Screening for causative structural variants in neurological disorders using long-read sequencing

J. Ekholm[1], S. Kujawa[1], Y. Tsai[1], D. Greenberg[1], T. Hon[1], K. Eng[1], A. Wenger[1], E. Tseng[1], J. Wang[2], M. Jarosz[2], K. Giorda[2], T. Clark[1]
[1]PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025, CA, United States, [2]Integrated DNA technologies, Coralville, IA, United States

Introduction

Over the past decades neurological disorders have been extensively studied using positional cloning and association studies, both producing a large number of candidate genomic regions and candidate genes. The SNPs identified in these studies rarely represent the true disease-related functional variants. However, more recently a shift in focus from SNPs to larger structural variants(SV) has yielded breakthroughs in our understanding of neurological and neuromuscular disorders.

In fact, due to their larger size, SV events in the human genome account for a greater number of variable bases than SNPs. However, they have not been extensively studied mainly due to the difficulty in accessing large and complex genomic regions using genotyping or short-read sequencing approaches.

Here we discuss a candidate gene screening method that is optimized for SV discovery using enrichment of long DNA fragments and PacBio long-read sequencing. Single Molecule, Real-Time (SMRT) Sequencing combines single-molecule observation, long reads, and low degree of bias to fully characterize genetic complexity of disease associated structural variants in neurological disorders. We will also discuss our development of a novel, amplification-free enrichment technique using the CRISPR/Cas9 system and demonstrate the targeting of large repeat expansions in individuals with neuromuscular disorders.

Screening of 35 Alzheimer's disease candidate genes

A custom panel of 35 Alzheimer's disease (AD) GWAS candidate genes[1] (Table 1) was designed using IDT xGen Lockdown probes[2]. Probes were spaced 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 GWAS candidate genes[1].

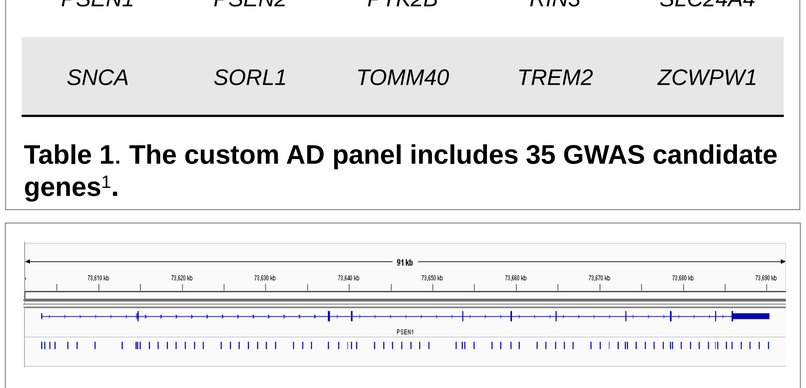


Figure 1. Probe design for *PSEN1*. 77 probes were evenly spaced across the ~90 kb gene.

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 8 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 4 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

Reads from the gDNA from Subjects 1 and 2 were mapped to the hg38 reference genome using NGM-LR[4]. SV >50 bp were called using PBHoney Spots[5] (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.

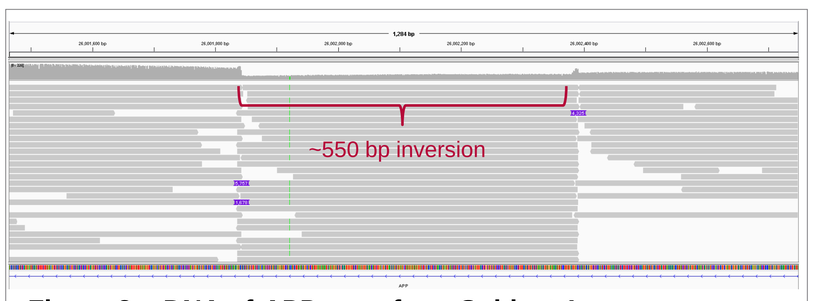


Figure 2. gDNA of *APP* gene from Subject 1. Approximately 550 bp inversion in intron 6 of the *APP* gene.

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:

- Alignment coverage ≥99%
- Alignment identity ≥95%
- At least 5 full-length reads to support
- Not a 5 prime degraded product
- Overlaps the probe target region

This resulted in a total of 515 isoforms from Subject 1 and 507 isoforms from Subject 2. When comparing with all existing annotation from Gencode v25 transcripts from the target genes with an annotated transcript support level of 1 very few overlapped.

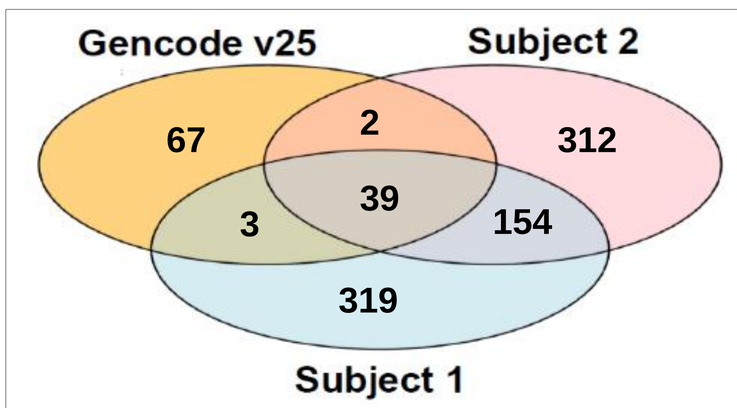


Figure 3. Comparison of isoforms observed in Subjects 1 & 2 with Level 1 isoforms in Gencode v25.

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.

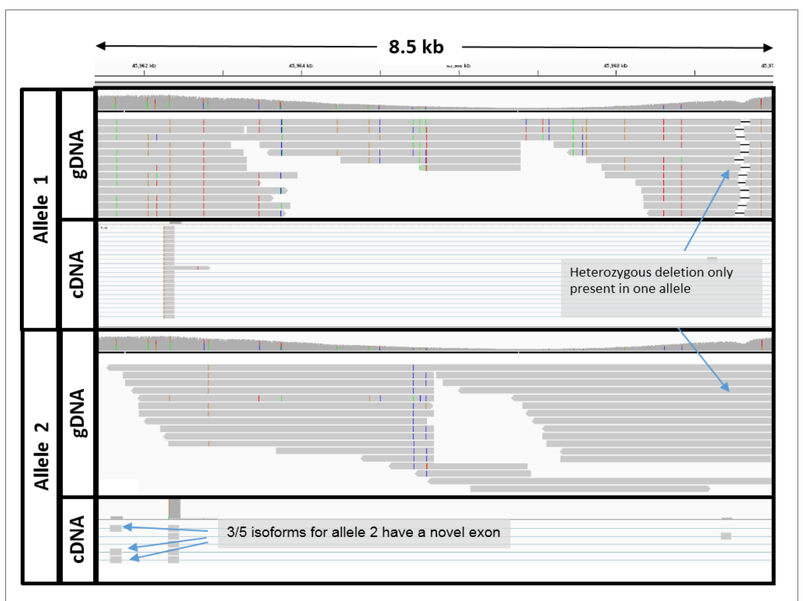
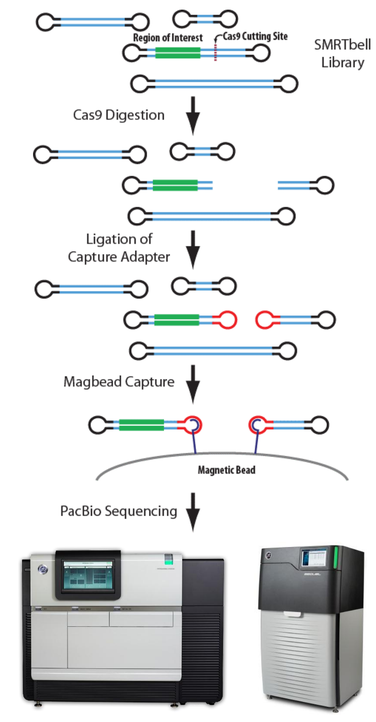


Figure 4. Phased Genes & Transcripts of *MAPT* from Subject 1. Heterozygous SNPs can be used to phase the genomic DNA and transcripts to their appropriate haplotype. Once phased, variants such as this 100 bp heterozygous deletion (blue arrows upper right) can be studied to better understand their potential 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.

Amplification-free enrichment using CRISPR/Cas9

Figure 5. Overview of the CRISPR/Cas9 method.

- Prepare standard SMRTbell library
- Design guide RNA (crRNA) to target the adjacent to the region of interest
- Digest SMRTbell library using Cas9
- Ligate capture adapter to a SMRTbell template to enrich for templates containing the targeted region of interest
- Sequence the SMRTbell templates on a PacBio RS II or Sequel System



Capturing and sequencing of ALS candidate gene *C9orf72*

Using the novel PCR-free CRISPR-Cas9 enrichment method, we screened several disease associated loci including the hexanucleotide repeat expansion C9orf72 that is associated with 40% of familiar ALS cases. Because there is no amplification, this method avoids bias or errors from PCR of a hard-to-amplify region. In addition, base modification is preserved in a single-molecule fashion allowing one to detect potential sample mosaicism[6].

Target Gene	Associated Disease(s)	Target Size	Repeat	Molecules on Target
HTT	Huntington's Disease	1125 bp	CAG	1197
C9orf72	**Familial Frontotemporal Dementia (FTD) and Amyotrophic Lateral Sclerosis (ALS)**	**1261 bp**	**CCCCGG**	**213**
SCA10	Spinocerebellar Ataxia Type 10	1019 bp	Variable ATTCT	2119
FMR1	Fragile X and Fragile X-associated Tremor/Ataxia Syndrome (FXTAS)	1013 bp	CGG	1105

Table 4. CRISRP/Cas9 targets. Guide RNAs designed to capture 4 repeat expansion loci were multiplexed together. While the number of molecules on-target varied between the different repeat expansions, the molecules on target for each loci were not affected by the multiplexing.

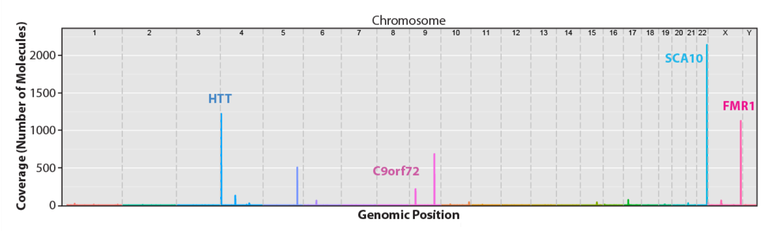


Figure 6. Coverage of molecules on target across the entire genome. Some off-target signals were seen but they all mapped to regions of high homology to the target of interest.

Conclusion

Combining xGen Lockdown probes with SMRT Sequencing provides a method to completely sequence candidate genes and their corresponding full-length transcripts allowing:

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

Amplification-free enrichment with CRISPR/Cas9 and SMRT Sequencing achieves the base-level resolution required to understand the underlying biology of repeat expansion disorders through:

- Ability to analyze genomic regions regardless of sequence content
- Repeat counting and identification of interruption sequences
- Exclusion of PCR bias and PCR errors
- Detection of mosaicism
- Simultaneous detection of epigenetics signals

References

1. Van Cauwenberghe C, et al. (2015). The genetic landscape of Alzheimer disease: clinical implications and perspectives. *Genetics in Medicine*, 18(5), 421-430.
2. http://www.pacb.com/wp-content/uploads/Unsupported-Protocol-Target-Sequence-Capture-Using-IDT-Library-PacBio-Barcoded-Adapters.pdf
3. http://www.pacb.com/wp-content/uploads/Unsupported-Protocol-Full-length-cDNA-Target-Sequence-Capture-IDT-xGen-Lockdown-Probes.pdf
4. Rescheneder P, et al. (2017). https://github.com/philres/ngmlr
5. English A. (2014) https://www.hgsc.bcm.edu/software/honey
6. Clark, T. et al. (February, 2017) Targeted SMRT Sequencing of difficult regions of the genome using a Cas9, non-amplification based method. *Poster presented at Advances in Genome Biology and Technology,* Florida, USA.

Authors
PacBio, 1305 O’Brien Drive, Menlo Park, CA 94025