A Combination of Long-range PCR and Long-read Sequencing Resolves Common, High-risk Dark and Camouflaged Carrier Screening Genes

Sarah Statt[1], **Julie R Thibert**[1], **Jon Kemppainen**[1], **Cody Edwards**[1], **Liangjing Chen**[1], **Theodore Markulin**[1], **Jacob Wisotsky**[1], **Brian C Haynes**[1], **Cheryl Heiner**[2], **John Harting**[2] and **Gary J Latham**[1]

[1]Asuragen, Inc., Austin, TX; [2]Pacific Biosciences of California, Inc., Menlo Park, CA

Poster Number: eP371

Summary

- High-prevalence carrier genes, associated with disorders such as Fragile X Syndrome and Hemophilia A, include GC-rich repeats, complex structural variants, and/or pseudogenes that derail conventional sequencing methods.
- We established a prototype streamlined, multi-gene, sample-multiplexed AmplideX® PCR/SMRT sequencing assay* that presents a solution for carrier screening.
- We accurately detected different classes of pathogenic variation across 24 samples, including large structural variants (such as inversions), Single Nucleotide Variants (SNVs), Short Tandem Repeats (STRs), Insertions/Deletions (INDELs) and Copy Number Variants (CNVs).

Introduction

The human genome contains “dark” and “camouflaged” gene regions that cannot be adequately assessed using standard PCR and short-read sequencing technologies, ultimately preventing researchers from identifying mutations within these gene regions that are relevant to human disease. There are two primary causes for the obfuscation of these regions: 1) low complexity DNA that is hard-to-amplify by traditional PCR; and 2) duplicated loci that cannot be confidently aligned to a unique location using short sequencing reads. Such “dark” and “camouflaged” gene regions are especially problematic in Expanded Carrier Screening (ECS), where six of the top ten carrier screening disorders by reproductive risk to couples are difficult or intractable using conventional short-read sequencing methods (Figure 1). These six disorders (seven genes) are: Fragile X Syndrome (*FMR1*), Alpha Thalassemia (*HBA1/HBA2*), Duchenne and Becker Muscular Dystrophy (*DMD*), Hemophilia A (*F8*), Congenital Adrenal Hyperplasia (*CYP21A2*), and Spinal Muscular Atrophy (*SMN1*). The analysis of these high-prevalence carrier screening genes requires several different, specialized non-NGS methods that burden laboratories with inefficient workflows, disjointed results, and subpar detection rates.

Materials and Methods

We investigated combining novel, long-range AmplideX® PCR chemistry and long-read Single Molecule Real-Time (SMRT®) Sequencing to consolidate non-NGS workflows and provide more comprehensive analysis of ECS targets. Specifically, we sought to create a unified workflow to improve clinical sensitivity by resolving ≥95% of pathogenic variants in the following “hard-to-do” genes: *FMR1*, *HBA1/2*†, *DMD*, *F8*, *GBA*, *CYP21A2*, and *SMN1*. Samples tested were amplified with AmplideX PCR, barcoded, pooled, prepped with SMRTbell® Express TPK 2.0 (PacBio®) and ran using a 10hr movie time on a Sequel® SMRT® Cell 1M v3 (PacBio). Pre-analysis was performed on SMRT® Link v9.0 to achieve multiplexed CCS HiFi reads for each sample, followed by alignment to Hg38 via minimap2[1]. *FMR1* associated reads were processed through an in-house pipeline to resolve CGG genotypes and AGG status. Other gene targets underwent a bespoke sequence deconvolution method to resolve paralogous sequences, where applicable, and further analyzed for pathogenic variants.

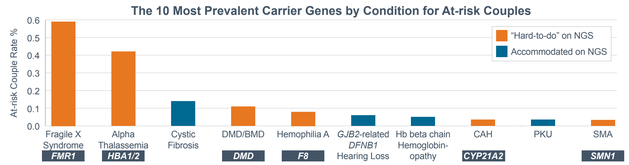

Figure 1. The 10 Most Prevalent Carrier Genes by Condition for At-risk Couples. Dark and camouflaged genes are especially problematic in Expanded Carrier Screening (ECS), where six (shown in orange) of the top ten carrier screening disorders by reproductive risk to couples[2-4] are difficult or intractable using conventional short-read sequencing methods.

**This product is under development. Future availability and performance to be determined.*
†Data for updated HBA1/2 designs were pending at the time of the poster submission deadline.
#Research Use Only. Not for use in diagnostic procedures.
Presented at ACMG 2021, April 13-16, virtual.

Results

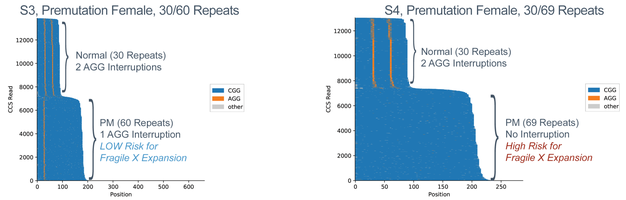

Figure 2. AmplideX PCR/SMRT Prototype Assay Resolves *FMR1* CGG Repeats & AGG Interruptions that Inform the Risk of Expansion for Fragile X. The number of AGG interruptions modify expansion risk for premutation (55-200 CGG) samples. As shown in the above waterfall plots, one sample (S3) had one AGG in the premutation allele (low risk having a child with a further expansion) while the other (S4) had no AGG interruptions in the 69-CGG allele (high risk for further expansion).

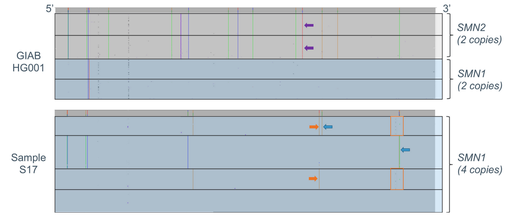

Figure 3. Sequence Deconvolution by Paralog-specific Variants (PSVs) Allows Copy Number and INDEL/SNV Identification Within *SMN1/SMN2* Alleles. As seen in Genome in a Bottle (GIAB) HG001, the colored vertical lines in *SMN2* showcase differences between *SMN1* and *SMN2*, with the purple arrows identifying the c.840C>T variant that affects splicing of exon 7. Sample S17 shows a complex example with 4 copies of *SMN1, SMN2* gene conversions (blue arrows) and silent carrier gene duplication variants c.*3+80T>G (orange arrows) and c.*211_*212del (orange boxes).

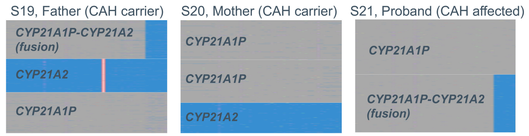

Figure 4. Accurate Resolution of Copy Number and Pseudogene fFusions are Identified in *CYP21A2* Gene Cluster Alignments. Two CAH carriers are shown above: A mother with a whole gene deletion and a father with a 30kb deletion that results in a fusion of pseudogene and gene. Analysis of the proband correctly identifies both pathogenic alleles from the parents. Pink line in the father represents a non-pathogenic SNP.

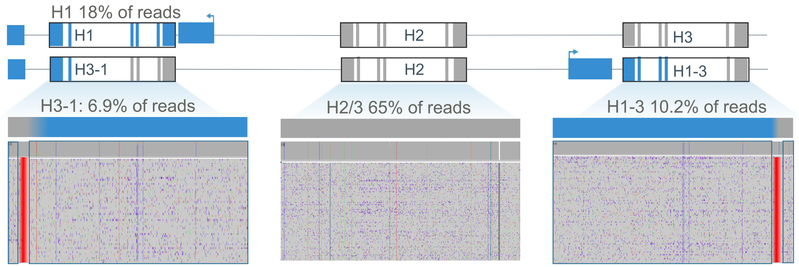

Figure 5. AmplideX PCR/SMRT Prototype Assay Detects Complex Inversions in a Hemophilia A Carrier. Utilizing sequence deconvolution by PSVs, we were able to identify a 0.6-Mb inversion in *F8* resulting in a fusion of Intron 22 H1 and H3 in sample S9. Red line identifies breakpoint between H1 and H3. Blue lines represent H1 specific PSVs while grey lines represent H2/H3 specific PSVs. Percentages showcase fraction of total F8 reads that aligned to region.

Table 1. AmplideX PCR/SMRT Prototype Assay Results for 24 Samples as Compared to Reference Truth. As seen in **A)**, *FMR1* CGG repeats across all disease categories (intermediate/INT, premutation/PM and full mutation) were called within previously established AmplideX PCR/CE *FMR1* Kit# precision, and AGG interruptions were identified in all samples in agreement with the reference Xpansion Interpreter® assay, noting * indicates a sample outside of the PCR/CE assay resolution. As seen in **A)** and **B)**, INDEL and/or SNV carriers and/or affected individuals of *GBA*, *DMD*, and *CYP21A2* variants were detected consistent with reference calls. Complex *CYP21A2* gene conversions, pseudogene fusions, and *CYP21A2/CYP21A1P1* gene copy number were called using as few as 50 reads per allele. *SMN1* analysis revealed copy numbers ranging from 0 to ≥4 along with gene conversions, and silent carrier-associated variants that were concordant with orthogonal AmplideX PCR/CE *SMN1/2* Plus Kit# calls, noting * indicates a sample outside of the PCR/CE assay resolution. Lastly, the method identified 0.6-Mb inversions in *F8* despite the challenges in resolving the highly homologous and GC-rich structural perturbation.

A

Sample #	Sample ID	Sex	Allele 1	Allele 2	Category	Match Reference?
S1	NA20236	F	*FMR1* CGG: 31 *FMR1* AGG: 10+9+10	*FMR1* CGG: 53 *FMR1* AGG: NO	Fragile X Carrier, INT	YES
S2	WB #1	F	*FMR1* CGG: 32 *FMR1* AGG: 9+12+9	*FMR1* CGG: 54 *FMR1* AGG: 9+44	Fragile X Carrier, INT/PM Border	YES
S3	WB #2	F	*FMR1* CGG: 30 *FMR1* AGG: 10+9+9	*FMR1* CGG: 60 *FMR1* AGG: 9+50	Fragile X Carrier, PM	YES
S4	WB #3	F	*FMR1* CGG: 30 *FMR1* AGG: 10+9+9	*FMR1* CGG: 69 *FMR1* AGG: NO	Fragile X Carrier, PM High Risk	YES
S5	NA20240	F	*FMR1* CGG: 30 *FMR1* AGG: 10+9+9	*FMR1* CGG: 80 *FMR1* AGG: NO	Fragile X Carrier, PM High Risk	YES
S6	NA06903	F	*FMR1* CGG: 24 *FMR1* AGG: 14+9	*FMR1* CGG: 92 *FMR1* AGG: 10+81	Fragile X Carrier, PM	YES
S7	NA20239	F	*FMR1* CGG: 20 *FMR1* AGG: 10+9	*FMR1* CGG: 202 *FMR1* AGG: 9+9+182	Fragile X, Full Mutation	YES
S8	NA07537	F	*FMR1* CGG: 29 *FMR1* AGG: 9+9+9	*FMR1* CGG: 321 *FMR1* AGG: 10+310	Fragile X, Full Mutation	YES*
S9	NIBSC 06/204	F	*F8*: normal	*F8*: Intron22 inversion	Hemophilia A Carrier	YES
S10	NIBSC 07/116	M	*F8*: Intron22 inversion	N/A	Hemophilia A Affected	YES
S11	ND14143	F	*GBA*: Normal	*GBA*: c.1226A>G	*GBA* Carrier	YES
S12	NA20270	F	*GBA*: c.115+G>A	*GBA*: c.1448T>C	*GBA* Affected	YES
S13	NA05263	M	*DMD* c.7893delC (p.Q2632Sfs6)	N/A	*DMD* Affected	YES
S14	NA04364	M	*DMD* Exon 51-55 DEL	N/A	*DMD* Affected	YES

B

Sample #	Sample ID	Sex	Observed SNV(s)/INDEL(s)	Copy Number	Category	Match Reference?
S15	NA09677	M	No Mutations	*SMN1*: 0 copy *SMN2*: 3 copies	SMA Affected	YES
S16	NA19984	M	c.*3+80T>G, c.*211_*212del	*SMN1*: 2 copies *SMN2*: 1 copy	SMA Silent Carrier	YES
S17	NA19360	M	c.*3+80T>G, c.*211_*212del, *SMN2* gene conversion	*SMN1*: 4 copies *SMN2*: 0 copies	SMA Silent Carrier	YES
S18	NA03814	F	No Mutations	*SMN1*: 1 copy *SMN2*: 5 copies	SMA Carrier	YES*
S19	NA14733	M	No Mutations	*CYP21A2*: 1 copy *CYP21A1P*: 1 copy *CYP21A2-CYP21A1P*: 1 copy	CAH Carrier	YES
S20	NA14732	F	No Mutations	*CYP21A2*: 1 copy *CYP21A1P*: 2 copies	CAH Carrier	YES
S21	NA14734	M	No Mutations	*CYP21A2*: 0 copies *CYP21A1P*: 1 copy *CYP21A2-CYP21A1P*: 1 copy	CAH Affected	YES
S22	NA11781	F	*CYP21A2* Homozygous c.293-13C>G	*CYP21A2*: 2 copies *CYP21A1P*: 2 copies	CAH Affected	YES
S23	NA12217	M	*CYP21A2* c.515T>A (p.Ile172Asn)	*CYP21A2*: 1 copy *CYP21A1P*: 2 copies *CYP21A2-CYP21A1P*: 1 copy	CAH Affected	YES
S24	NA20530	F	*CYP21A2* gene conversion (*CYP21A2* + *TNXA*)	*CYP21A2*: 3 copies *CYP21A1P*: 2 copies	Normal	YES

Conclusions

- We demonstrated that a workflow combining unbiased, long-read PCR enrichment with high-fidelity, long-read SMRT sequencing can accurately resolve pathogenic variants in “hard-to-do” genes.
- Pathogenic variants that were detected included >300 CGG repeats and AGG interruptions in *FMR1*, large 0.6Mb intron 22 inversions in *F8*, small and large deletions in *DMD* and *CYP21A2*, and copy numbers for *SMN1* and *CYP21A2* contrasted with their pseudogenes.
- This method has the potential to address real-world gaps in carrier screening by streamlining the Expanded Carrier Screening workflow and covering the full spectrum of clinically-relevant dark and camouflaged gene regions that contribute the largest reproductive risk to couples.

References

1. Li, H., Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 2018. 34(18): p. 3094-3100.
2. Balzotti, M., et al., Clinical validity of expanded carrier screening: Evaluating the gene-disease relationship in more than 200 conditions. Hum Mutat, 2020. 41(8): p. 1365-1371.
3. Ben-Shachar, R., et al., A data-driven evaluation of the size and content of expanded carrier screening panels. Genet Med, 2019. 21(9): p. 1931-1939.
4. Capalbo, A., et al., Optimizing clinical exome design and parallel gene-testing for recessive genetic conditions in preconception carrier screening: Translational research genomic data from 14,125 exomes. PLoS Genet, 2019. 15(10): p. e1008409..

Asuragen®

PACBIO®