A Method for the Identification of Variants in Alzheimer's Disease Candidate Genes and Transcripts Using Hybridization Capture Combined with Long-Read Sequencing

Steve Kujawa[1], Jenny Ekholm[1], Kevin Eng[1], Ting Hon[1], Elizabeth Tseng[1], Aaron Wenger[1], Kristina Giorda[2], Jiashi Wang[2] & Mirna Jarosz[2]
[1]PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025
[2]Integrated DNA Technologies, 1710 Commercial Park, Coralville, IA, 52241

Introduction

Alzheimer's disease (AD) is a devastating neurodegenerative disease that is genetically complex. Although great progress has been made in identifying fully penetrant mutations in genes that cause early-onset AD, these still represent a very small percentage of AD cases. Large-scale, genome-wide association studies (GWAS) have identified at least 20 additional genetic risk loci for the more common form of late-onset AD. However, the identified SNPs are typically not the actual risk variants, but are in linkage disequilibrium with the presumed causative variants[1].

Long-read sequencing together with hybrid-capture targeting technologies provides a powerful combination to target candidate genes/transcripts of interest. Here we present a method for capturing genomic DNA (gDNA) and cDNA from two AD subjects using a panel of probes targeting 35 AD candidate genes. By combining xGen® Lockdown® probes with SMRT Sequencing, we provide completely sequenced candidate genes as well as their corresponding full-length transcripts. Furthermore, we are able to take advantage of heterozygous variants to phase the genes and their corresponding transcript isoforms into their respective haplotypes.

Materials and Methods

A custom panel of 35 AD genes (Table 1) was designed using IDT xGen Lockdown probes. Probes were placed approximately every 1 kb (Figure 1) and designed to cover the entire gene (exons, introns and regulatory regions).

Genes Included in the Panel				
ABCA7	*APH1*	*APOE*	*APP*	*BACE1*
BIN1	*BSG*	*CASS4*	*CD2AP*	*CD33*
CELF1	*CLU*	*CR1*	*EPHA1*	*FERMT2*
GRN	*HLA-DRB1*	*HLA-DRB5*	*INPP5D*	*MAPT*
MEF2C-AS1	*MS4A6A*	*NCSTN*	*NME8*	*PICALM*
PSEN1	*PSEN2*	*PTK2B*	*RIN3*	*SLC24A4*
SNCA	*SORL1*	*TOMM40*	*TREM2*	*ZCWPW1*

Table 1. The custom AD panel includes 35 genes.

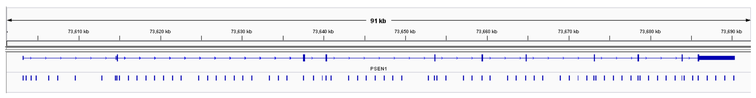

Figure 1. Probe design for *PSEN1*. 77 probes were evenly spaced across the ~90 kb gene.

Two subjects were sequenced during this experiment (Table 2). For each subject, gDNA was captured with the custom AD panel according to the published protocol[2] and sequenced on eight PacBio RS II SMRT Cells. Separately, for each subject, RNA was converted to cDNA, captured with the custom AD panel according to the published protocol[3] and sequenced on four PacBio RS II SMRT Cells.

	Subject	Source of Genomic DNA	Source of Total RNA
#1	87 year-old male	Brain, Frontal Lobe	Brain, Temporal Lobe
#2	93 year-old female	Skeletal Muscle	Brain, Temporal Lobe

Table 2. gDNA and total RNA from two AD subjects were purchased from BioChain Institute, Inc.

Results - Genes

Reads from the gDNA from Subjects 1 and 2 were mapped to the hg38 reference genome using NGM-LR. Structural variants >50 bp were called using PBHoney Spots (Table 3).

	# Events	# Unique Genes
Deletions >50 bp	15	10
Insertions >50 bp	16	8

Table 3. SVs >50 bp Observed in the 35 AD Genes from Subjects 1 & 2. 31 unique SVs were observed, ranging in size from 65 bp to multiple kilobases.

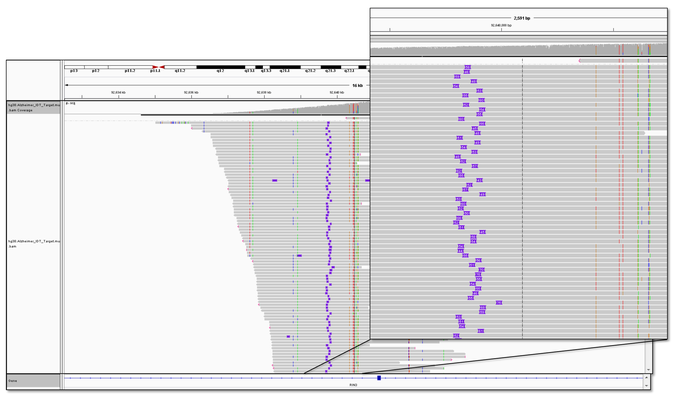

Figure 2. gDNA of *RIN3* gene from Subject 2. Approximately 50 bp insertion (purple bars) found in intron 4 of the *RIN3* gene.

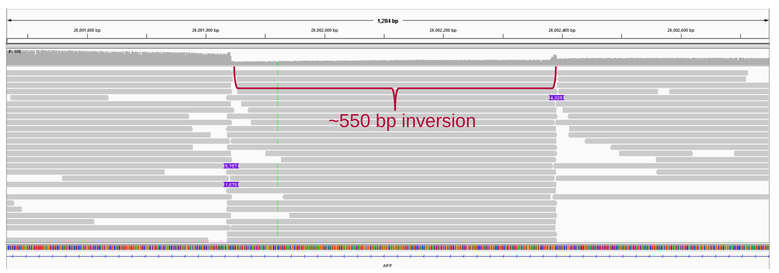

Figure 3. gDNA of *APP* gene from Subject 1. Approximately 550 bp inversion in intron 6 of the *APP* gene.

Results - Transcripts

The captured cDNA from Subjects 1 and 2 were run through the Iso-Seq (ToFU) bioinformatics pipeline to obtain Quiver-polished, full-length, high-quality transcript sequences. Sequences were then mapped to the hg38 genome and filtered with criteria: (1) alignment coverage ≥99%; (2) alignment identity ≥95; (3) at least 5 FL read support; (4) is not a 5' degraded product; and (5) overlaps the probe target region. This resulted in a total of 515 isoforms from Subject 1 and 507 isoforms from Subject 2. To compare with existing annotation, we selected all Gencode v25 transcripts from the target genes with an annotated transcript support level of 1 (most reliable annotation, all junctions supported by at least one mRNA evidence), resulting in 111 isoforms.

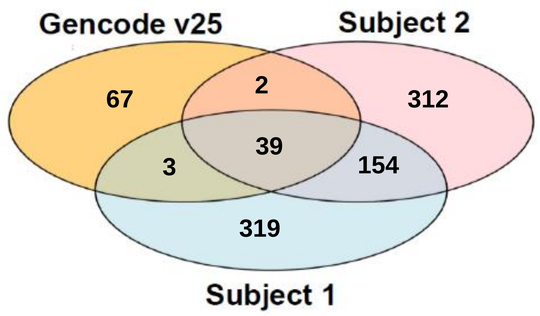

Figure 4. Comparison of isoforms observed in Subjects 1 & 2 with Level 1 isoforms in Gencode v25.

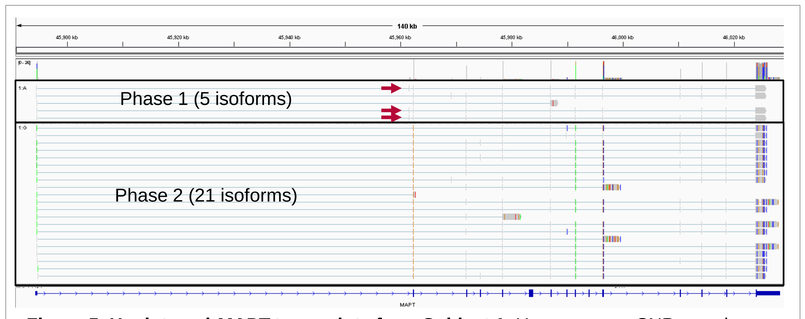

Figure 5. Haplotyped *MAPT* transcripts from Subject 1. Heterozygous SNPs can be used to haplotype the transcripts. A novel exon (red arrows) was observed in three of the five isoforms in Phase 1 and not observed in any of the 21 isoforms in Phase 2.

Results – Haplotyped Variants

After alignment to the hg38 genome, heterozygous variants can be used to further assign the gDNA and transcripts to their appropriate haplotype. As the average fragment size of the captured gDNA is ~6 kb, it is possible to phase regions that are multiple, tens of kilobases in length. Full-length transcripts are easily phased if a heterozygous SNP is captured in an exon or retained intron.

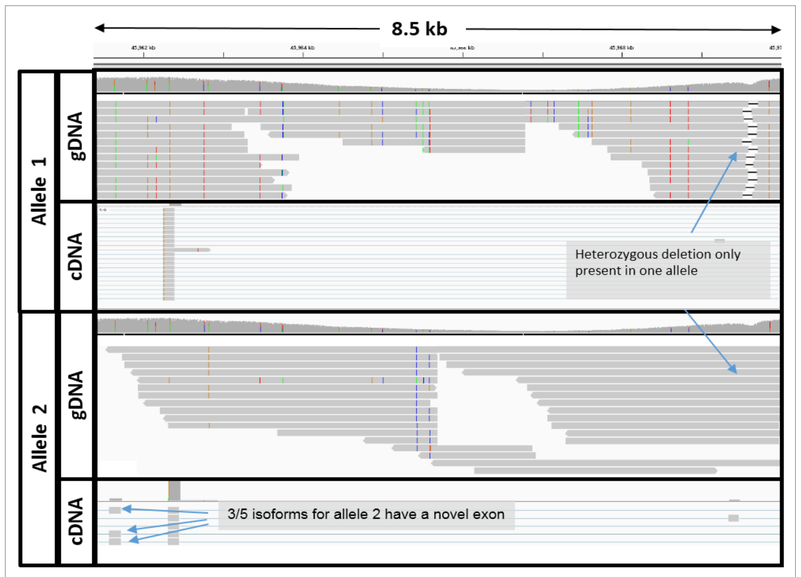

Figure 6. Phased Genes & Transcripts of *MAPT* from Subject 1. Heterozygous SNPs can be used to phase the genomic DNA and transcripts to their appropriate haplotype. Once phased, variants such as this 100 bp heterozygous deletion (blue arrows upper right) can be studied to better understand their potential impact on transcript isoform production. Five unique isoforms were observed from allele 2. Three of these isoforms contained a novel exon (blue arrows lower left) that was only present in allele 2. These exons were flanked by the canonical "AG" and "GT" splice sites in the gDNA.

Conclusion

Combining xGen Lockdown probes with SMRT Sequencing provides a method for completely sequenced candidate genes and their corresponding full-length transcripts.

This method enables:

- Detection of a broad range of genomic variants, from SNPs to multi-kilobase insertions and deletions
- Detection of novel transcript isoforms, including novel exons
- Assignment of variants and transcripts isoforms to their specific alleles

References

1. Van Cauwenberghe C, et al. (2015). The genetic landscape of Alzheimer disease: clinical implications and perspectives. *Genet Med*, 18(5), 421-430.
2. Target Sequence Capture Using IDT Library with PacBio® Barcoded Adapters
3. Full-length cDNA Target Sequence Capture Using IDT xGen® Lockdown® Probes

Acknowledgements
The authors would like to thank everyone who helped generate data for the poster.