Long-Range PCR and Nanopore Sequencing Method Resolves *F8*, *GBA*, *CYP21A2*, *SMN1*, and *TNXB* Variants Using a Single Streamlined Workflow

Cody Edwards, Bryan Killinger, Theodore Markulin, Monica Roberts, Ryan Routsong, Jon Kemppainen, Adrian Lara, Sarah Statt, Bradley Martin, Gary Latham, and Bradley Hall

Asuragen, a Bio-Techne Brand, Austin, TX

Summary

- **High-prevalence carrier genes, associated with disorders such as Gaucher Disease (GD) and Hemophilia A (HA), include complex structural variants and pseudogenes that confound conventional sequencing methods.**
- **We established a prototype single-platform, multi-gene AmplideX® PCR based nanopore sequencing assay with custom informatics pipelines, that can replace disparate workflows and provide additional information for variant interpretation.**
- **The assay utilizes sequence deconvolution, amplicon read depth, Paralog Specific Variants (PSVs)-based copy number, and machine learning models to automate and streamline identification of key genetic variants specific to each disease.**
- **More than 90% of known variants across the five genes were detected in reference cell-line samples. We also discovered 27 carriers from our presumed normal whole blood testing cohort, consistent with known carrier rates.**

Introduction

There are many “dark” or “camouflaged” regions of the genome that are critically important to assessing risk in population carrier screening (CS). Six of the top ten disorders by reproductive risk to couples can be difficult or intractable using short-read sequencing, and, more broadly, 20.4% of pathogenic/likely pathogenic variants in ClinVar have been reported to be “technically challenging” [1]. Furthermore, analyses of some of the highest prevalence screening targets rely on multiple bespoke non-NGS methods that derail NGS workflows and require highly trained personnel. Carrier screening can be particularly difficult for *GBA*, *CYP21A2*, *TNXB* and *SMN1* due to the presence of highly homologous pseudogenes (*GBAP1*, *CYP21A1P*, *TNXA*, and *SMN2*, respectively). Furthermore, inversions found within *F8* intron 1 and 22 require distinct technologies that often result in subpar detection rates for at-risk couples, with the focus being limited only to more easily detected variants.

Here we describe a single-tube workflow using long-range PCR amplification and long-read nanopore sequencing to detect multiple classes of genetic variation within a multigene panel of these challenging genes. Targeted variants include structural variants (SVs, such as inversions), single-nucleotide variants (SNVs), insertion/deletions (INDELs), and copy number variants (CNVs).

Materials and Methods

We optimized performance of the assay utilizing 90 cell-lines with variants across representative classes and investigated performance by screening 232 whole blood samples. Samples were amplified in multiplexed gene-specific PCR reactions using novel long-range amplification reagents based on AmplideX PCR chemistry. Samples were then barcoded, pooled, prepared by ligation sequencing kit SQK-LSK114 (ONT) and run on R10.4.1 flow cells (ONT) utilizing the MinION Mk1B device. Cell-line samples containing all major classes of variation were used to develop custom data analysis pipelines. Analyses were performed with custom software for sequence deconvolution and Clair3 for SNV/INDEL identification [2]. Performance was demonstrated for cell-line and whole-blood samples across 4 sequencing runs, multiplexing up to 95 samples in a single run. Orthogonal methods or reporting in known databases (e.g., Coriell, 1000 Genomes, custom PCR/capillary electrophoresis (CE), AmplideX® PCR/CE *SMN1/2* Plus Kit†, and various commercially available products) were utilized to determine performance.

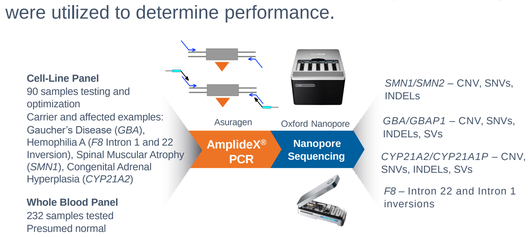

Figure 1. Amplification-based Nanopore Carrier Screening Panel Design and Workflow Identifies Pathogenic Variants for High-Prevalence Carrier Genes that are Problematic for Conventional Sequencing. The combination of proprietary amplification-based technology and nanopore sequencing enables amplification of large highly GC-rich regions, such as those present in *F8*, as well as copy number and structural variant detection of highly homologous sequences, which can be particularly difficult in *GBA*, *CYP21A2*, *TNXB* and *SMN1*.

This product is under development. Future availability and performance to be determined.
†For Research Use Only. Not for use in diagnostic procedures.
All authors have the financial relationship to disclose: Employment by Asuragen.
Presented at ACMG 2023

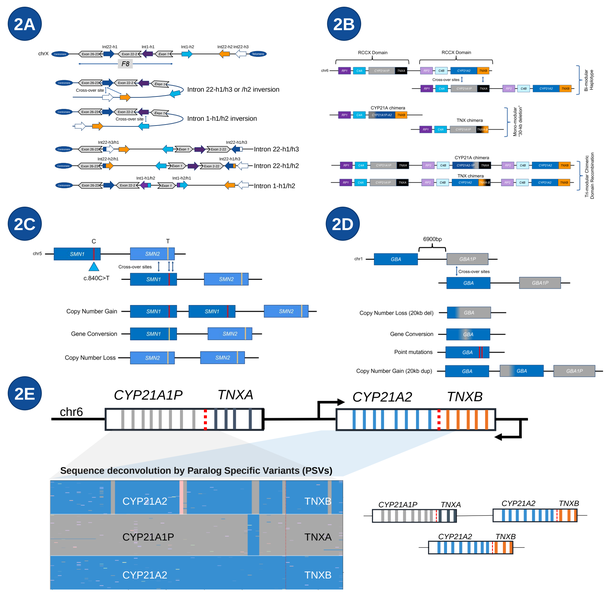

Figure 2. Complex Structural Alterations due to Highly Homologous Regions Make Identification of Carriers Difficult for *F8*, *CYP21A2*, *SMN1*, and *GBA*. A) *F8* introns 1 and 22 can recombine with regions upstream of *F8* (intron 1-H2, intron 22-H2/H3) causing inversion of *F8* exons leading to severe hemophilia A. **B)** The RCCX domain, where *CYP21A2* and *TNXB* reside, can recombine across any of four genes causing copy number alterations and micro-conversions between the genes and pseudogenes (*CYP21A1P* and *TNXA*). **C)** *SMN1* and *SMN2* are > 99% similar, with 1 nucleotide of importance in their coding region (c.840C>T) where the T in exon 7 of *SMN2* causes exon skipping and reduces the amount of functional *SMN2* protein by ~90%. **D.)** *GBA* and its pseudogene *GBAP1* are separated by only 6.9kb, leading to copy number changes and gene conversions. **E)** Methods were developed to deconvolve sequences using paralog specific variants across each of these genes. *CYP21A2*/*CYP21A1P* clustering of HG001 is shown as an example.

Results

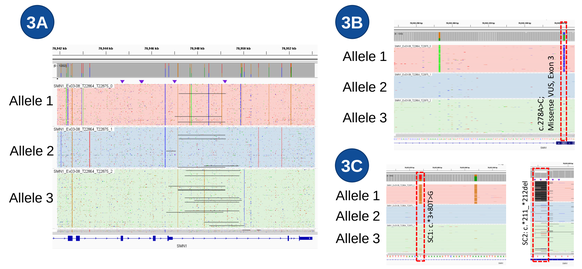

Figure 3. Sequence Deconvolution Followed by Downstream Variant Calling Allows Phasing of Variants to Distinct Copies of *SMN1*. PCR enrichment was performed using two >10 kb amplicons. Independently, *SMN1/2* CN analysis was performed using a larger number of shorter amplicons; see poster P477 for more details. Cell-line HG02818 is shown, which contains **A)** three copies of *SMN1* and one copy of *SMN2* (data not shown). **B)** A rare missense variant of unknown significance (VUS), c.278A>C, in Exon 3 was phased to *SMN1* within allele 1 that also **C)** contains silent carrier mutations, SC1: c.*3+80T>G and SC2: c.*211_*212del.

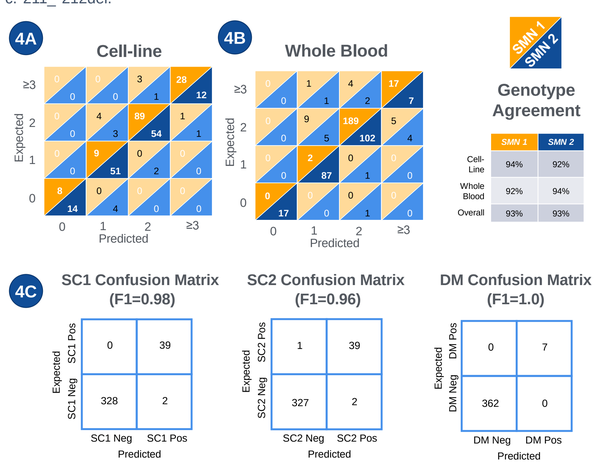

Figure 4. *SMN1/2* Amplification-based Nanopore Assay Classifies Copy Number, Silent Carrier, and Disease Modifier Status with 92-100% Accuracy. We utilized the prototype assay for *SMN1/2* copy number (CN) predictions across **A)** 90 unique cell-lines in 154 samples (142 passed QC) and **B)** 232 unique whole blood samples (227 passed QC). Overall, copy number predictions resulted in 342/369 (93%) *SMN1* and 344/369 (93%) *SMN2* accuracy compared with orthogonal methods (data not shown). **C)** A heuristic analysis detected *SMN1* silent carrier variants (SC1=c.*3+80T>G, SC2=c.*211_*212del) with 99% accuracy and a positive disease modifier (DM=c.859G>C) with 100% accuracy.

SMN1/2 CN analysis can also be performed using a series of shorter amplicons rather than two long amplicons as shown here. To compare performance, see poster P477.

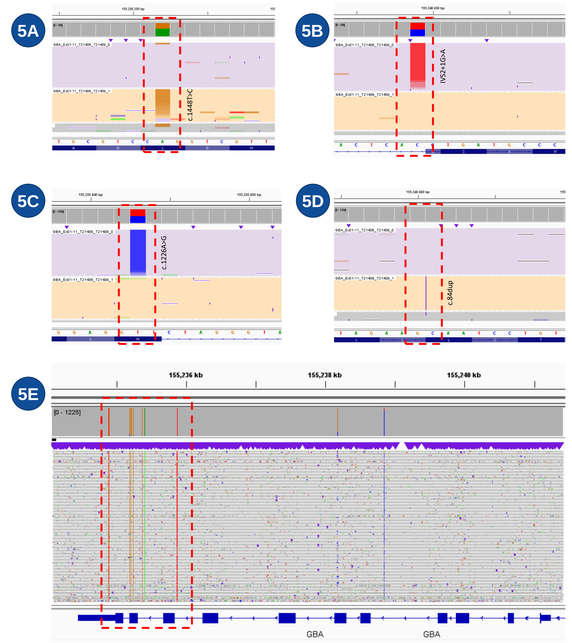

Figure 5. Sequence Data Reveals Diverse Pathogenic Variants Across the *GBA* gene. Clinically affected cell-line NA20270 is a compound heterozygote and contains two SNVs within the *GBA* gene; **A)** one allele has a T>C transition at nucleotide 1448 in exon 10 (L444P, c.1448T>C), and **B)** the other allele has a splice site mutation in intron 2 (IVS2+1G>A). **C)** Cell-line ND14143 is heterozygous for missense variant N370S, c.1226A>G. Cell-line NA00852 is a compound heterozygous sample with **D)** an insertion of a G at c.84 (c.84dup), and N370S (c.1226A>G) (data not shown). **E)** A homozygous *GBA*-*GBAP1* fusion was identified in cell-line NA20273. The variants in the red box are concordant with *GBAP1* paralogue specific variants. All samples were verified with orthogonal methods (data not shown).

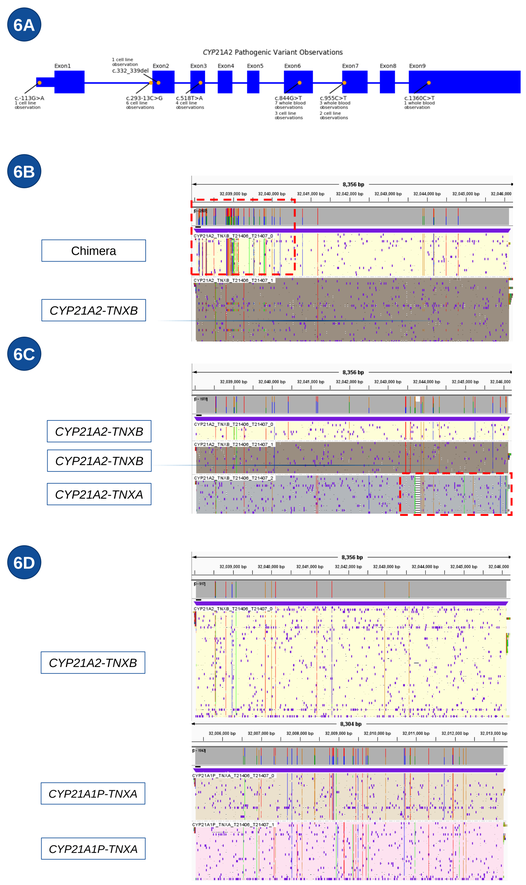

Figure 6. Accurate Resolution of Copy Number and Pseudogene Fusions in the *CYP21A2* Gene Cluster Utilizing Sequencing Deconvolution. **A)** 28 pathogenic variants were identified across cell-line and whole blood samples and mapped to their location in *CYP21A2*. **B)** A sample with a *CYP21A2* and *CYP21A1P* fusion, denoted by a 30 kb deletion. The variants in red box align with *CYP21A1P* PSVs. **C)** A whole blood sample with a duplication of a *CYP21A2*, associated with *TNXA* instead of *TNXB*. The assay identified a 121 bp deletion (red box) between *CYP21A2* and *TNXB* as well as *TNXA* PSVs to differentiate allele. **D)** A presumed normal whole blood sample identified as a carrier, with a single copy of *CYP21A2* and two copies of *CYP21A1P*. All samples were verified with orthogonal methods (data not shown).

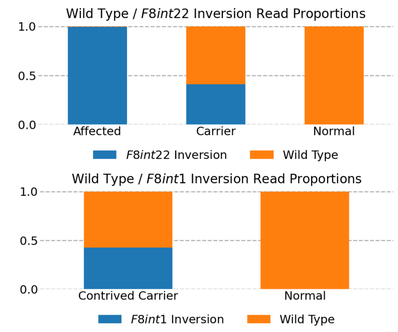

Figure 7. Prototype Assay Amplifies and Identifies >100 kb *F8* Inversions. *F8* inversions were successfully amplified and aligned across 3 samples (1 affected intron 22 inversion, 1 intron 22 carrier, and 1 contrived intron 1 carrier). The results show the proportion (y-axis) of *F8* reads that aligned to either the reference genome (Wild Type) or an *F8* inversion. Inversion reads were not identified in the wild type sample. Both intron 1 and intron 22 inversions were verified by orthogonal methods (data not shown).

Table 1. Identification of 27 Carriers in a Presumed Normal Cohort (n=232) of Whole Blood Samples using the Prototype Assay. Two donor samples (bolded text) were identified as carriers for both *SMN1* (including *SMN1* SC) and *CYP21A2*. No Hemophilia A (*F8*) carriers were identified in line with expected carrier rates. All *CYP21A2* and *GBA* variants are pathogenic or likely pathogenic. All variants were confirmed by orthogonal methods; QC = QC failure.

Sample ID	*SMN1* CN	*SMN2* CN	*SMN2* DM	*SMN1* SC1	*SMN1* SC2	*CYP21A2*	*GBA*	*F8*	Identified Carrier
SID395	**1**	**3**	**(-)**	**(-)**	**(-)**	**Q318X (c.955C>T)**	**WT**	**WT**	**SMN1, CYP21A2**
SID405	2	2	(-)	(-)	(-)	30 Kb Deletion	WT	WT	CYP21A2
SID412	2	1	(-)	(-)	(-)	V281L (c.844G>T)	WT	WT	CYP21A2
SID433	2	2	(-)	(-)	(-)	WT	R496H (c.1604G>A)	WT	GBA
SID434	2	2	(-)	(-)	(-)	WT	N370S (c.1226A>G)	WT	GBA
SID439	2	2	(-)	(-)	(-)	V281L (c.844G>T)	WT	WT	CYP21A2
SID449	2	2	(-)	(-)	(-)	Q318X (c.955C>T)	WT	WT	CYP21A2
SID478	QC	QC	QC	QC	QC	V281L (c.844G>T)	WT	WT	CYP21A2
SID514	3	1	(-)	(-)	(-)	WT	N370S (c.1226A>G)	WT	GBA
SID539	2	1	(-)	(-)	(-)	Q318X (c.955C>T)	WT	WT	CYP21A2
SID554	QC	QC	QC	QC	QC	Q318X (c.955C>T)	WT	WT	CYP21A2
SID563	3	1	(-)	(-)	(-)	V281L (c.844G>T)	WT	WT	CYP21A2
SID574	2	1	(-)	(-)	(-)	P453S (c.1360C>T)	WT	WT	CYP21A2
SID606	2	1	(-)	(-)	(-)	V281L (c.844G>T)	WT	WT	CYP21A2
SID617	2	1	(-)	(-)	(-)	V281L (c.844G>T)	WT	WT	CYP21A2
SID546	1	2	(-)	(-)	(-)	WT	WT	WT	SMN1
SID461	1	1	(-)	(-)	(-)	WT	WT	WT	SMN1
SID589	2	2	(-)	Positive	Positive	WT	WT	WT	SMN1 SC
SID578	2	2	(-)	Positive	Positive	WT	WT	WT	SMN1 SC
SID510	3	1	(-)	Positive	Positive	WT	WT	WT	SMN1 SC
SID633	2	1	Positive	(-)	(-)	WT	WT	WT	SMN1 DM
SID564	4	0	(-)	Positive	Positive	WT	WT	WT	SMN1 SC
SID530	2	1	(-)	Positive	Positive	WT	WT	WT	SMN1 SC
SID420	3	1	(-)	Positive	Positive	WT	WT	WT	SMN1 SC
SID414	**3**	**1**	**(-)**	**Positive**	**Positive**	**V281L (c.844G>T)**	**WT**	**WT**	**SMN1 SC, CYP21A2**
SID409	3	1	(-)	Positive	Positive	WT	WT	WT	SMN1 SC
SID441	2	2	Positive	(-)	(-)	WT	WT	WT	SMN1 DM

Conclusions

- **The prototype AmplideX PCR-based nanopore assay accurately resolves multiple challenging variants in *F8*, *GBA*, *CYP21A2*, *TNXB* and *SMN1*, genes that are critical for assessing risk in population carrier screening but are difficult to resolve due to highly homologous pseudogenes.**
- **The single-tube streamlined workflow allows multiplexing of at least 95 samples in a sequencing run, replacing four independent assays that target fewer variants and provide less information than comprehensive sequencing.**
- **Detection of two potential dual carriers (*CYP21A2* and *SMN1*) highlights the importance of a unified carrier screening approach.**
- **Combining long amplicons with sequence deconvolution allows variant phasing (as shown in *SMN1*) and identification of complex fusions (as shown in *CYP21A2* and *GBA*), resulting in >90% detection of known variants in cell-line samples.**

REFERENCES

1. Lincoln SE, et al. Genet Med 2021;23:1673-1680
2. Zheng, Z., Li, S., Su, J. et al. Nat Comput Sci 2, 797–803 (2022).

Asuragen a bio-techne brand