



The Landscape of SNCA Transcripts Across Synucleinopathies: New Insights From Long Reads Sequencing Analysis

Elizabeth Tseng¹, William J. Rowell¹, Omolara-Chinue Glenn^{2,3}, Ting Hon¹, Julio Barrera^{2,3}, Steve Kujawa¹ and Ornit Chiba-Falek^{2,3*}

¹Pacific Biosciences, Menlo Park, CA, United States, ²Department of Neurology, Duke University Medical Center, Durham, NC, United States, ³Center for Genomic and Computational Biology, Duke University Medical Center, Durham, NC, United States

OPEN ACCESS

Edited by:

Ioannis Dragatsis,
University of Tennessee Health
Science Center (UTHSC),
United States

Reviewed by:

Katrin Beyer,
Fundación Instituto de Investigación
Biomédica “Germans Trias
i Pujol”, Spain
Jordi Clarimon,
Sant Pau Institute for Biomedical
Research, Spain

*Correspondence:

Ornit Chiba-Falek
o.chibafalek@duke.edu

Specialty section:

This article was submitted to
Neurogenomics,
a section of the journal
Frontiers in Genetics

Received: 26 February 2019

Accepted: 04 June 2019

Published: 09 July 2019

Citation:

Tseng E, Rowell WJ, Glenn O-C,
Hon T, Barrera J, Kujawa S and
Chiba-Falek O (2019) The Landscape
of SNCA Transcripts Across
Synucleinopathies: New
Insights From Long Reads
Sequencing Analysis.
Front. Genet. 10:584.
doi: 10.3389/fgene.2019.00584

Dysregulation of alpha-synuclein expression has been implicated in the pathogenesis of synucleinopathies, in particular Parkinson’s Disease (PD) and Dementia with Lewy bodies (DLB). Previous studies have shown that the alternatively spliced isoforms of the SNCA gene are differentially expressed in different parts of the brain for PD and DLB patients. Similarly, SNCA isoforms with skipped exons can have a functional impact on the protein domains. The large intronic region of the SNCA gene was also shown to harbor structural variants that affect transcriptional levels. Here, we apply the first study of using long read sequencing with targeted capture of both the gDNA and cDNA of the SNCA gene in brain tissues of PD, DLB, and control samples using the PacBio Sequel system. The targeted full-length cDNA (Iso-Seq) data confirmed complex usage of known alternative start sites and variable 3’ UTR lengths, as well as novel 5’ starts and 3’ ends not previously described. The targeted gDNA data allowed phasing of up to 81% of the ~114 kb SNCA region, with the longest phased block exceeding 54 kb. We demonstrate that long gDNA and cDNA reads have the potential to reveal long-range information not previously accessible using traditional sequencing methods. This approach has a potential impact in studying disease risk genes such as SNCA, providing new insights into the genetic etiologies, including perturbations to the landscape the gene transcripts, of human complex diseases such as synucleinopathies.

Keywords: alternative splicing, targeted sequencing, long read sequencing, Parkinson’s Disease, PacBio, Iso-Seq, isoforms

INTRODUCTION

Transcriptional and posttranscriptional programs control gene expression levels and/or production of multiple distinct mRNA isoforms, and changes in these mechanisms result in dysregulation of gene expression and differential expression profiles. Aberrant transcriptional and posttranscriptional gene regulation is abundant in human nervous system tissues and contributes to phenotypic differences within and between individuals in health and disease.

Dysregulation of alpha-synuclein expression has been implicated in the pathogenesis of synucleinopathies, in particular Parkinson's Disease (PD) and Dementia with Lewy bodies and (DLB). While the role of *SNCA* overexpression in synucleinopathies, mainly PD, has been well established, here we focused on determination of the complete repertoire of *SNCA* transcript isoforms in different synucleinopathies. Previously, several different *SNCA* transcript isoforms have been described for *SNCA* gene, arisen from alternative splicing, transcriptional start sites (TSSs), and selection of polyadenylation sites (McLean et al., 2012; Xu et al., 2014). Alternative splicing of the coding exons gives rise to *SNCA* 140, *SNCA* 112, *SNCA* 126, and *SNCA* 98, resulting in four protein isoforms (Beyer and Ariza, 2012). Alternative TSSs of *SNCA* gene results in four different 5'UTRs, and alternative selection of different polyadenylation sites determines three major lengths of the 3'UTR, with no impact on the composition of the protein product (Beyer and Ariza, 2012). Our overarching goal is to gain new insights into the contribution of the different *SNCA* mRNA species, known and novel, to the pathogenesis and heterogeneity of synucleinopathies.

To date, most studies have used short read sequencing technologies to interrogate the transcriptome complexity in human brains. The availability of third generation long read technologies provides an unprecedented and nearly complete picture of isoform structures. However, existing long read transcript sequencing for human disease genes has used an amplicon-based approach (Treutlein et al., 2014; Kohli, 2017; Tseng et al., 2017). While this approach has been successful in identifying complex alternative splicing in human disease genes, it is limited to the PCR primer design and will not uncover alternative start and end sites. Targeted enrichment, such as through the use of IDT probes, can deliver comprehensive isoform view of genes of interest at low sequencing cost. Further, highly accurate full-length transcript reads enable isoform-specific haplotyping.

Here, we present the first known study using targeted capture of gDNA and cDNA of the *SNCA* gene region using PacBio SMRT sequencing. The *SNCA* gene region is ~114 kb long, consisting of six exons with transcript lengths around 3 kb. We multiplexed 12 human brain samples from PD, DLB, and normal control samples and sequenced the gDNA and cDNA library on the PacBio Sequel system. We describe the bioinformatics analyses used to identify SNPs, indels, and short tandem repeats for the gDNA capture, and isoform-level haplotyping for the cDNA data. We show that targeted capture is a cost-effective way of jointly studying genomic variation and alternative splicing in a disease-related neural gene.

MATERIALS AND METHODS

Study Samples

The study cohort ($N = 12$) consisted of individuals with three autopsy-confirmed neuropathological diagnoses: (1)

PD ($N = 4$); (2) DLB ($N = 4$); and (3) clinically and neuropathologically normal subjects ($N = 4$). Frontal cortex brain tissues were obtained through the Kathleen Price Bryan Brain Bank (KPBBB) at Duke University, the Banner Sun Health Research Institute Brain and Body Donation Program (Beach et al., 2015), and Layton Aging and Alzheimer's Disease Center at Oregon Health and Science University. Neuropathologic phenotypes were determined in *postmortem* examination following standard well-established methods following the method and clinical practice recommendations of McKeith and colleagues (McKeith et al., 1999, 2005). The density of the LB pathology (in a standard set of brain regions) received scores of mild, moderate, severe, and very severe. The study samples within each diagnosis group, PD and DLB, were carefully selected such that the severity of the clinicopathological phenotypes was similar within each pathology. All brains exhibited brainstem, limbic, and neocortical Lewy bodies (LBs), whereas PD showed severe to very severe McKeith scores in the sub-nigra and the amygdala. All brains indicate no AD according to CERAD criteria and Braak and Braak stage = II. The neurologically healthy brain samples were obtained from *postmortem* tissues of clinically normal subjects who were examined, in most instances, within 1 year of death and were found to have no cognitive disorder or parkinsonism and neuropathological findings insufficient for diagnosing PD, Alzheimer's disease (AD), or other neurodegenerative disorders. All samples were whites. Demographics and neuropathology for these subjects are summarized in **Supplementary Table 1**. The project was approved by the Duke Institution Review Board (IRB) which provided an ethical approval. The methods were carried out *in accordance with* the relevant guidelines and regulations.

Genomic DNA and RNA Extractions

Genomic DNA was extracted from brain tissues by the standard Qiagen protocol (Qiagen, Valencia, CA). Total RNA was extracted from brain samples (100 mg) using TRIzol reagent (Invitrogen, Carlsbad, CA) followed by purification with a RNeasy kit (Qiagen, Valencia, CA), following the manufacturer's protocol. gDNA and RNA concentration was determined spectrophotometrically, and the quality of the RNA samples and lack of significant degradation were confirmed by measurements of the RNA Integrity Number (RIN, **Supplementary Table 1**) utilizing an Agilent Bioanalyzer.

Library Preparation and Sequencing gDNA Capture Using IDT Xgen® Lockdown® Probes and Single-Molecule Sequencing

Approximately 2 μ g of each gDNA sample was sheared to 6 kb using the Covaris g-TUBE and ligated with barcoded adapters. An equimolar pool of 12-plex barcoded gDNA library (2 μ g total) was input into the probe based capture with a custom designed *SNCA* gene panel.

A SMRTBell library was constructed using 626 ng of captured and re-amplified gDNA¹.

cDNA Capture Using IDT Xgen® Lockdown® Probes and Single-Molecule Isoform-Sequencing (Iso-Seq)

About 100–150 ng of total RNA per reaction was reverse transcribed using the Clontech SMARTer cDNA synthesis kit and 12 sample specific barcoded oligo dT (with PacBio 16mer barcode sequences, see **Supplementary Methods**). Three reverse transcription (RT) reactions were processed in parallel for each sample. PCR optimization was used to determine the optimal amplification cycle number for the downstream large-scale PCR reactions. A single primer (primer IIA from the Clontech SMARTer kit 5' AAG CAG TGG TAT CAA CGC AGA GTA C 3') was used for all PCR reactions post-RT. Large-scale PCR products were purified separately with 1X AMPure PB beads, and the bioanalyzer was used for QC. An equimolar pool of 12-plex barcoded cDNA library (1 µg total) was input into the probe-based capture with a custom designed SNCA gene panel.

A SMRTBell library was constructed using 874 ng of captured and re-amplified cDNA². One SMRT Cell 1M (6 h movie) was sequenced on the PacBio Sequel platform using 2.0 chemistry.

gDNA Analysis

Sequencing of the barcoded gDNA data was run on three SMRT Cells 1M using 2.0 chemistry. The data were demultiplexed by running the Demultiplex Barcodes application in PacBio SMRT Link v6.0.

Short Variant Analysis and Phasing

Circular Consensus Sequence (CCS) reads were generated using SMRT Analysis 6.0 from each demultiplexed data set and aligned to the hg19 reference genome using minimap2. PCR duplicates from post-capture amplification were identified by mapping endpoints and tagged using a custom script. Short variants were called using GATK4 HaplotypeCaller (GATK4 HC) (Poplin et al., 2018). After a first pass of filtering using coverage depth and quality metrics, variants were manually inspected in IGV³. If variants did not phase with nearby SNPs, they were manually filtered. The variant sites that passed manual curation were used in conjunction with the deduplicated CCS alignments for read-backed phasing with WhatsHap (Martin et al., 2016).

Clustering and Determining Haplotypes for CT-Rich Region

Subsequences aligned to chr4: 90742331-90742559 (hg19) were extracted for each sample. After inspecting the size distribution of these subsequences, they were clustered by

size and sequence similarity using a combination of python and MUSCLE (Edgar, 2004), and a consensus sequence was generated independently for each cluster.

Custom scripts and workflows further described in <https://github.com/williamrowell/Long-reads-Sequencing-of-SNCA-in-Diseases>.

Isoform Analysis

Sequencing of the barcoded cDNA data was on one SMRT Cell 1M on the PacBio Sequel system using 2.0 chemistry. Bioinformatics analysis was done using the IsoSeq3 application in the PacBio SMRT Analysis v6.0.0 to obtain high-quality, full-length isoform sequences (see **Supplementary Methods** for more information).

Isoform SNP Calling

Full-length reads associated with the final 41 isoforms from all 12 samples were aligned to the hg19 genome to create a pileup. Bases with QV less than 13 were excluded. Then, at each position with at least 40 base coverage, a Fisher exact test with Bonferroni correction is applied with a *p* cutoff of 0.01. Only substitution SNPs not close to homopolymer regions (stretches of 4 or more of the same nucleotide) were called. After SNP calling, the genotype for each sample was determined by tallying the number of supporting sample-specific full length (FL) reads. If a sample had 5+ FL reads supporting both reference and alternative base, it was heterozygous. If a sample had 5+ FL reads supporting one allele and fewer than 5 FL reads for the other, it was homozygous. Otherwise, it was inconclusive. Scripts are available at: https://github.com/Magdoll/cDNA_Cupcake.

RESULTS

We designed custom probes for the SNCA gene and performed targeted capture of both gDNA and cDNA on a multiplexed library consisting of 12 human brain samples from PD, DLB, and normal controls (**Figure 1, Supplementary Table 1**). The gDNA and cDNA libraries were sequenced on the PacBio Sequel platform. Bioinformatics analysis was done using PacBio software followed by custom analysis.

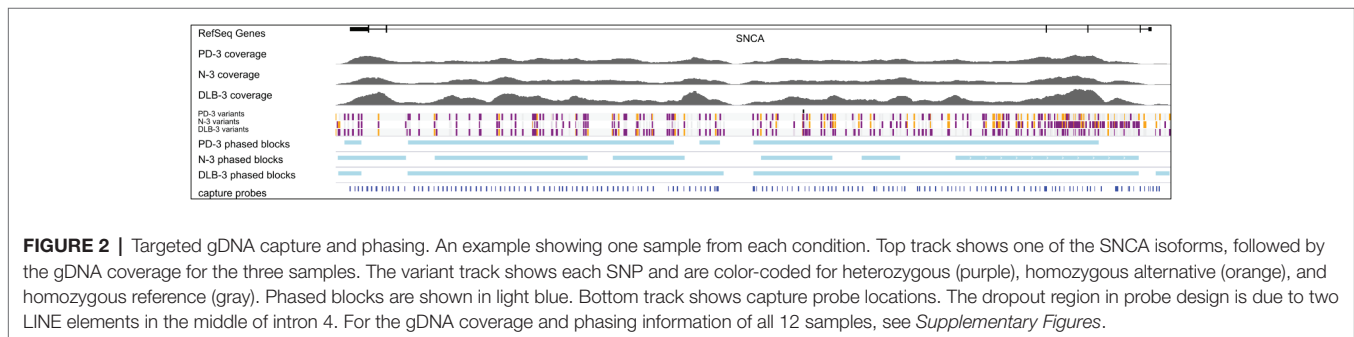
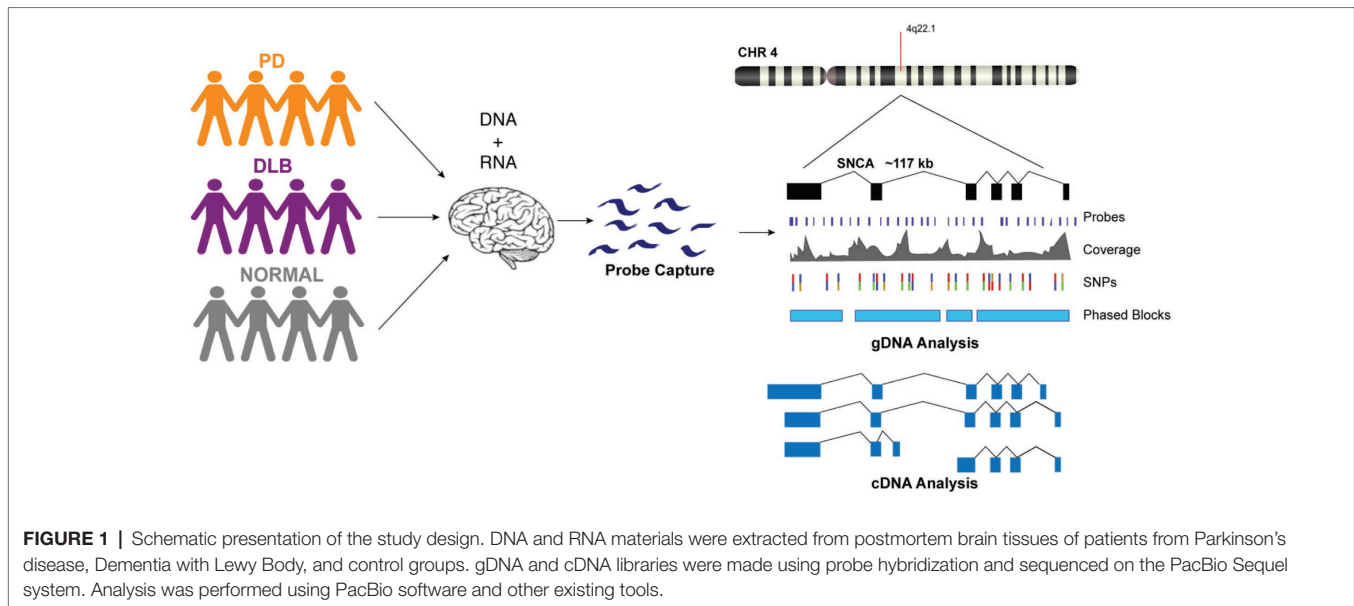
Targeted gDNA Capture Identified Known and Novel Variations

After generating circular consensus sequences (CCS) and removing PCR duplicates (**Supplemental Methods**), we obtained 16- to 71-fold mean unique coverage of the SNCA gene region. The CCS reads had a mean insert length of 2.9 kb and a mean read accuracy of 98.9%. With the exception of a 5 kb region intentionally uncovered by probes due to the presence of LINE elements (hg19 chr4: 90697216-90702113) and a 2.1 kb region of high GC content around exon 1, there was sufficient coverage to genotype both haplotypes for each of the 212 samples (**Figure 2, Supplementary Figure 1**).

¹<https://www.pacb.com/wp-content/uploads/Procedure-Checklist-%E2%80%93-Multiplex-Genomic-DNA-Target-Capture-Using-IDT-xGen-Lockdown-Probes.pdf>

²<https://www.pacb.com/wp-content/uploads/Procedure-Checklist-%E2%80%93-cDNA-Capture-Using-IDT-xGen-Lockdown-Probes.pdf>

³<http://www.broadinstitute.org/igv>



Using GATK4 HC, quality-based filtering, and manual curation, we identified 282 SNPs and 35 indels, including 8 SNPs and 13 indels not found in dbSNP (dbSNP Build ID: human_9606_b150_GRCh37p13) (**Supplementary Table 2**). No variants were identified in the coding region for SNCA, although eight variants were identified in untranslated regions. The majority of the identified variants, including several short tandem repeats (STR), fall within introns 2, 3, and 4.

We have previously described a highly polymorphic CT-rich region in intron 4 of SNCA with four observed haplotypes (Lutz et al., 2015). While this highly repetitive and structurally variable region proved difficult to genotype with GATK4 HC, we were able to construct consensus sequences for all 12 samples and observed all 4 of the previously discovered haplotypes (**Supplementary Figure 2**). Additionally, we identified a novel STR in intron 4 consisting of a three-base unit repeated 16 times in the reference. Within the 12 samples, we identified three haplotypes, with 9, 12, and 15 copies of the TTG repeat unit. GATK HC correctly genotyped all of these except for one haplotype for PD-4, which had fairly low coverage in this region. However, with the given data for this sample, the genotype can be determined by visual inspection (**Table 1**).

TABLE 1 | A novel triplet tandem repeat in intron 4 (chr4: 90713442).

Sample	Genotype
PD-1	(TTG) ₁₅ /(TTG) ₁₅
PD-2	(TTG) ₁₂ /(TTG) ₁₂
PD-3	(TTG) ₁₂ /(TTG) ₁₂
PD-4	(TTG) ₁₅ */(TTG) ₁₂
N-1	(TTG) ₁₅ /(TTG) ₁₂
N-2	(TTG) ₁₂ /(TTG) ₁₂
N-3	(TTG) ₁₂ /(TTG) ₁₂
N-4	(TTG) ₁₂ /(TTG) ₁₂
DLB-1	(TTG) ₁₂ /(TTG) ₁₂
DLB-2	(TTG) ₁₂ /(TTG) ₁₂
DLB-3	(TTG) ₁₅ /(TTG) ₉
DLB-4	(TTG) ₁₂ /(TTG) ₁₂

*PD-4 is incorrectly genotyped by GATK4HC but can be genotyped by visual inspection.

The reference has 16 repeats. The table shows the repeat number of both haplotypes for each sample.

We used the short variants detected by GATK HC in conjunction with the read-based phasing tool WhatsHap (Martin et al., 2016) to phase the CCS reads across the locus, with a range of success driven mostly by the heterozygous variant

density over the locus. Samples PD-1, PD-4, N-4, DLB-1, and DLB-4 had long stretches of low heterozygosity, with very few, short phase blocks, while the other samples yielded phase blocks ranging from 7 to 18 times the mean read length, up to 54 kb (Supplementary Figure 3).

Targeted cDNA Capture Identified Novel Start and End Sites

We processed the PacBio cDNA (Iso-Seq) data using the PacBio SMRT Analysis software. After mapping the Iso-Seq data to hg19 and removing artifacts (Supplementary Table 3, Supplementary Figure 4), we obtained a final set of 41 SNCA isoforms (Figure 3). All final isoforms have all canonical splice sites (GT-AG or GC-AG) and are supported by a total of 20 or more full-length reads. The majority of the isoforms (28 out of 41) have all six exons, differing only in the use of alternative 5' start sites and 3' UTR lengths. The 3' UTR lengths varied between 300 and 2.6 kb. The use of highly diverse alternative 5' start site in SNCA is known; what is less known is the variable 3' UTR length, which had been previously studied using RNA-seq data that did not resolve full-length isoform structures (Rhinn et al., 2012). The Iso-Seq data show that the variable 3' UTR length

seems paired with all possible combinations of 5' start sites with no preferential coupling. Almost none of the variability in start and end site changes the predicted open reading frame (Supplementary Figure 5) and is predicted to translate to the canonical 141 amino acid sequence.

We further validated the novel (but canonical) junctions using publicly available short read junction data. The Intropolis (v1, <https://github.com/nellore/intropolis>) database combines over 21,000 publicly available RNA-seq. Due to the high volume of junction data supported by only a single short read, for this study, we require a minimum of 10 short read support (combined from all >21,000 RNA-seq datasets) to confirm our Iso-Seq novel junctions. With the exception of the novel junctions for PB.1016.253 and PB.1016.296 (Figure 3), all other novel junctions are supported by the Intropolis data set. Interestingly, these novel junctions have significantly less short read support than the Gencode-annotated junctions. For example, the two novel junctions in PB.1016.139 introduced by the novel exon are supported by 2,519 and 44 Intropolis short read counts, respectively, whereas the other four known junctions are supported by over 1 million short read counts. This shows the power of targeted enrichment using full-length transcriptome sequencing for detecting rare, novel isoforms.

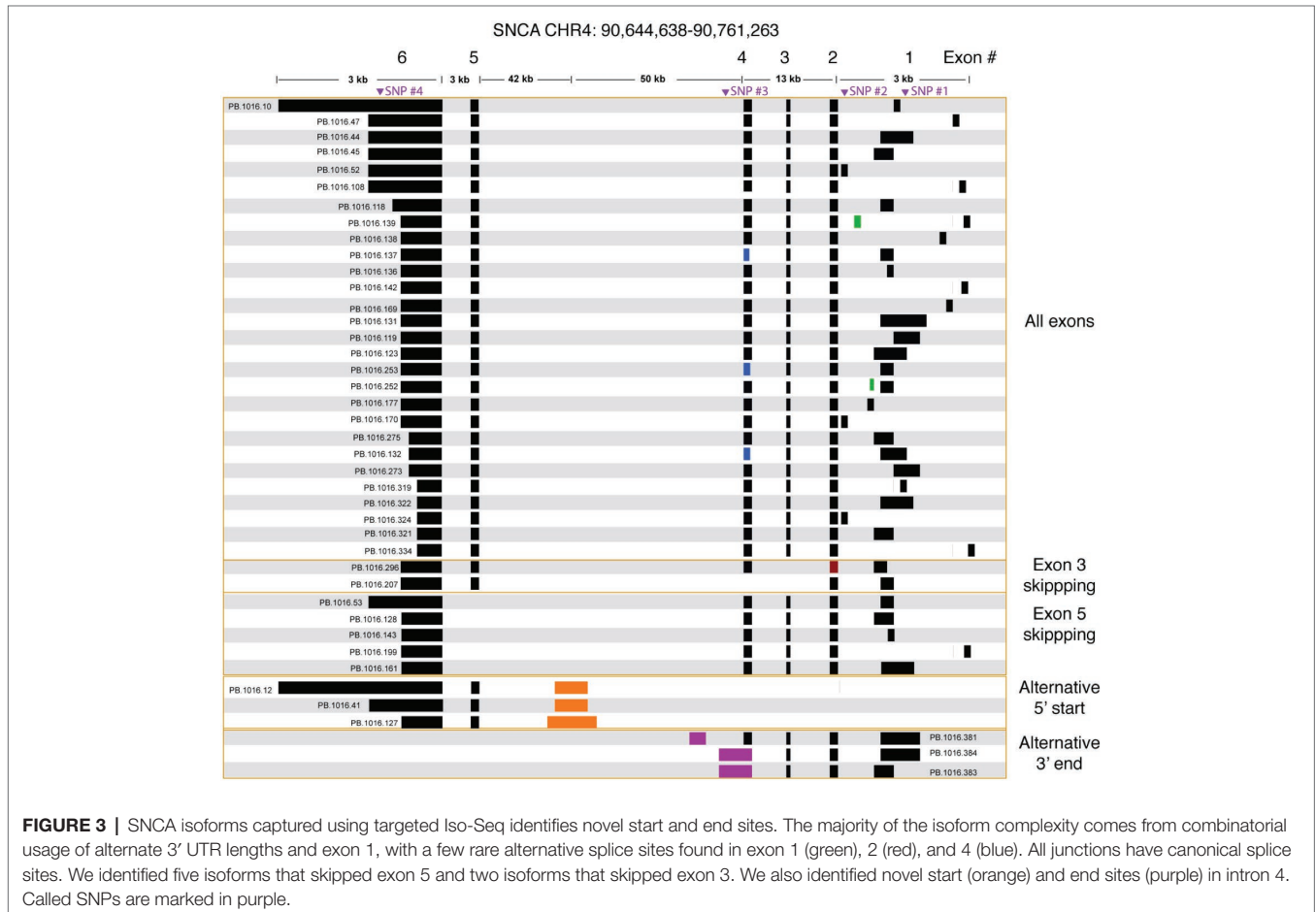


TABLE 2 | SNCA isoform abundance for each sample, aggregated by splicing patterns.

GROUP	PD-1 (%)	PD-2 (%)	PD-3 (%)	PD-4 (%)	N-1 (%)	N-2 (%)	N-3 (%)	N-4 (%)	DLB-1 (%)	DLB-2 (%)	DLB-3 (%)	DLB-4 (%)
AllExons	95.92	97.06	95.82	95.76	98.10	97.73	96.63	96.68	96.00	94.59	94.58	96.21
Skip3	0.00	0.00	0.03	0.08	0.03	0.09	0.02	0.25	0.08	0.13	0.24	0.00
Skip5	2.50	1.47	1.68	1.10	0.79	0.93	1.00	1.32	2.04	1.42	1.56	1.89
Alt5	1.02	0.00	1.61	2.76	0.64	0.77	1.85	0.08	0.94	2.83	3.16	0.79
Alt3	0.56	1.47	0.85	0.30	0.44	0.49	0.51	1.67	0.94	1.03	0.47	1.10

The abundance for each isoform is the fraction of on-target, full-length reads associated with that isoform.

Isoforms are grouped by their splice patterns (see **Figure 3**).

AllExons, isoforms expressing all 6 exons; Skip3, isoforms skipping exon 3; Skip5, isoforms skipping exon 5; Alt5, isoforms with the alternative start site in intron 4; Alt3, isoforms with the alternative end site in intron 4.

The full abundance for each isoform is shown in Supplementary Table and Supplementary Data.

We observed two isoforms with exon 3 skipping (SNCA126) and five isoforms with exon 5 skipping (SNCA112). Again, the splicing diversity in these two exon skipping groups mostly comes from the diverse use of alternative 5' start sites and variable 3' UTR length. ORF prediction shows skipping exon 3 or exon 5 shortens the ORF but maintains the reading frame. Three isoforms have novel 3' end sites located in intron 4. ORF prediction shows that this results in truncated protein product.

We identified a previously unannotated 5' start site located in intron 4 (hg19 chr4: 90692548-90693045, **Figure 3**). The three isoforms associated with this novel start consists of the novel start site, exon 5, and variable 3' UTR lengths. Interestingly, while publicly downloaded short read data from GTEx and Sandor et al. (2017) and CAGE peak data (FANTOM5) did not support this novel start site, a recent public NA12878 direct RNA data set⁴ contained only one SNCA transcript that confirmed this alternative start site. Further, the novel junction between exon 5 and the novel start site is confirmed by Intropolis short read junction data. Interestingly, this novel 5' start site is predicted to introduce novel peptides while maintaining the reading frame in exon 5.

We also identified three SNCA transcripts with new end sites (**Figure 3**). Two isoforms (PB.1016.383, PB.1016.384) used an extended 3' UTR in exon 4, while the third isoform (PB.1016.381) used a novel 3' exon in intron 4. The novel junctions between the novel last exon and the previous exon are supported by public short read junction data (Intropolis). The novel 3' UTRs result in a truncated ORF prediction.

Using the normalized full-length read count as a proxy for isoform abundance, we find one of the canonical SNCA isoforms (PB.1016.131) to be the most abundant, with an abundance of 50–60% across all subject samples (**Supplementary Table 4**). We further grouped the 41 isoforms by their splicing patterns (**Table 2**). Isoforms that have all six exons account for 95–97% of the abundance. Previous studies have shown a marked expression increase in isoforms missing exon 3 (SNCA126) in the frontal cortex of DLB samples compared to normal (Beyer et al., 2008); our aggregated isoform counts shows that three of the DLB samples have a slightly elevated count level compared to the normal samples as well as the SNCA112 (exon 5 skipping) variants for PD and DLB against normal samples.

Full-Length cDNA Enables Isoform-Level Phasing Information

We called SNPs using cDNA by piling up all full-length reads from the 12 samples to call variants (see Section “Methods”). A total of four SNPs were called and all were previously annotated in dbSNP (**Table 3, Figure 3**). The four SNPs are all located in non-CDS regions, one in the 3' UTR (exon 6), one in intron 4, and two in the 5' UTR (exon 1). The 3' UTR SNP (chr4: 90646886) is only covered by isoforms with a 3' UTR that is at least ~1 kb long, and hence, not all canonical isoforms cover this SNP. The intron 4 SNP (chr4: 90743331) is only covered by the novel alternative 3' end isoforms (PB.1016.383, PB.1016.384) and is not connected to any of the other SNPs. The two 5' UTR SNPs (chr4: 90757312 and chr4: 90758389) are covered by two mutually exclusive exon 1 usage and hence are also not linked.

Our current approach is limited to calling only substitution variants in transcribed regions with sufficient coverage. Comparing the list of our SNPs with the hg19 dbSNP annotation shows that most of the SNPs or variants missed were either less than 1% frequency in the population, were not single nucleotide substitutions, or adjacent to low complexity regions. For example, rs77964369 (chr4: 90646532) is reported to have 50/50 frequency of T/A; however, this T is adjacent to a stretch of 11 genomic As downstream. Manual inspection of the Iso-Seq read pileup, which has ~1,300 reads at this site, does not suggest evidence of variation at least amongst our 12 samples.

Using the sample-specific reads, we call the genotype of each sample at each SNP location (**Table 3**). Besides PD-2 having too few reads and is inconclusive for all four SNPs, we were able to call the genotype for most other samples. Notably, DLB-3 was the only sample that is heterozygous at all SNP locations. Otherwise, we did not observe any condition-specific pattern of preferring one genotype to other.

DISCUSSION

We describe the first study using targeted enrichment of the SNCA gene on multiplexed gDNA and cDNA libraries for studying neurological diseases using long read sequencing. The long read lengths of the PacBio Sequel system facilitated the sequencing of the full-length transcript isoforms repertoire of

⁴<https://github.com/nanopore-wgs-consortium/NA12878/>

TABLE 3 | cDNA SNP information.

#	Coord	dbSNP	Annotation	Ref	Alt	Homo_Ref	Homo_Alt	Het	Inconclusive
1	90758389	rs2301135	exon1 (5' UTR)	G	C	PD-1, DLB-1	PD-3, DLB-1, DLB-2	PD-4, N-1, N-2, N-3, DLB-3	PD-2, DLB-4
2	90757312	rs2870027	exon1 (5' UTR)	C	T		PD-3, PD-4, N-1, N-2, N-3, N-4, DLB-1, DLB-2	DLB-3	PD-1, PD-2, DLB-4
3	90743331	rs10005233	intron4 (alt 3' UTR)	C	T	PD-1, PD-4, N-4, DLB-4	PD-3, DLB-1, DLB-2	N-1, N-2, N-3, DLB-3	PD-2
4	90646886	rs356165	exon6 (3' UTR)	G	A	N-4, DLB-4	PD-1, DLB-1	PD-3, N-1, N-2, N-3, DLB-2, DLB-3	PD-2

SNPs were called using full-length reads from the Iso-Seq data. For each sample, the number of FL reads supporting either the reference or alternative base was tabulated. If both alleles have 5+ FL read support, it is called heterozygous; if one allele has 5+ reads and the other has <5, homozygous; otherwise, inconclusive.

the SNCA gene. We revealed the diversity in the use of alternative 5' start sites and variable 3' UTR lengths and observed known exon skipping events, such as exon 3 deletion (SNCA126) and exon 5 deletion (SNCA112). Additionally, new alternative start and end sites within the large intron 4 were identified that are predicted to be translated to novel proteins. It is likely that the high depth of sequencing coverage of targeted capture, in combination with the ability to sequence complete transcripts, allowed us to detect these previously undescribed isoforms.

The biological and pathological significance of the different SNCA protein isoforms has yet to be fully discovered. However, specific SNCA post translational modification and splicing isoforms have been associated with intracellular aggregation propensities (Kalivendi et al., 2010) and are differently expressed in human synucleinopathies (Beyer et al., 2008; Beyer and Ariza, 2012). Studies of SNCA post translational modification showed that Lewy bodies, the pathological hallmark of synucleinopathies, contain abundant phosphorylated, nitrated, and monoubiquitinated SNCA (Kim et al., 2014). The effects of post-transcriptional modification on SNCA aggregation have also been studied. Alternative splicing was suggested to affect SNCA aggregation. A deletion of either exon 3 or 5 predicts functional consequences: while exon 3 deletion (SNCA126) leads to the interruption of the N-terminal protein-membrane interaction domain which may lead to less aggregation, and exon 5 deletion (SNCA112) may result in enhanced aggregation

due to a significant shortening of the unstructured C-terminus (Lee et al., 2001; Beyer, 2006). In the frontal cortex of DLB, SNCA112 is increased markedly compared to the controls (Beyer et al., 2008), while SNCA126 levels are decreased in the prefrontal cortex of DLB patients (Beyer et al., 2006). In contrast, SNCA126 expression showed increased in the frontal cortex of PD brains and no significant differences in MSA (Beyer et al., 2008). SNCA98 is a brain specific splice variant that lacks both exon 3 and 5 and exhibits different expression levels in various areas of fetal and adult brain. Overexpression of SNCA98 has been reported in DLB, PD (Beyer et al., 2007), and MSA (Beyer et al., 2008) frontal cortices compared with controls. In addition, the post transcriptional process resulting in alternative 3'UTR usage was reported to have effects on the mRNA stability and localization (Fabian et al., 2010; Rhinn et al., 2012; Yeh and Yong, 2016). Further investigation regarding the aggregation propensities of the different known SNCA protein isoforms and the composition of Lewy bodies are warranted. Furthermore, our study laid the groundwork for mRNA quantification analyses of the previously known and novel transcripts in a larger sample size comprised of subjects with a range of clinicopathological stages using several brain regions from each subject. These analyses of the brain region-specific transcriptomic landscape of SNCA in the context of neuropathological severity will be informative with respect to the role of specific SNCA transcript isoforms in the progression

of the neuropathological stages and the severity of the Lewy bodies and Lewy neurites density.

In this paper, we focused on creating a sequencing and analysis standard for analyzing targeted gDNA and cDNA data generated from the same subjects. This is a powerful approach that potentially allows the phasing of the gDNA sequences across the complete region of a particular gene based on heterozygosity in the sequence of the full-length transcript isoforms. The PacBio targeted gDNA data in this study produced phased blocks that covered 81% of the 114 kb region centered on SNCA, with the longest phased block exceeding 54 kb. As gDNA phasing is limited by read length and heterozygosity, increasing read lengths will likely generate larger phase blocks.

gDNA variant analysis confirmed known and identified novel short tandem repeats (STRs) in the intronic regions. For example, previously, using phased sequencing by cloning and Sanger sequencing, we discovered four distinct haplotypes within an intronic CT-rich region that comprised of a cluster of variable repetitive sequences (Lutz et al., 2015). We showed that a specific haplotype, termed haplotype 3, conferred risk to develop Lewy body pathology in Alzheimer's patients. Here, we validated the sequence of this highly polymorphic low-complexity region and its four defined haplotypes. Although our sample size was small, "haplotype 3" was present exclusively in disease patients (one PD patient, two DLB patients), consistent with our previous findings. The pilot results and our previous publication provide the premise to repeat the association analyses of synucleinopathies with accurately defined, i.e., by long reads, STRs and structural haplotypes using a larger sample size.

Our paper demonstrated the ability of the PacBio Sequel system to discover novel full-length transcripts and characterize the complete full-length transcript repertoire of a gene implicated in a disease. Furthermore, we also showed that long reads gDNA define more accurately short structural variants and haplotypes including STRs and by that can facilitate the discovery and validation of disease associated variants other than SNPs. Collectively, this new knowledge is highly valuable and applicable in advancing our understanding of the genetic etiologies, that may involve perturbations in the transcript landscape, underlying complex human diseases including age-related neurodegenerative disorders such as synucleinopathies.

REFERENCES

- Beach, T. G., Adler, C. H., Sue, L. I., Serrano, G., Shill, H. A., Walker, D. G., et al. (2015). Arizona study of aging and neurodegenerative disorders and brain and body donation program. *Neuropathology* 35, 354–389. doi: 10.1111/neup.12189
- Beyer, K. (2006). A-synuclein structure, posttranslational modification and alternative splicing as aggregation enhancers. *Acta Neuropathol.* 112, 237–251. doi: 10.1007/s00401-006-0104-6
- Beyer, K., and Ariza, A. (2012). Alpha-synuclein posttranslational modification and alternative splicing as a trigger for neurodegeneration. *Mol. Neurobiol.* 47, 509–524. doi: 10.1007/s12035-012-8330-5
- Beyer, K., Domingo-Sábat, M., Humbert, J., Carrato, C., Ferrer, I., and Ariza, A. (2008). Differential expression of alpha-synuclein, parkin, and synphilin-1 isoforms in Lewy body disease. *Neurogenetics* 9, 163–172. doi: 10.1007/s10048-008-0124-6

DATA AVAILABILITY

The three SMRT cells of gDNA raw data is available at Zenodo.org with doi: 10.5281/zenodo.1560688. The one SMRT cell of cDNA raw data is available at Zenodo.org with doi: 10.5281/zenodo.1581809. The processed gDNA and cDNA results, including gDNA variants and cDNA isoforms, are available at Zenodo.org with doi: 10.5281/zenodo.3261805.

AUTHOR CONTRIBUTIONS

OC-F contributed conception and design of the study. ET and WR organized sequences databases, performed the sequencing analyses and prepared all figures and tables. O-CG and JB handled the brain tissues and nucleic sample preparations. TH generated the sequencing data sets. SK designed and obtained the reagents. OC-F, ET and WR wrote the first draft of the manuscript. OC-F obtained funding. All authors contributed to manuscript preparations, read and approved the submitted version.

FUNDING

This work was funded in part by the National Institutes of Health/National Institute of Neurological Disorders and Stroke (NIH/NINDS) [R01 NS085011 to OC-F].

ACKNOWLEDGMENTS

This manuscript has been released as a pre-print at BioRxiv (Tseng et al., 2019). <https://www.biorxiv.org/content/10.1101/524827v1>.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2019.00584/full#supplementary-material>

- Beyer, K., Domingo-Sábat, M., Lao, J. I., Carrato, C., Ferrer, I., and Ariza, A. (2007). Identification and characterization of a new alpha-synuclein isoform and its role in Lewy body diseases. *Neurogenetics* 9, 15–23. doi: 10.1007/s10048-007-0106-0
- Beyer, K., Humbert, J., Ferrer, A., Lao, J. I., Carrato, C., pez, D. L., et al. (2006). Low alpha-synuclein 126 mRNA levels in dementia with Lewy bodies and Alzheimer disease. *Neuroreport* 17, 1327–1330. doi: 10.1097/01.wnr.0000224773.66904.e7
- Fabian, M. R., Sonenberg, N., and Filipowicz, W. (2010). Regulation of mRNA translation and stability by microRNAs. *Annu. Rev. Biochem.* 79, 351–379. doi: 10.1146/annurev-biochem-060308-103103
- Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340
- Kalivendi, S. V., Yedlapudi, D., Hillard, C. J., and Kalyanaraman, B. (2010). Oxidants induce alternative splicing of A-synuclein: implications for Parkinson's disease. *Free Radic. Biol. Med.* 48, 377–383. doi: 10.1016/j.freeradbiomed.2009.10.045

- Kim, W. S., Gedal, K. K., and Halliday, G. M. (2014). Alpha-synuclein biology in Lewy body diseases. *Alzheimers Res. Ther.* 6, 1–9. doi: 10.1186/s13195-014-0073-2
- Kohli, M. (2017). Androgen receptor variant AR-V9 is coexpressed with AR-V7 in prostate cancer metastases and predicts abiraterone resistance. *Clin. Cancer Res.* 23, 1–13. doi: 10.1158/1078-0432.CCR-17-0017
- Lee, H.-J., Choi, C., and Lee, S. J. (2001). Membrane-bound A-synuclein has a high aggregation propensity and the ability to seed the aggregation of the cytosolic form. *J. Biol. Chem.* 277, 671–678. doi: 10.1074/jbc.M107045200
- Lutz, M. W., Saul, R., Linnertz, C., Glenn, O.-C., Roses, A. D., and Chiba-Falek, O. (2015). A cytosine-thymine (CT)-rich haplotype in intron 4 of SNCA confers risk for Lewy body pathology in Alzheimer's disease and affects SNCA expression. *Alzheimers Dement.* 11, 1133–1143. doi: 10.1016/j.jalz.2015.05.011
- Martin, M., Patterson, M., Garg, S., Fischer, S. O., Pisanti, N., Klau, G. W., et al. (2016). WhatsHap: fast and accurate read-based phasing. *bioRxiv* [Preprint]. doi: 10.1101/085050
- McKeith, I. G., Dickson, D. W., Lowe, J., Emre, M., O'Brien, J. T., Feldman, H., et al. (2005). Diagnosis and management of dementia with Lewy bodies: third report of the DLB consortium. *Neurology* 65, 1863–1872. doi: 10.1212/01.wnl.0000187889.17253.b1
- McKeith, I. G., Perry, E. K., and Perry, R. H. (1999). Report of the second dementia with Lewy body international workshop: diagnosis and treatment. Consortium on dementia with Lewy bodies. *Neurology* 53, 902–905. doi: 10.1212/WNL.53.5.902
- McLean, J. R., Hallett, P. J., Cooper, O., Stanley, M., and Isacson, O. (2012). Transcript expression levels of full-length alpha-synuclein and its three alternatively spliced variants in Parkinson's disease brain regions and in a transgenic mouse model of alpha-synuclein overexpression. *Mol. Cell. Neurosci.* 49, 230–239. doi: 10.1016/j.mcn.2011.11.006
- Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Van der Auwera, G. A., et al. (2018). Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* [Preprint]. doi: 10.1101/201178
- Rhinn, H., Qiang, L., Yamashita, T., Rhee, D., Zolin, A., Vanti, W., et al. (2012). Alternative A-synuclein transcript usage as a convergent mechanism in Parkinson's disease pathology. *Nat. Commun.* 3, 889–821. doi: 10.1038/ncomms2032
- Sandor, C., Robertson, P., Lang, C., Heger, A., Booth, H., Vowles, J., et al. (2017). Transcriptomic profiling of purified patient-derived dopamine neurons identifies convergent perturbations and therapeutics for Parkinson's disease. *Hum. Mol. Genet.* 54, ddw412–ddw415. doi: 10.1093/hmg/ddw412
- Treutlein, B., Gokce, O., Quake, S. R., and Südhof, T. C. (2014). Cartography of neurexin alternative splicing mapped by single-molecule long-read mRNA sequencing. *Proc. Natl. Acad. Sci.* 111, E1291–E1299. doi: 10.1073/pnas.1403244111
- Tseng, E., Rowell, W. J., Omolara-Chinue, G., Hon, T., Barrera, J., Kujawa, S., et al. (2019). The landscape of SNCA transcripts across synucleinopathies: new insights from long reads sequencing analysis. *bioRxiv* [Preprint]. doi: 10.1101/524827
- Tseng, E., Tang, H.-T., AlOlaby, R. R., Hickey, L., and Tassone, F. (2017). Altered expression of the FMR1 splicing variants landscape in premutation carriers. *Biochim. Biophys. Acta Gene. Regul. Mech.* 1860, 1117–1126. doi: 10.1016/j.bbagr.2017.08.007
- Xu, W., Tan, L., and Yu, J.-T. (2014). The link between the SNCA gene and Parkinsonism. *Neurobiol. Aging* 36, 1–14. doi: 10.1016/j.neurobiolaging.2014.10.042
- Yeh, H.-S., and Yong, J. (2016). Alternative polyadenylation of mRNAs: 3'-untranslated region matters in gene expression. *Mol. Cell* 39, 281–285. doi: 10.14348/molcells.2016.0035

Conflict of Interest Statement: ET, WR, TH, and SK are or were employees of Pacific Biosciences at the time of the study.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Tseng, Rowell, Glenn, Hon, Barrera, Kujawa and Chiba-Falek. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.