		c.1614C>T p.(=)	rs138268387 (<0.01)	synonym (2)	1	not analyzed	
64		c.2821-41G>T	rs113608895 (<0.01)	no differences between wt and mut	1	not analyzed	
65		c.4195G>A p.(Arg1399His)	rs1800382	missense (8)	1	no	←
66		c.5049A>C p.(=)	rs200368994 (nr)	synonym (0)	1	not analyzed	
67		c.6303C>A p.(=)	rs115914543 (<0.001)	synonym (2)	1	not analyzed	
68		c.6798+11G>A	rs139999346 (<0.01)	generation of new potential acceptor splice site: score 72.53	1(IT21)	yes	←
60		c.7548+66C>G	nr	generation of new potential donor site, score: 73.86	1(IT8)	yes	←
70	WAS (NM_000377.2)	c.1349-18G>A	rs370010448 (nr)	no differences between wt and mut	1	not analyzed	

§present in CLINVAR and OMIM

*present in CLINVAR, OMIM and HGVD

*present in CLINVAR, HGMD and LOVD

#present in HGMD

‡present CLINVAR, LOVD, OMIM and HGVD

¶present in CLINVAR

Red arrows indicate the 75 (40 different) variants validate with Sanger sequencing because considered potentially pathogenetic due to their presence in mutation database or their high pathogenicity score (see materials and methods). The 30 (27 different) pathogenetic variants confirmed by Sanger sequencing are highlighted in blue.

Table 11: MANN-WHITNEY TEST BETWEEN IT MALES AND IT FEMALES

Sample	Gender	FLNA		GATA1		WAS	
		<i>P</i> (FLNA sample / FLNA ♂ control)	<i>P</i> (FLNA sample / FLNA ♀ control)	<i>P</i> (GATA1 sample / GATA1 ♂ control)	<i>P</i> (GATA1 sample / GATA1 ♀ control)	<i>P</i> (WAS sample / WAS ♂ control)	<i>P</i> (WAS sample / WAS ♀ control)
ALL MALE	♂	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000
ALL FEMALE	♀	0.0000	1.0000	0.0000	1.0000	0.0000	1.0000
IT1	♂	0.0347	<0,0001	0.8341	0.0001	0.9033	<0,0001
IT2	♂	0.0632	<0,0001	1.0000	0.0001	0.3721	<0,0001
IT5	♂	<0,0001	<0,0001	0.0594	0.0001	0.0305	<0,0001
IT7	♂	0.0146	<0,0001	0.4980	0.0001	0.1555	<0,0001
IT8	♂	<0,0001	<0,0001	0.0501	0.0001	<0,0001	<0,0001
IT9	♂	0.0274	<0,0001	0.4091	0.0001	0.0251	<0,0001
IT10	♂	<0,0001	<0,0001	0.6663	0.0001	0.6097	<0,0001
IT14	♂	0.5737	<0,0001	0.9117	0.0001	0.4895	<0,0001
IT17	♂	0.4005	<0,0001	0.6308	0.0001	0.4104	<0,0001
IT18	♂	<0,0001	<0,0001	0.6663	0.0086	0.2363	0.0001
IT19	♂	0.2665	<0,0001	0.6843	0.0001	0.4386	<0,0001
IT20	♂	0.0001	<0,0001	0.8534	0.0001	0.9363	<0,0001
IT3	♀	<0,0001	<0,0001	0.0001	0.4011	<0,0001	0.0473
IT4	♀	<0,0001	<0,0001	0.0001	0.4168	<0,0001	0.6275
IT6	♀	<0,0001	0.2821	0.0001	0.7726	<0,0001	0.5371
IT11	♀	<0,0001	0.3843	0.0001	0.6497	<0,0001	0.6331
IT12	♀	<0,0001	0.1439	0.0001	0.3019	<0,0001	0.8008
IT13	♀	<0,0001	0.0066	0.0001	1.0000	<0,0001	0.0173
IT15	♀	<0,0001	0.0035	0.0001	0.6497	<0,0001	0.2522
IT16	♀	<0,0001	<0,0001	0.0048	0.9014	<0,0001	0.3910
IT21	♀	<0,0001	0.1705	0.0001	0.7516	<0,0001	0.8370

Since the distribution of the data was non-normal, we have performed a Mann-Whitney test to exclude that the difference observed between males and females was due to random variations of the *FLNA*, *GATA1* and *WAS* coverage. For each gene, samples were split in two control groups (males and females) against which we compared all the intersample normalisation ratios of *FLNA*, *GATA1* and *WAS* amplicons of each patient. The difference between median of all patients and the median of controls of the opposite gender was statistically significant, while the difference between median of all patients and the median of true gender was not significant, confirming the reliability of the attribution of gender test.

BIBLIOGRAPHY

Albers CA, Paul DS, Schulze H, et al. Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in exon-junction complex subunit RBM8A causes TAR syndrome. *Nature genetics* 2012;44:435-439, S431-432.

Auerbach AD. Fanconi anemia diagnosis and the diepoxybutane (DEB) test. *Experimental hematology* 1993;21:731-733.

Auerbach AD. Fanconi anemia and its diagnosis. *Mutation research* 2009;668:4-10.

Balduini CL, Cattaneo M, Fabris F, et al. Inherited thrombocytopenias: a proposed diagnostic algorithm from the Italian Gruppo di Studio delle Piastrine. *Haematologica* 2003;88:582-592.

Balduini CL, Pecci A, Noris P. Inherited thrombocytopenias: the evolving spectrum. *Hamostaseologie* 2012;32:259-270.

Balduini CL, Pecci A, Noris P. Diagnosis and management of inherited thrombocytopenias. *Seminars in thrombosis and hemostasis* 2013;39:161-171.

Balduini CL, Savoia A. Genetics of familial forms of thrombocytopenia. *Human genetics* 2012;131:1821-1832.

Bogliolo M, Schuster B, Stoepker C, et al. Mutations in ERCC4, encoding the DNA-repair endonuclease XPF, cause Fanconi anemia. *American journal of human genetics* 2013;92:800-806.

Bottega R, Marconi C, Faleschini M, et al. ACTN1-related thrombocytopenia: identification of novel families for phenotypic characterization. *Blood* 2015;125:869-872.

Bottega R, Pecci A, De Candia E, et al. Correlation between platelet phenotype and NBEAL2 genotype in patients with congenital thrombocytopenia and alpha-granule deficiency. *Haematologica* 2013;98:868-874.

Bragg LM, Stone G, Butler MK, et al. Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. *PLoS computational biology* 2013;9:e1003031.

Buermans HP, den Dunnen JT. Next generation sequencing technology: Advances and applications. *Biochimica et biophysica acta* 2014;1842:1932-1941.

Castella M, Pujol R, Callen E, et al. Origin, functional role, and clinical impact of Fanconi anemia FANCA mutations. *Blood* 2011;117:3759-3769.

De Rocco D, Bottega R, Cappelli E, et al. Molecular analysis of Fanconi anemia: the experience of the Bone Marrow Failure Study Group of the Italian Association of Pediatric Onco-Hematology. *Haematologica* 2014;99:1022-1031.

Gross M, Hanenberg H, Lobitz S, et al. Reverse mosaicism in Fanconi anemia: natural gene therapy via molecular self-correction. *Cytogenetic and genome research* 2002;98:126-135.

Johnsen JM, Nickerson DA, Reiner AP. Massively parallel sequencing: the new frontier of hematologic genomics. *Blood* 2013;122:3268-3275.

Kottemann MC, Smogorzewska A. Fanconi anaemia and the repair of Watson and Crick DNA crosslinks. *Nature* 2013;493:356-363.

Kunishima S, Okuno Y, Yoshida K, et al. ACTN1 mutations cause congenital macrothrombocytopenia. *American journal of human genetics* 2013;92:431-438.

Loman NJ, Misra RV, Dallman TJ, et al. Performance comparison of benchtop high-throughput sequencing platforms. *Nature biotechnology* 2012;30:434-439.

Machlus KR, Italiano JE, Jr. The incredible journey: From megakaryocyte development to platelet formation. *The Journal of cell biology* 2013;201:785-796.

Manchev VT, Hilpert M, Berrou E, et al. A new form of macrothrombocytopenia induced by a germ-line mutation in the PRKACG gene. *Blood* 2014;124:2554-2563.

McClellan J, King MC. Genetic heterogeneity in human disease. *Cell* 2010;141:210-217.

Merriman B, Ion Torrent R, Team D, et al. Progress in ion torrent semiconductor chip based sequencing. *Electrophoresis* 2012;33:3397-3417.

Moldovan GL, D'Andrea AD. How the fanconi anemia pathway guards the genome. *Annual review of genetics* 2009;43:223-249.

Morgan NV, Tipping AJ, Joenje H, et al. High frequency of large intragenic deletions in the Fanconi anemia group A gene. *American journal of human genetics* 1999;65:1330-1341.

Patel SR, Hartwig JH, Italiano JE, Jr. The biogenesis of platelets from megakaryocyte proplatelets. *The Journal of clinical investigation* 2005;115:3348-3354.

Pippucci T, Savoia A, Perrotta S, et al. Mutations in the 5' UTR of ANKRD26, the ankirin repeat domain 26 gene, cause an autosomal-dominant form of inherited thrombocytopenia, THC2. *American journal of human genetics* 2011;88:115-120.

Rothberg JM, Hinz W, Rearick TM, et al. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 2011;475:348-352.

Savoia A, Dufour C, Locatelli F, et al. Congenital amegakaryocytic thrombocytopenia: clinical and biological consequences of five novel mutations. *Haematologica* 2007;92:1186-1193.

Soulier J. Fanconi anemia. *Hematology / the Education Program of the American Society of Hematology American Society of Hematology Education Program* 2011;2011:492-497.

Stevenson WS, Morel-Kopp MC, Ward CM. Platelets are not all gray in GFI1B disease. *Clinical genetics* 2014.

Tischkowitz MD, Hodgson SV. Fanconi Anemia. *Journal of Medical Genetics* 2003.

Verver E, Pecci A, De Rocco D, et al. R705H mutation of MYH9 is associated with MYH9-related disease and not only with non-syndromic deafness DFNA17. *Clinical genetics* 2014.

Waisfisz Q, Morgan NV, Savino M, et al. Spontaneous functional correction of homozygous fanconi anaemia alleles reveals novel mechanistic basis for reverse mosaicism. *Nature genetics* 1999;22:379-383.

Zhang MY, Churpek JE, Keel SB, et al. Germline ETV6 mutations in familial thrombocytopenia and hematologic malignancy. *Nature genetics* 2015;47:180-185.