Characterization of the Poly-T Variants in the *TOMM40* Gene Using PacBio Long Reads

Anand Sethuraman, Kevin Eng, Jenny Gu, Primo Baybayan, Steve Kujawa
PacBio, 1380 Willow Road, Menlo Park, CA 94025

Abstract

Genes associated with several neurological disorders have been shown to be highly polymorphic. Targeted sequencing of these genes using NGS technologies is a powerful way to increase the cost-effectiveness of variant discovery and detection. However, for a comprehensive view of these target genes, it is necessary to have complete and uniform coverage across regions of interest. Unfortunately, short-read sequencing technologies are not ideal for these types of studies as they are prone to mis-mapping and often fail to span repetitive regions. Targeted sequencing with PacBio long reads provides the unique advantage of single-molecule observations of complex genomic regions. PacBio long reads not only provide continuous sequence data though polymorphic or repetitive regions, but also have no GC bias.

Here we describe the characterization of the poly-T locus in *TOMM40*, a gene known to be associated with late onset Alzheimer's disease, using PacBio long reads. Probes were designed to capture the entire *TOMM40* gene, including the poly-T variant of interest. This region was captured from twelve different cell lines and sequencing libraries made using standard sample preparation methods. Here we present our results on the poly-T structural variant that was observed in *TOMM40* in these cell lines. We also present our results on probe design and barcoding strategies for a cost-effective solution.

Workflow

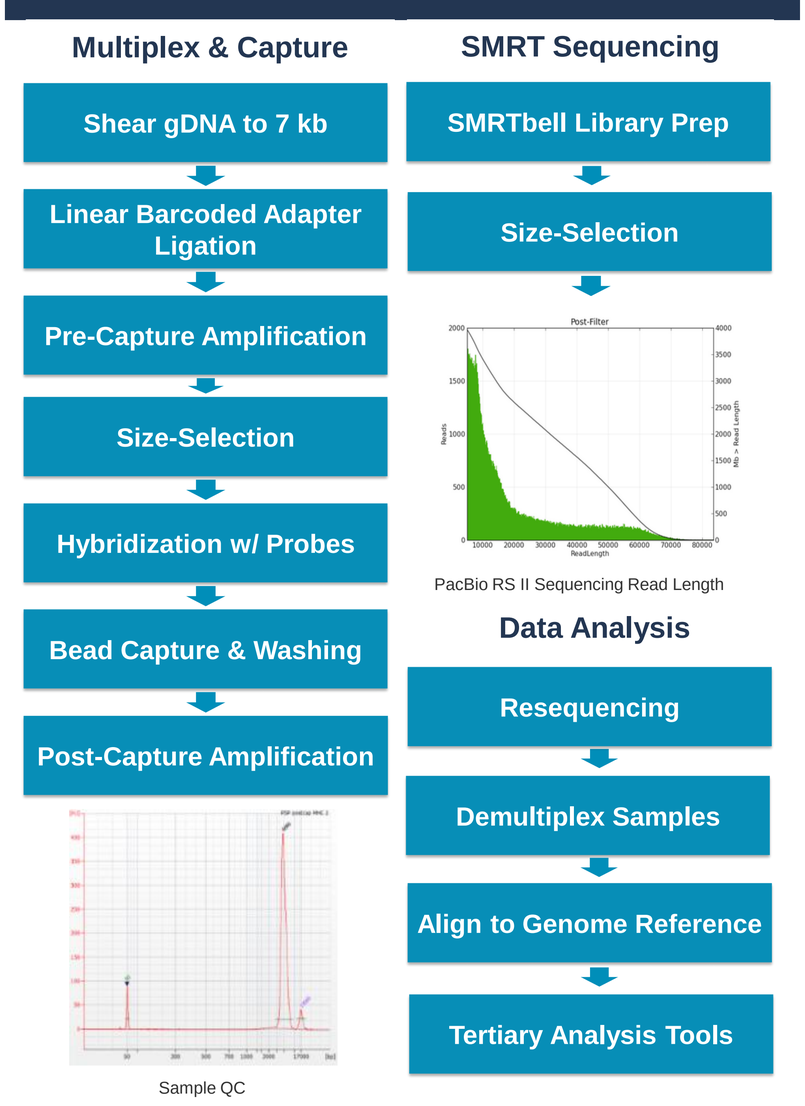

Figure 1. Multiplexed target-capture workflow incorporates the linear barcoded adapters immediately after gDNA shearing, allowing pooling of samples prior to probe hybridization and SMRTbell library preparation.

Genes and Samples

Target Genes
FLG
FLT3
HTT
FMR1
PIK3CA
PMS2
CD274(PDL1)
HLA-DRB1
C9orf72
BACE1
TOMM40
HLA-A
CYP2D6
APP
PSEN1
APOE

Table 1. IDT xGen lockdown probes were designed for 16 pharmaceutically-relevant genes.

Barcode	Cell Line
1	A431
9	HeLa
17	Jurkat
26	K562
38	MCF7
40	Raji
48	NA11922
52	NA18527
58	NA18942
59	NA18484
62	NA18526
70	NA12878

Table 2. 12 cell lines were barcoded and pooled before sequencing.

Barcoding

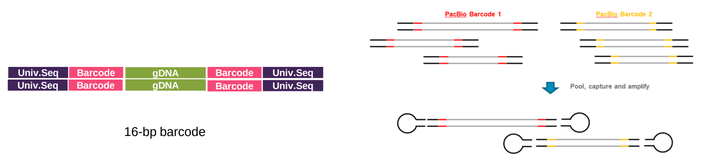

Figure 2. Barcoded linear adapters are blunt-end ligated to PCR amplicons. By adding the barcode to the standard SMRTbell adapter, researchers can use their existing primer designs to generate amplicons.

TOMM40

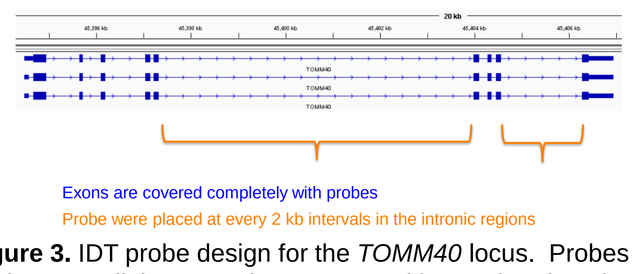

Figure 3. IDT probe design for the *TOMM40* locus. Probes were tiled across all the exons in *TOMM40* with a probe placed at every 2 kb intervals in the intronic regions.

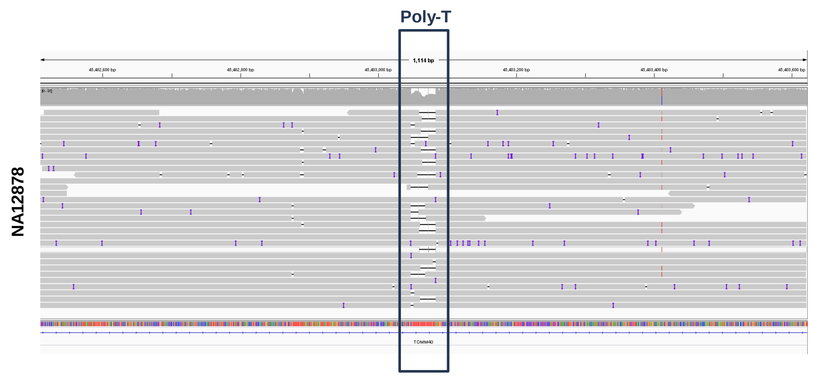

Figure 4. IGV screenshot showing the poly-T locus in *TOMM40* in NA12878

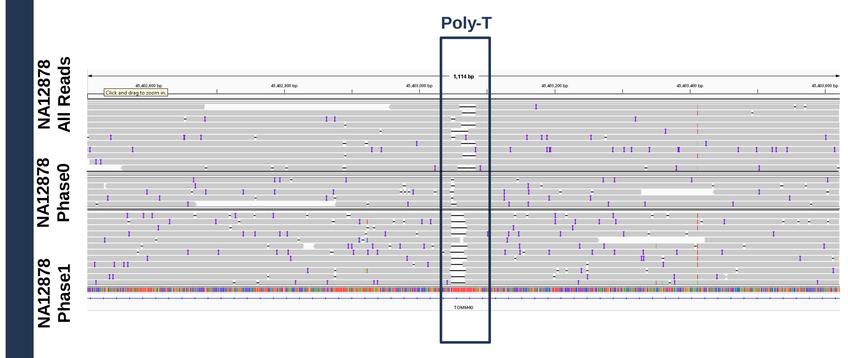

Figure 5. IGV screenshot showing the phased subreads for a 1kb region comprising the poly-T locus in *TOMM40* in NA12878

TOMM40 Poly-T Size Distribution

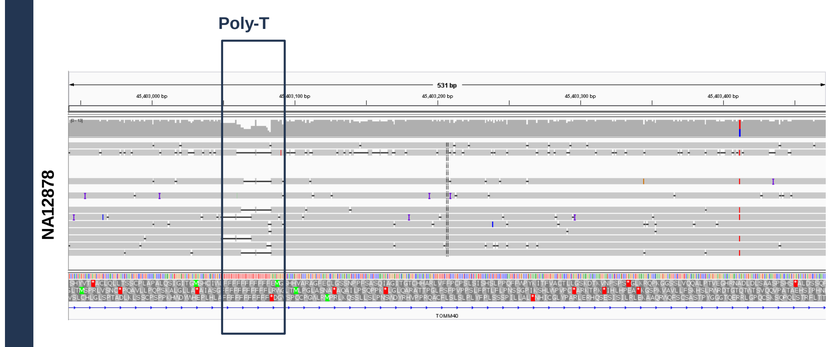

Figure 6. IGV screenshot showing Circular Consensus Sequencing (CCS) reads aligned to *TOMM40* poly-T region

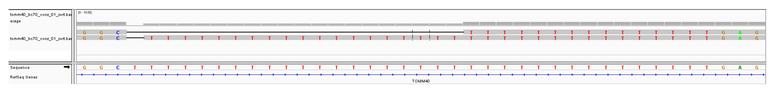

Figure 7. IGV screenshot showing alignment of Quiver consensus of the *TOMM40* gene poly-T haplotypes.

Sample	CCS	Quiver
NA12878	15 / 33	15 / 34

Table 3. Summary table of *TOMM40* poly-T homopolymer length as determined by either using the CCS reads or the Quiver consensus for the two haplotypes in NA12878.

Conclusions

- Using custom designed IDT xGen probes, we could successfully target and pull-down 16 pharmaceutically-relevant genes.
- By using PacBio barcodes, we could multiplex 12 cell lines for targeted sequencing experiments using the IDT capture probes.
- Using Targeted Phasing and Consensus Analysis pipeline, we could successfully haplotype and estimate high quality consensus for various genes.
- Finally, using highly quality Quiver consensus, we could accurately determine the precise length of *TOMM40* poly-T homopolymer length in the haplotypes from different cell lines.
- Multiplexing with PacBio barcodes for targeted sequencing experiments using capture probes saves time, money, and cost.

References

1. PacBio Barcodes Order Sheet:

Barcodes for multiplex targeted sequencing

2. Github Data Analysis Pipeline:

Analyzing your multiplexed targeted capture data
Other Targeted Sequencing & Barcode Solutions

3. PacBio Targeted Sequencing Applications Website:

PacBio Targeted Sequencing Applications

4. Shared Protocol:

Target Sequence Capture Using IDT Library with PacBio Barcoded Adapter

Acknowledgements

The authors would like to thank our collaborators at Integrated DNA Technologies (IDT) for their help with this study.