Diagnostic Single Gene Analyses Beyond Sanger

Economic high-throughput sequencing of small genes involved in congenital coagulation and platelet disorders

Juliane Najm¹, Matthias Rath¹, Winnie Schröder¹, Ute Felbor¹

¹Department of Human Genetics, University Medicine Greifswald, and Interfaculty Institute of Genetics and Functional Genomics, University of Greifswald, Greifswald, Germany


Address for correspondence: Prof. Dr. Ute Felbor, MD, Department of Human Genetics, University Medicine Greifswald, Fleischmannstr. 43, 17475 Greifswald, Germany (e-mail: felbor@uni-greifswald.de).

Keywords
► congenital coagulation and platelet disorders
► LR-PCR
► NGS
► high throughput sequencing

Summary
Molecular testing of congenital coagulation and platelet disorders offers confirmation of clinical diagnoses, supports genetic counselling, and enables predictive and prenatal diagnosis. In some cases, genotype-phenotype correlations are important for predicting the clinical course of the disease and adaptation of individualized therapy. Until recently, genotyping has been mainly performed by Sanger sequencing. While next generation sequencing (NGS) enables the parallel analysis of multiple genes, the cost-value ratio of custom-made panels can be unfavorable for analyses of specific small genes. The aim of this study was to transfer genotyping of small genes involved in congenital coagulation and platelet disorders from Sanger sequencing to an NGS-based method. A LR-PCR approach for target enrichment of the entire genomic regions of the genes F7, F10, F11, F12, GATA1, MYH9, TUBB1 and WAS was combined with high-throughput sequencing on a MiSeq platform. NGS detected all variants that had previously been identified by Sanger sequencing. Our results demonstrate that this approach is an accurate and flexible tool for molecular genetic diagnostics of single small genes.

Zusammenfassung
Die molekulare Diagnostik hereditärer Gerinnungs- und Thrombozytenstenosen ermöglicht die Bestätigung einer klinischen Verdachtsdiagnose, ist aber auch für die genetische Beratung und die prädiktive und pränatale Diagnostik notwendig. In einigen Fällen ist eine Genotyp-Phänotyp-Korrelation für die Prognose des klinischen Krankheitsverlaufs wichtig und erlaubt eine individuelle Therapieplanung. Im
Hämorrhagische Diathesen können durch Mutationen in primären Hämostasemustern (u. a. Blutgerinnungsstörungen), sekundären hämostasemustern (z. B. Gerinnungsstörungen) oder Plättchenstörungen verursacht werden. Sie präsentieren sich in einem heterogenen Muster von Blutungsvorfällen mit unterschiedlichen klinischen Manifestationen, von leichten bis hin zu schweren Verläufen.


**Material and Methods**

**DNA Isolation and Conventional Mutation Analyses**

DNA Samples of all 43 study probands were extracted with written informed consent according to the German Gene Diagnostic Act from peripheral blood lymphocytes using standard techniques. The coding regions and adjacent exon-intron boundaries (±50 bp) of the following genes were amplified by PCR:

- F7 (Locus Reference Genomic sequence LRG_554; transcript LRG_554t1; number of probands [n] = 6)
- F10 (LRG_548, LRG_548t1; n = 5)
- F11 (LRG_583, LRG_583t1; n = 3)
- F12 (LRG_145, LRG_145t1; n = 5)
- GATA1 (LRG_559, LRG_559t1; n = 5)
- TUBB1 (LRG_581, LRG_581t1; n = 6)
- WAS (LRG_125, LRG_125t1; n = 5)
- MYH9 (LRG_567, LRG_567t1; n = 8)

**Target Enrichment and High-Throughput Sequencing**

Primers for LR-PCRs were designed with Primer3 (v4.0.0; http://bioinfo.ut.ee/primer3/) and Table S1. PrimerSTAR® GXL DNA polymerase (Takara Biotech, Otsu, Japan). Sequencing data were analyzed with the SeqPilot software (Version 4.3.1, JSI Medical Systems, Ettenheim, Germany). SALSA MLPA Kits P207, P440, P432, and P432 were used for detection of copy number variations in F7, F10, F11, and MYH9 (MRC-Holland, Amsterdam, The Netherlands).

**Long-range PCR (LR-PCR)**

LR-PCR has the advantage that it does not require a customized design by commercial vendors. In this proof-of-principle study, we intended to establish LR-PCR target enrichment for single gene analyses of F7, F10, F11, F12, GATA1, TUBB1, and WAS on a MiSeq platform. This approach proved to be highly flexible, also appropriate for even larger target regions such as MYH9.
Equimolar amounts of the purified PCR amplicons were combined in 14 pools containing non-overlapping target regions of up to seven individual probands. The Nextera XT kit was used to prepare sequencing libraries with 1 ng of each pool (Illumina®, San Diego, USA). Individually barcoded DNA libraries were combined and sequenced on a MiSeq instrument with 2 × 150 or 2 × 250 cycles (Reagent Nano Kit v2 or Reagent Kit v3; Illumina®, San Diego, USA).

Bioinformatics Analyses

The MiSeq Reporter Software was used for demultiplexing and primary data analysis (MSR v2.5.1; Illumina®, San Diego, USA). The quality of sequencing reads was analyzed with the FASTQC toolkit (http://www.bioinformatics.babraham.ac.uk/index.html) and sequencing coverage was visualized with the GVIZ package for R software (https://www.r-project.org/). The SeqNext module of SeqPilot software (JSI Medical Systems) was used for read alignment against the human reference assembly GRCh37/hg19 and variant calling. Only coding regions, exon-intron boundaries (±50 bp), and known promoter regions of the genes were analyzed (diagnostic region of interest). 30 × was defined as the minimum diagnostic sequencing depth.

Results

Fast and Comprehensive Enrichment of Various Genomic Target Regions with LR-PCR

With only minor optimization of the reaction conditions, the entire genomic loci of F7 (14.9 kb), F10 (26.7 kb), F11 (23.8 kb), F12 (14.4 kb), GATA1 (7.8 kb), TUBB1 (7.4 kb), and WAS (14.8 kb) could be successfully amplified by LR-PCR. Only one or two PCR amplicons were necessary for a specific and highly efficient enrichment of each gene (►Fig. 1, ►Table S1). Target enrichment, purification of the PCR products, pooling, and library preparation for 35 probands were completed in only two regular working days. Due to its large size of 106.8 kb, the entire genomic region of the MYH9 gene was independently amplified without any PCR dropouts for another eight probands with eight overlapping LR-PCR amplicons.

![Fig. 1](image.png)

**Fig. 1** Comprehensive coverage and characteristic, gene-specific patterns of sequencing depth for the entire genomic target regions. The genomic loci of F7, F10, F11, F12, GATA1, TUBB1, WAS and MYH9 were amplified with 17 LR-PCR amplicons and separated by gel electrophoresis. The robust and specific amplification of all amplicons is shown for a healthy control (upper left subpanel). Representative gene-specific patterns of read depth across the entire genomic regions are shown as coverage plots (dark blue). Horizontal bold and broken lines indicate a read depth of 6000 × and 3000 ×, respectively. Chromosome ideograms with the cytogenetic locations of the genes (bold green lines) are depicted in the upper part of each subpanel. Exon-intron structures of their respective reference transcripts are shown in the lower parts. The size of its entire genomic region is given below each gene. kb = 1000 base pairs.
High-throughput sequencing of the pooled DNA libraries produced an average output of $1.9 \times 10^6$ sequencing reads (SD: $0.3 \times 10^6$) and consistently high mean coverages for the specific target regions (Min: $837 \times$, Max: $3434 \times$, SD: $578 \times$; – Fig. 2). Interestingly, characteristic patterns of sequencing depth were found for each gene after read alignment (– Fig. 1). These unique patterns were recapitulated in all probands tested for the respective genetic loci (– Fig. S1). The NGS data of the pooled DNA libraries were also checked for hints of PCR cross-contaminations, but significant read depths ($>10$ reads) were only seen in regions previously enriched by PCR.

**Broad Coverage of Coding Regions and Exon-Intron Boundaries**

The mean coverage for the coding regions, canonical splice sites, exon-intron boundaries ($\pm 50$ bp) and known promoter regions were:
- F7: $1292 \times$,
- F10: $1687 \times$,
- F11: $2583 \times$, and
- F12: $1455 \times$.

The sequencing depths of thrombocytopenia-associated genes were also sufficiently high:
- GATA1: $2175 \times$
- TUBB1: $2768 \times$
- WAS: $1928 \times$
- MYH9: $2420 \times$

Overall, 94% of the cumulative target region was covered with a sequencing depth of more than $30 \times$ (180.6 of 191.6 kb). NGS achieved complete coverage for the diagnostic target regions of F10, F11, F12, GATA1, and TUBB1. Only a single exonic gap in exon 10 of the WAS gene (coverage $<30 \times$; LRG_125t1: c.932_1338) was consistently seen in all five probands sequenced for this genomic locus (– Fig. 1). A minor sequencing problem with a mean coverage of $46 \times$ was also found in exon 36 of MYH9 (LRG_567t1: c.5062_5150). Both regions have a high GC-content of nearly 70%. A complete sequencing dropout was observed for only one of the 107 LR-PCR products sequenced in this study (dropout rate = 0.9%). This single PCR amplicon was included and completely covered in a second sequencing run.

**Reliable, Sensitive and Specific Mutation Detection by NGS**

The main purpose of this study was to establish a novel NGS-based screen for missense, nonsense, splice and small frameshift mutations. For an evaluation of the analytic sensitivity and specificity of this approach, DNA samples of 43 probands previously screened with conventional Sanger sequencing were re-analyzed with LR-PCR and NGS. All 25 known pathogenic or likely pathogenic variants (– Table 1) and 128 polymorphisms in the eight analyzed genes were re-identified. As no false-positives were detected in the regions of interest, analytical sensitivity and specificity of this NGS approach were calculated with 100% each.

Interestingly, two large deletions previously found using multiplex ligation-dependent probe amplifications (MLPA) could be re-identified by analyses of the gene-specific coverage patterns. A homozygous deletion spanning exon 2 of the F7 gene (proband P5; – Fig. S1) and a heterozygous deletion of exon 6 of the F10 gene (P7; – Fig. 3) were consistently detected. NGS sequence analysis gave a rather precise estimation of the respective size of the F7 [4.35 kb; c.64 + 430.131-6delinsTCTGTA] and F10 [approximately 4.8 kb; c.(502 + 1_503-1)_747 + 1_748-1] deletions but could not fine-map the breakpoints down to single nucleotide level.
Table 1 | Distinct pathogenic and likely pathogenic variants included in this study. Variants identified in more than one proband are only listed once

<table>
<thead>
<tr>
<th>Gene</th>
<th>Nucleotide Change</th>
<th>Protein Change</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>F7</td>
<td>c.911C&gt;T</td>
<td>p.(Ala304Val)</td>
<td>Tamary et al. 1996 18</td>
</tr>
<tr>
<td></td>
<td>c.1061C&gt;T</td>
<td>p.(Ala354Val)</td>
<td>Bernardi et al. 1994 19</td>
</tr>
<tr>
<td></td>
<td>c.1391delC</td>
<td>p.(Pro464Hisfs532)</td>
<td>Arbini et al. 1994 20</td>
</tr>
<tr>
<td></td>
<td>c.64 + 430_131-6delinsTCGTA</td>
<td>p.7</td>
<td>Rath et al. 2015 21</td>
</tr>
<tr>
<td>F10</td>
<td>c.413A&gt;T</td>
<td>p.(Gln138Leu)</td>
<td>Rath et al. 2015 21</td>
</tr>
<tr>
<td></td>
<td>c.979C&gt;T</td>
<td>p.(Arg327Trp)</td>
<td>Millar et al. 2000 22</td>
</tr>
<tr>
<td></td>
<td>c.1097G&gt;A</td>
<td>p.(Arg366His)</td>
<td>Rath et al. 2015 21</td>
</tr>
<tr>
<td></td>
<td>c.1159C&gt;T</td>
<td>p.(Arg387Cys)</td>
<td>Hermann et al. 2006 23</td>
</tr>
<tr>
<td></td>
<td>c.1247A&gt;C</td>
<td>p.(Glu416Pro)</td>
<td>unpublished</td>
</tr>
<tr>
<td></td>
<td>deletion of exon 6 [c.(502 + 1_503−1)(747 + 1_748−1)]</td>
<td>p.7</td>
<td>Hainmann et al. 2009 24</td>
</tr>
<tr>
<td>F11</td>
<td>c.644_649delTCGACA</td>
<td>p.(Ile215_Asp216del)</td>
<td>Zadra et al. 2004 25</td>
</tr>
<tr>
<td></td>
<td>c.803G&gt;A</td>
<td>p.(Arg268His)</td>
<td>Duncan et al. 2008 26</td>
</tr>
<tr>
<td>F12</td>
<td>c.-57G&gt;C</td>
<td>p.7</td>
<td>Hofferbert et al. 1996 27</td>
</tr>
<tr>
<td></td>
<td>c.-62C&gt;T</td>
<td>p.7</td>
<td>Lombardi et al. 2008 28</td>
</tr>
<tr>
<td></td>
<td>c.1668delC</td>
<td>p.(Asp557Metfs107)</td>
<td>unpublished</td>
</tr>
<tr>
<td></td>
<td>c.1681G&gt;A</td>
<td>p.7</td>
<td>Schloesser et al. 1993 30</td>
</tr>
<tr>
<td>GATA1</td>
<td>c.622G&gt;A</td>
<td>p.(Gly208Arg)</td>
<td>Del Vecchio et al. 2005 31</td>
</tr>
<tr>
<td>WAS</td>
<td>c.101delG</td>
<td>p.(Arg34Hisfs11)</td>
<td>unpublished</td>
</tr>
<tr>
<td></td>
<td>c.256C&gt;T</td>
<td>p.Arg86Cys</td>
<td>Kolluri et al. 1995 32</td>
</tr>
</tbody>
</table>

Fig. 3 | Identification of a known exon-spanning heterozygous deletion in F10. Analysis of the F10 coverage pattern of proband P7 (A) compared with the characteristic control pattern (B) revealed an approximately 4.8 kb spanning heterozygous deletion (indicated by red broken lines) which includes exon 6 of F10 and part of the flanking introns. The exon-intron structure of the reference transcript is shown below. Result of F10 MLPA analysis for P7 also shows the heterozygous deletion of exon 6 (red arrow). Green bars indicate normalized gene dosages of P7 compared with three healthy controls (grey bars). F10-specific probes are depicted on the left, reference probes on the right.
Discussion

The implementation of NGS in most laboratories has already changed standard workflows. Nevertheless, NGS can also be challenging for diagnostic laboratories. Its high sequencing capacity often implies the economic need for parallel analyses of large target regions or a higher number of patients. These requirements can usually be best met with capture-based NGS multigene panels. Hybridization probes minimize the risk of allele dropouts, can be easily combined and are perfectly suitable for multiplexing. Technically, all genes tested in this study could also have been covered with comprehensive capture-based panels. However, if a diagnosis can most likely be confirmed by analysis of a single small disease gene such as F7 or F10, additional focused strategies are more cost-effective.

As an alternative to time-consuming conventional Sanger sequencing, LR-PCR combined with NGS is advantageous in this context since PCR primers are less expensive than hybridization probes and PCR reactions can be realized with standard laboratory skills and equipment. Additionally, there is only a limited need for PCR optimization, and there are a lot of safety points in the workflow which allow repetition of a failed reaction without extensive additional costs. Targeted single gene analyses by LR-PCR also reduce the risk of incidental findings and have higher enrichment specificities for targets with pseudogenes or repetitive regions compared to capture-based multigene panels. Analogous strategies for the enrichment of en bloc genomic target regions have been described not only for BRCA1- and BRCA2-associated hereditary cancer predisposition syndromes and HLA genotyping but also for molecular analysis in hemophilia A.

Our study further expands the spectrum of genes that can be efficiently analyzed by LR-PCR amplification and NGS in a diagnostic context.

With an analytic sensitivity and specificity of 100%, our approach proved to be highly reliable. Furthermore, it is also quite flexible as different target regions can be individually combined. Sequencing gaps or regions with a relatively low read depth that were observed in this study in WAS and MYH9 are mainly restricted to regions with a high GC-content. Those regions are known to cause trouble in NGS sequencing. Currently they still need to be completed by Sanger sequencing.

Obviously, the specific assay for the eight genes studied here might not be suitable for all diagnostic laboratories. However, it can be individually adapted and serve as a valuable addition rather than as an alternative to existing NGS multigene panels. Even combined sequencing with DNA samples enriched with other panels is feasible and further decreases sequencing costs.

Of course, PCR-based enrichments are always hampered by the risk of allele dropout due to sequence variants in primer binding sites or large deletions that can result in a selective amplification of a single allele. For example, we and others have noticed difficulties in conventional DNA sequencing of exon 2 of MYH9 with a biased amplification due to a poly-guanine repeat. Thus, several optimization steps may be necessary for each PCR to find the best primer binding sites. Nonetheless, one has to keep in mind that allele dropout could never be completely excluded and is hard to detect by Sanger sequencing because heterozygous SNPs are often missing in amplified regions due to their small size. LR-PCR significantly increases the chance of finding a heterozygous SNP in the genomic target region that can be used to exclude an allele dropout. Therefore, a screen for heterozygous positions has to be an integral part of data analysis if PCR-based enrichment strategies are chosen.

Multiplex ligation-dependent probe amplification (MLPA) analysis still is the method of choice for copy number detection in routine diagnostics. This technique requires high quality DNA of patients as well as controls and a separate laboratory workflow.

The LR-PCR approach described here can replace the standard method to speed up the overall diagnostic process in particular in male patients with X-linked diseases such as GATA1- and WAS-associated disorders.

A hemizygous intragenic deletion can easily be detected with NGS especially in such small genes that are covered with only one LR-PCR amplicon. Even breakpoint mapping to the approximate location within the amplicon is possible in men. But also in female carriers of X-linked diseases and autosomal disorders, a heterozygous intragenic deletion can be detected due to the unique gene specific NGS pattern in particular in male patients (Fig. S1). In these cases LR-PCR might assess the size of a deletion more accurately than MLPA.

One further advantage of a LR-PCR strategy covering the entire genomic region of a specific gene of interest is the possibility to directly search for deep-intronic mutations in patients that had remained mutation-negative after sequencing of the coding regions and invariant splice sites of the respective disease-associated gene. However, this procedure is not remunerable in a diagnostic setting according to the reimbursement catalogue of Germany’s statutory health insurance system (EBM) and can therefore be performed only in a research context.

International research collaborations such as the ThromboGenomics Consortium or the BRIDGE-BPD study have been formed to unravel the genetic basis in probands with congenital bleeding and platelet disorders. They apply comprehensive NGS multigene panels which work well for rather clear phenotypes. For cases without a clear phenotypic lead, whole exome or genome sequencing might be more successful. However, this requires a careful clinical characterization of family members, their participation after informed consent including the management of incidental findings and a well-established bioinformatics pipeline.

Conclusion

Taken together, our LR-PCR approach combined with NGS resulted in a success rate of 100% for the re-identification of variants in eight different exemplary genes involved in congenital coagulation and platelet disorders. In addition...
to multigene panels which are particularly useful for disorders with genetic heterogeneity, it appears to be a practical alternative method to Sanger sequencing for molecular diagnostics of any small gene to confirm a specific clinical diagnosis.

What is known about this topic?
- Molecular testing of genes involved in congenital coagulation and platelet disorders has been routinely done by Sanger sequencing.
- NGS-based molecular testing became remunerable for routine diagnostics of hereditary coagulation and platelet disorders according to the reimbursement catalogue of Germany’s statutory health insurance system (EBM).

What does this paper add?
- We here present a new method to transfer mutation analyses of genes involved in congenital coagulation and platelet disorders from Sanger to NGS-based sequencing.
- The LR-PCR strategy covering the entire genomic region of a specific gene of interest raises the possibility to directly search for deep-intronic mutations in patients that had remained mutation-negative after sequencing of the coding regions.

Conflicts of Interest
The authors stated that they had no interests which might be perceived as posing a conflict or bias.

Acknowledgments
Study participants and their families are thanked for their participation. K. Hartwig is thanked for excellent technical assistance.

References
3 Altmüller J, Budde BS, Nürnberg P. Enrichment of target sequences for next-generation sequencing applications in research and diagnostics. Biol Chem 2014;395(02):231–237