Title

Bioinformatic characterization of genomic and transcriptomic diversity in the human brain

Author

Liu, Christine Sara

Publication Date

2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA SAN DIEGO

Bioinformatic characterization of genomic and transcriptomic diversity in the human brain

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Biomedical Sciences

by

Christine Sara Liu

Committee in charge:

Professor Jerold Chun, Chair
Professor William Joiner, Co-Chair
Professor Vineet Bafna
Professor Jin Zhang
Professor Kun Zhang

2022

The Dissertation of Christine Sara Liu is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

DEDICATION

To my mom, dad, and sister Andrea. Thank you for all your love and support and for never letting me doubt myself.

EPIGRAPH

"Science has taught me that everything is more complicated than we first assume, and that being able to derive happiness from discovery is a recipe for a beautiful life."

Hope Jahren

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank Dr. Jerold Chun for welcoming me into the lab with much aloha and enthusiasm. His encouragement and trust through the years has given me the confidence to pursue my ideas and learn new skills.

I'd also like to acknowledge my committee for their support and for lending an ear when I needed to figure out next steps.

Many many thanks the entire Chun Lab past and present for creating and being part of a wonderful community of colleagues I consider to be good friends. When I started grad school, I strongly prioritized finding a friendly lab environment to call home, and I couldn't have asked for a better group of people to work with day in and day out.

I'd also like to thank my family for all their love and support. Without them, I wouldn't be who I am today, and I am grateful for their encouragement and continual confirmation that I've found where I belong. I'd especially like to thank my sister Andrea for being my "wombate-turned-roommate" and listening to all my complaints and dealing with all my stress. I want to also thank Nick for all his support and encouragement through our respective med and grad school journeys.

Chapter 2, in full, is a reprint of the material as it appears in *Nature* 2020. Lee, M-H.*, Liu, C.S.*, Zhu, Y., Kaeser, G.E., Rivera, R., Romanow, W.J., Kihara, Y., Chun, J., Reply: Evidence that *APP* gene copy number changes reflect recombinant vector contamination. The dissertation author was a co-primary researcher and author of this paper.

Chapter 3, in part, is currently being prepared for submission for publication of the material. Liu, C.S., Zhu, Y., Chun, J. The dissertation author was the primary researcher and author of this material.

Chapter 4, in full, is a reprint of the material as it appears in *PNAS* 2021. Palmer, C.R.*, Liu, C.S.*, Romanow, W.J., Lee, M-H., Chun, J. Altered cell and RNA isoform diversity in aging Down syndrome brains revealed by snRNA-seq. The dissertation author was a co-primary researcher and author of this paper.

Chapter 6, in part, is currently being prepared for submission for publication of the material. Liu, C.S., Chun, J. The dissertation author was the primary researcher and author of this material.

Chapter 7, in part, is currently being prepared for submission for publication of the material. Park, C., Liu, C.S., Ngo, T., Saikumar, J., Palmer, C.R., Costantino, I., Romanow, W.J., Chun, J. The dissertation author was a co-primary researcher and author of this material.

VITA

2016	Bachelor of Science in Chemical Biology, Minor in Computer Science, University of California, Berkeley
2022	Doctor of Philosophy in Biomedical Sciences, University of California San Diego

PUBLICATIONS

Palmer, C.R.*, **Liu, C.S.***, Romanow, W.J., Lee, M-H., Chun, J. Altered cell and RNA isoform diversity in aging Down syndrome brains revealed by snRNA-seq. *PNAS*. **2021**, *118*(47), e2114326118. **[*co-first authors]**

Bakken, T.E., Jorstad, N.L., Hu, Q., Lake, B.B., [et al. including **Liu, C.S.**]. Comparative cellular analysis of motor cortex in human, marmoset and mouse. *Nature*. **2021**, *598*, 111-119.

BRAIN Initiative Cell Census Network (BICCN)., A multimodal cell census and atlas of the mammalian primary motor cortex. *Nature*. **2021**, *598*, 86-102.

Lee, M-H.*, **Liu, C.S.***, Zhu, Y., Kaeser, G.E., Rivera, R., Romanow, W.J., Kihara, Y., Chun, J. Reply: Evidence that *APP* gene copy number changes reflect recombinant vector contamination. *Nature*. **2020**, *584*, E29-E33 **[*co-first authors]**

Rohrback, S., April, C., Kaper, F., Rivera, R.R., **Liu, C.S.**, Siddoway, B., Chun, J. Submegabase copy number variations arise during cerebral cortical neurogenesis as revealed by single-cell whole-genome sequencing. *PNAS*. **2018**, *115*(42), 10804-10809.

Lee, M-H., Siddoway, B., Kaeser, G.E., Segota, I., Rivera, R.R., Romanow, W., **Liu, C.S.**, Park, C., Kennedy, G., Long, T., Chun, J. Somatic *APP* gene recombination in Alzheimer's disease and normal neurons. *Nature*. **2018**, *563*, 639-645.

Rohrback, S., Siddoway, B., **Liu, C.S.**, Chun, J. Genomic mosaicism in the developing and adult brain. *Dev. Neurobiol*. **2018**, *78*(11), 1026-1048.

Xiao, T., Ackerman, C.A., Carroll, E.C., Jia, S., Hoagland, A., Chan, J., Thai, B., **Liu, C.S.**, Isacoff, E.Y., Chang, C.J. Copper regulates rest-activity cycles through the locus coeruleus-norepinephrine system. *Nat. Chem. Biol*. **2018**, *14*, 655-663.

ABSTRACT OF THE DISSERTATION

Bioinformatic characterization of genomic and transcriptomic diversity in the human brain

by

Christine Sara Liu

Doctor of Philosophy in Biomedical Sciences

University of California San Diego, 2022

Professor Jerold Chun, Chair
Professor William Joiner, Co-Chair

The human brain can be organized using various different layers of information about the cells: epigenetic, genomic, transcriptomic, proteomic, etc. Recent endeavors have put tremendous effort into mapping the brain cell-by-cell using these layers of information. A challenge associated with these multi-modal approaches is being able to parse through the giga- to terabyte scale amount of data that is generated. My thesis work has focused on investigating the diversity of the brain's genome (DNA) and transcriptome (RNA) and developing bioinformatic tools to make that possible. My work can be broken into two general categories, addressing the genome and the transcriptome.

On the genomic side, I focused on identifying novel features known as gencDNAs (genomic cDNAs). gencDNAs are hypothesized to result from transcription of a highly expressed gene which is then spliced, reverse-transcribed, and inserted back into the genome at the site of a DNA strand break. These novel sequences are predicted to be functional, resulting in additional translation of a protein. *APP*, the amyloid precursor protein gene, was the first gene to be identified as a gencDNA and was determined to be more prevalent in neurons of Alzheimer's disease (AD) patient brains. I developed an unbiased approach to identify additional gencDNAs in the genome from short-read sequencing data.

The transcriptome can be studied at various resolutions. Through several projects, I examined gene expression at the single-cell level, and I additionally characterized full-length isoforms using long-read sequencing technologies. Recent advances in sequencing have made it possible to sequence the entire lengths of mRNA transcripts. This technology is relatively new, and bioinformatic tools need to be developed to handle this type of data. While several packages and tools exist for quality control, alignment, reduction of redundancy, and annotation, a tool for comparing isoforms (known and novel) across multiple samples and groups is not available. I made a database-driven tool for this purpose that is compatible with current analysis pipelines. The applications of this software were demonstrated by examining a dataset from the 1000 Genomes Project in addition to a large single-cell dataset investigating gene and isoform expression changes in several neurodegenerative diseases.

CHAPTER 1

INTRODUCTION

The brain contains multitudes - there is immense diversity on many levels from cell morphology to gene expression. Recent endeavors have aimed to characterize and organize the cell types of the brain using epigenetics, genomics, transcriptomics, proteomics, etc. Advances in sequencing technology have made it easier to assess variations in the genome and transcriptome that contribute to the overall diversity in cells of the brain.

On the genomic level, variations can exist on the population level as well as within an individual. Genomic mosaicism, a phenomenon in which cells do not share an identical genome, has been well characterized in the human brain (*1-3*). It can be observed and measured in many forms through various experimental approaches. Early, low-resolution studies examined DNA content variation (DCV) and aneuploidies using flow cytometry and imaging respectively. Initial studies comparing neuronal and non-neuronal cells demonstrated that neurons typically had increased DNA content in comparison to lymphocytes (*4-6*). Aneuploidies that were detected showed both increased and decreased numbers of chromosomes compared to the expected 23 (*7-10*). Smaller features like copy number variations (CNVs) and single nucleotide variations (SNVs) can also be detected (*2, 11-20*). It is predicted that rearrangements in the genome that cause the individual variation from cell to cell can occur during development but also later in life. Mutations in progenitor populations can be propagated to daughter cells, however in neurons, which are postmitotic, these alterations will only occur in that single cell.

More recently, a novel form of genomic mosaicism in the brain has been identified and termed genomic cDNA (gencDNA)(*21*). gencDNAs can be categorized as a specific type of structural variation (SV) where the genomic sequence resembles a spliced mRNA. While gencDNA sequences often resemble highly expressed mRNA isoforms, they can also contain novel combinations of exons or splice sites. The initial study defining gencDNAs also observed sequences with intra-exonic junctions (IEJs)(*21*). These new splice junctions join together the middle of one exon to the middle of another (not necessarily adjacent) exon with all intervening sequence spliced out. These new splice sites and junctions are hypothesized to be the result of a low-fidelity reverse transcriptase involved in the formation of the gencDNA. The mechanism

for the generation of gencDNAs has been determined to involve gene expression, reverse transcriptase activity, and DNA strand breaks. First, a gene is expressed, and the mRNA is processed, splicing out the introns. Next, the gene is reverse transcribed back into DNA. Finally, that sequence is inserted back into the genome at the site of a DNA strand break.

The first gencDNA to be discovered was from the gene *APP*, the amyloid precursor protein gene (*21*). This gene is strongly associated with Alzheimer's disease (AD), and its gencDNAs were found to be prevalent in frontal cortex tissue from individuals diagnosed with sporadic AD. Many different forms of *APP* sequences were identified through several types of experiments including genomic DNA PCR, targeted short-read sequencing, and PacBio long-read sequencing. The targeted short-read sequencing experiment was later determined to contain contamination from an *APP* plasmid (*22*). The plasmid sequence was discovered by identifying the pGEM-T Easy backbone sequence flanking the terminal exons of *APP*. While impossible to prove, the argument was made that all the exon junction-spanning reads originated not from gencDNA sequence but the plasmid contamination. The presence of plasmid contamination was acknowledged, and "clean" datasets that had reads indicating the presence of gencDNAs were cited as evidence that the plasmid contamination could not account for all the exon junction-spanning reads (Chapter 2)(*23*).

Additional evidence supporting the existence of gencDNAs was provided in the form of insertion sites (*23*). gencDNAs are inserted back into the genome, and should therefore be flanked by sequences that differ from the native gene locus. To look for evidence of insertion sites, short-read sequencing data was analyzed to identify specific read structures. These paired read structures were expected to contain a single read that spanned the UTR of the gene (*APP* in this case) and the novel insertion site, while the mate was expected to map to the novel insertion site as well. These types of read pairs do not provide direct evidence of the cDNA-like sequence, however read mates of exon-junction spanning reads were additionally examined to see if they mapped to potential insertion sites as well. Long-read sequencing approaches through PacBio or Oxford Nanopore Technologies would provide a fuller picture of what the insertion sites of

gencDNAs look like if they could be captured. A read from either approach on average would be long enough to capture the entire cDNA-like sequence in addition to flanking sequences upstream and downstream of it.

One challenge of sequencing these structures is their rarity. Estimates of frequency in AD brain regions in the original gencDNA report indicated that they were present in ∼60% of neurons (*21*). This is predicted to be frequent enough to be captured in sequencing approaches (targeted and whole genome) but depends on the sample itself. Neurons are largely outnumbered by glial cells in the brain, which could affect the ability to sequence gencDNAs that are predominantly expressed in neurons (*24*). However, the prediction that other genes could generate gencDNA sequences does not preclude the possibility that certain gencDNAs are primarily created in non-neuronal cells.

Chapter 3 describes a bioinformatic pipeline used to analyze over 3,000 sequencing datasets with the intent of identifying additional gencDNAs. The premise of this project was based on the estimate from the original gencDNA report that gencDNAs occur in approximately 60% of neurons from AD. We predicted that gencDNAs could account for some of the DCV increase observed in neurons and that genes associated with particular diseases could potentially have deleterious effects through gencDNA formation. The results of this study did not indicate any enrichment of potential gencDNA-forming genes in brain regions, cell types (neuronal vs non-neuronal), or disease. Moreover, *APP* gencDNAs were not detected in any AD samples, neuronal or non-neuronal. These results were inconsistent with the publication identifying and defining gencDNAs in AD, and we expected many samples to reads indicating the presence of *APP* gencDNAs (*21*). The primary observation from this study was that evidence of gencDNA reads in the form of exon-exon junction spanning reads is extremely rare and occur with a frequency of approximately 1/500,000,000 reads. The non-standardized datasets that were all previously generated for other studies with different purposes could have potentially limited our ability to identify gencDNAs, and this study highlights the need for an approach that can more reliably capture rare sequences.

Another layer of heterogeneity in the brain is its transcriptome. While variants that occur in the genome can then be propagated into RNA sequence, other factors can create diversity in the transcriptome. Mechanisms like alternative splicing and gene expression regulation create transcriptional heterogeneity across different cells in the brain.

Gene expression on a global level can be measured through RNA-sequencing. The relative abundances of genes are estimated through the number of sequencing reads that map to them. Relatively modern techniques like single-cell sequencing allow scientists to profile gene expression in individual cells. Instead of averaging gene expression across all the cells in a tissue sample, single-cell sequencing typically uses molecular barcoding to indicate which genes were expressed together in an individual cell (*25, 26*). The resulting transcriptional profiles are used to label each cell with its predicted cell type. With this approach, comparisons can be made across various cell types in different conditions, and changes that only occur in a small population of cells can actually be detected. Single-cell studies have provided detail about cellular gene expression in health and identified cells that appear to be more affected or involved in disease.

Alternative splicing is a process that creates variably spliced mRNAs that can result in many different gene products from a single gene. This mechanism can create a huge amount of isoform diversity. Typical RNA-sequencing studies use short-read sequencing technologies which are not able to capture splicing differences without ambiguity. Short reads can only span one or two splice junctions, making it possible to infer that a few exons were present in the same transcript but not providing enough information about additional exons not covered by the read. In contrast, long-read sequencing can resolve the entire isoform making it possible to read through all the different splice sites and exon combinations (*27-29*).

Several studies have made use of long-read sequencing approaches for examining RNA diversity. Scientists have discovered numerous novel transcripts when examining the human brain in health and disease and also in cancers (*30-35*). The consistency with which these studies identify isoforms that are not part of the reference annotation highlights the great advance in sequencing technology to be able to capture these rare structures.

More recently, efforts have been made to combine the use of single-cell technologies with long-read sequencing to characterize cell-type-specific transcript expression. The Tilgner group combined 10X Genomics sequencing with PacBio Iso-Seq using mouse cerebellar cells to establish the protocol (*30*). To our knowledge, our group published the first application of this protocol to human brain (Chapter 4), sequencing 16 samples from healthy controls and individuals with Down syndrome (*36*). The main takeaways from this approach were: 1) $\sim$50% of the sequenced isoforms were novel, indicating that the human brain contains a high level of transcriptomic diversity; 2) the proportion of isoforms that were novel varied amongst different cell types; 3) isoform switching, or a change in the dominantly expressed isoform, occurred between different cell types, but not between control and DS groups as a whole; and 4) cell type identification can be performed using only long reads (without the support of corresponding short-read data), but is limited by depth. While we were still able to make some interesting observations regarding isoform expression amongst cell types in the brain, sequencing depth was a major limitation in identifying trends and statistically significant changes in expression. Typical short-read sequencing methods result in hundreds of millions of reads and can reach gene-level sequencing saturation. The significantly lower throughput of long-read sequencing methods only yields a few hundred thousand high quality reads after processing and filtering for quality control. It is not certain how much sequencing is required to reach isoform-level saturation. The seemingly infinite number of isoforms that can be created from alternative splicing makes it hard to predict a read threshold.

Technology advances quickly, and in the weeks leading up to the finalization of this dissertation, PacBio released a new RNA sequencing kit, MAS-seq, specifically designed to improve the throughput of single-cell isoform sequencing $\sim$16X (*37, 38*). The polymerase reads of PacBio's sequencing instrument far exceed the length of a typical spliced mammalian gene. MAS-seq concatenates several barcoded mRNAs together, and downstream bioinformatics tools demultiplex and separate out the individual reads. This method is likely to be the most effective for use in tissues with only a few cell types because the increased depth still does not attain

comparable throughput to that of short-read sequencing for cell type identification.

With advances in data generation methods, data analysis tool development must follow suit. Bioinformatic tools have been written for many steps of the process from processing to alignment to annotation. Downstream analyses tools have remained fairly limited, with many scientists encountering the dilemma of having no clear guidance or method for comparing multiple samples. The problem originates in the typical analysis workflow. Individual transcriptomes made up of known and novel isoforms are assembled for each sample by using various bioinformatics tools. Each transcriptome's isoform identifiers are unique to the particular sample, and "isoform 1" in sample A will not necessarily match "isoform 1" in sample B. Several "hacks" exist and make up the only current suggestions for how to unify isoform IDs across samples. Each approach has limitations, and while several publications, including ours, has made use of one or another, the field is in need of a tool specifically created to address this need.

Chapter 6 describes isoSeQL, a program I wrote to compare isoforms from different datasets. The program creates a SQLite database for storing information about each isoform identified in each sample. Supporting functions can be used to query the database to identify isoforms that overlap, examine isoform switching, and calculate differential expression using other bioinformatics packages. This tool has developed further through its application to other projects in the lab and will hopefully continue to develop as other scientists use this for their own studies.

References

1. Rohrback, S., Siddoway, B., Liu, C. S. and Chun, J., Genomic mosaicism in the developing and adult brain. *Dev Neurobiol* **78**, 1026-1048 (2018).

2. McConnell, M. J., Moran, J. V., Abyzov, A., Akbarian, S., Bae, T., Cortes-Ciriano, I., Erwin, J. A., Fasching, L., Flasch, D. A., Freed, D., Ganz, J., Jaffe, A. E., Kwan, K. Y., Kwon, M., Lodato, M. A., Mills, R. E., Paquola, A. C. M., Rodin, R. E., Rosenbluh, C., Sestan, N., Sherman, M. A., Shin, J. H., Song, S., Straub, R. E., Thorpe, J., Weinberger, D. R., Urban, A. E., Zhou, B., Gage, F. H., Lehner, T., Senthil, G., Walsh, C. A., Chess, A., Courchesne, E., Gleeson, J. G., Kidd, J. M., Park, P. J., Pevsner, J., Vaccarino, F. M. and Brain Somatic Mosaicism, Network, Intersection of diverse neuronal genomes and neuropsychiatric disease: The Brain Somatic Mosaicism Network. *Science* **356**, (2017).

3. Costantino, I., Nicodemus, J. and Chun, J., Genomic Mosaicism Formed by Somatic Variation in the Aging and Diseased Brain. *Genes (Basel)* **12**, (2021).

4. Bushman, D. M., Kaeser, G. E., Siddoway, B., Westra, J. W., Rivera, R. R., Rehen, S. K., Yung, Y. C. and Chun, J., Genomic mosaicism with increased amyloid precursor protein (APP) gene copy number in single neurons from sporadic Alzheimer's disease brains. *Elife* **4**, (2015).

5. Westra, J. W., Rivera, R. R., Bushman, D. M., Yung, Y. C., Peterson, S. E., Barral, S. and Chun, J., Neuronal DNA content variation (DCV) with regional and individual differences in the human brain. *J Comp Neurol* **518**, 3981-4000 (2010).

6. Fischer, H. G., Morawski, M., Bruckner, M. K., Mittag, A., Tarnok, A. and Arendt, T., Changes in neuronal DNA content variation in the human brain during aging. *Aging Cell* **11**, 628-633 (2012).

7. Peterson, S. E., Westra, J. W., Rehen, S. K., Young, H., Bushman, D. M., Paczkowski, C. M., Yung, Y. C., Lynch, C. L., Tran, H. T., Nickey, K. S., Wang, Y. C., Laurent, L. C., Loring, J. F., Carpenter, M. K. and Chun, J., Normal human pluripotent stem cell lines exhibit pervasive mosaic aneuploidy. *PLoS One* **6**, e23018 (2011).

8. Rehen, S. K., Yung, Y. C., McCreight, M. P., Kaushal, D., Yang, A. H., Almeida, B. S., Kingsbury, M. A., Cabral, K. M., McConnell, M. J., Anliker, B., Fontanoz, M. and Chun, J., Constitutional aneuploidy in the normal human brain. *J Neurosci* **25**, 2176-2180 (2005).

9. Yang, A. H., Kaushal, D., Rehen, S. K., Kriedt, K., Kingsbury, M. A., McConnell, M. J. and Chun, J., Chromosome segregation defects contribute to aneuploidy in normal neural progenitor cells. *J Neurosci* **23**, 10454-10462 (2003).

10. Knouse, K. A., Wu, J., Whittaker, C. A. and Amon, A., Single cell sequencing reveals low levels of aneuploidy across mammalian tissues. *Proc Natl Acad Sci U S A* **111**, 13409-13414 (2014).

11. Cai, X., Evrony, G. D., Lehmann, H. S., Elhosary, P. C., Mehta, B. K., Poduri, A. and Walsh, C. A., Single-cell, genome-wide sequencing identifies clonal somatic copy-number variation in the human brain. *Cell Rep* **8**, 1280-1289 (2014).

12. McConnell, M. J., Lindberg, M. R., Brennand, K. J., Piper, J. C., Voet, T., Cowing-Zitron, C., Shumilina, S., Lasken, R. S., Vermeesch, J. R., Hall, I. M. and Gage, F. H., Mosaic copy number variation in human neurons. *Science* **342**, 632-637 (2013).

13. Rohrback, S., April, C., Kaper, F., Rivera, R. R., Liu, C. S., Siddoway, B. and Chun, J., Submegabase copy number variations arise during cerebral cortical neurogenesis as revealed by single-cell whole-genome sequencing. *Proc Natl Acad Sci U S A* **115**, 10804-10809 (2018).

14. Bae, T., Fasching, L., Wang, Y., Shin, J. H., Suvakov, M., Jang, Y., Norton, S., Dias, C., Mariani, J., Jourdon, A., Wu, F., Panda, A., Pattni, R., Chahine, Y., Yeh, R., Roberts, R. C., Huttner, A., Kleinman, J. E., Hyde, T. M., Straub, R. E., Walsh, C. A., Brain Somatic Mosaicism Network section, sign, Urban, A. E., Leckman, J. F., Weinberger, D. R., Vaccarino, F. M. and Abyzov, A., Analysis of somatic mutations in 131 human brains reveals aging-associated hypermutability. *Science* **377**, 511-517 (2022).

15. Bizzotto, S., Dou, Y., Ganz, J., Doan, R. N., Kwon, M., Bohrson, C. L., Kim, S. N., Bae, T., Abyzov, A., Network, Nimh Brain Somatic Mosaicism, Park, P. J. and Walsh, C. A., Landmarks of human embryonic development inscribed in somatic mutations. *Science* **371**, 1249-1253 (2021).

16. Wang, Y., Bae, T., Thorpe, J., Sherman, M. A., Jones, A. G., Cho, S., Daily, K., Dou, Y., Ganz, J., Galor, A., Lobon, I., Pattni, R., Rosenbluh, C., Tomasi, S., Tomasini, L., Yang, X., Zhou, B., Akbarian, S., Ball, L. L., Bizzotto, S., Emery, S. B., Doan, R., Fasching, L., Jang, Y., Juan, D., Lizano, E., Luquette, L. J., Moldovan, J. B., Narurkar, R., Oetjens, M. T., Rodin, R. E., Sekar, S., Shin, J. H., Soriano, E., Straub, R. E., Zhou, W., Chess, A., Gleeson, J. G., Marques-Bonet, T., Park, P. J., Peters, M. A., Pevsner, J., Walsh, C. A., Weinberger, D. R., Brain Somatic Mosaicism, Network, Vaccarino, F. M., Moran, J. V., Urban, A. E., Kidd, J. M., Mills, R. E. and Abyzov, A., Comprehensive identification of somatic nucleotide variants in human brain tissue. *Genome Biol* **22**, 92 (2021).

17. Lodato, M. A., Rodin, R. E., Bohrson, C. L., Coulter, M. E., Barton, A. R., Kwon, M., Sherman, M. A., Vitzthum, C. M., Luquette, L. J., Yandava, C. N., Yang, P., Chittenden, T. W., Hatem, N. E., Ryu, S. C., Woodworth, M. B., Park, P. J. and Walsh, C. A., Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* **359**, 555-559 (2018).

18. Lodato, M. A. and Walsh, C. A., Genome aging: somatic mutation in the brain links age-related decline with disease and nominates pathogenic mechanisms. *Hum Mol Genet* **28**, R197-R206 (2019).

19. Lodato, M. A., Woodworth, M. B., Lee, S., Evrony, G. D., Mehta, B. K., Karger, A., Lee, S., Chittenden, T. W., D'Gama, A. M., Cai, X., Luquette, L. J., Lee, E., Park, P. J.

and Walsh, C. A., Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science* **350**, 94-98 (2015).

20. Luquette, L. J., Miller, M. B., Zhou, Z., Bohrson, C. L., Zhao, Y., Jin, H., Gulhan, D., Ganz, J., Bizzotto, S., Kirkham, S., Hochepied, T., Libert, C., Galor, A., Kim, J., Lodato, M. A., Garaycoechea, J. I., Gawad, C., West, J., Walsh, C. A. and Park, P. J., Single-cell genome sequencing of human neurons identifies somatic point mutation and indel enrichment in regulatory elements. *Nat Genet* **54**, 1564-1571 (2022).

21. Lee, M. H., Siddoway, B., Kaeser, G. E., Segota, I., Rivera, R., Romanow, W. J., Liu, C. S., Park, C., Kennedy, G., Long, T. and Chun, J., Somatic APP gene recombination in Alzheimer's disease and normal neurons. *Nature* **563**, 639-645 (2018).

22. Kim, J., Zhao, B., Huang, A. Y., Miller, M. B., Lodato, M. A., Walsh, C. A. and Lee, E. A., APP gene copy number changes reflect exogenous contamination. *Nature* **584**, E20-E28 (2020).

23. Lee, M. H., Liu, C. S., Zhu, Y., Kaeser, G. E., Rivera, R., Romanow, W. J., Kihara, Y. and Chun, J.., Reply to: APP gene copy number changes reflect exogenous contamination. *Nature* **584**, E29-E33 (2020).

24. Purves, Dale and Williams, S. Mark, *Neuroscience*. (Sinauer Associates, Sunderland, Mass., ed. 2nd, 2001), pp. xviii, 681, 616, 683, 625 p.

25. Stuart, T. and Satija, R., Integrative single-cell analysis. *Nat Rev Genet* **20**, 257-272 (2019).

26. Jovic, D., Liang, X., Zeng, H., Lin, L., Xu, F. and Luo, Y., Single-cell RNA sequencing technologies and applications: A brief overview. *Clin Transl Med* **12**, e694 (2022).

27. De Paoli-Iseppi, R., Gleeson, J. and Clark, M. B., Isoform Age - Splice Isoform Profiling Using Long-Read Technologies. *Front Mol Biosci* **8**, 711733 (2021).

28. Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E. and Gouil, Q., Opportunities and challenges in long-read sequencing data analysis. *Genome Biol* **21**, 30 (2020).

29. Hardwick, S. A., Joglekar, A., Flicek, P., Frankish, A. and Tilgner, H. U., Getting the Entire Message: Progress in Isoform Sequencing. *Front Genet* **10**, 709 (2019).

30. Gupta, I., Collier, P. G., Haase, B., Mahfouz, A., Joglekar, A., Floyd, T., Koopmans, F., Barres, B., Smit, A. B., Sloan, S. A., Luo, W., Fedrigo, O., Ross, M. E. and Tilgner, H. U., Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat Biotechnol* **36**, 1197-1202 (2018).

31. Hardwick, S. A., Hu, W., Joglekar, A., Fan, L., Collier, P. G., Foord, C., Balacco, J., Lanjewar, S., Sampson, M. M., Koopmans, F., Prjibelski, A. D., Mikheenko, A., Belchikov, N., Jarroux, J., Lucas, A. B., Palkovits, M., Luo, W., Milner, T. A., Ndhlovu, L. C., Smit,

A. B., Trojanowski, J. Q., Lee, V. M. Y., Fedrigo, O., Sloan, S. A., Tombacz, D., Ross, M. E., Jarvis, E., Boldogkoi, Z., Gan, L. and Tilgner, H. U., Single-nuclei isoform RNA sequencing unlocks barcoded exon connectivity in frozen brain tissue. *Nat Biotechnol* **40**, 1082-1092 (2022).

32. Joglekar, A., Prjibelski, A., Mahfouz, A., Collier, P., Lin, S., Schlusche, A. K., Marrocco, J., Williams, S. R., Haase, B., Hayes, A., Chew, J. G., Weisenfeld, N. I., Wong, M. Y., Stein, A. N., Hardwick, S. A., Hunt, T., Wang, Q., Dieterich, C., Bent, Z., Fedrigo, O., Sloan, S. A., Risso, D., Jarvis, E. D., Flicek, P., Luo, W., Pitt, G. S., Frankish, A., Smit, A. B., Ross, M. E. and Tilgner, H. U., A spatially resolved brain region- and cell type-specific isoform atlas of the postnatal mouse brain. *Nat Commun* **12**, 463 (2021).

33. Leung, S. K., Jeffries, A. R., Castanho, I., Jordan, B. T., Moore, K., Davies, J. P., Dempster, E. L., Bray, N. J., O'Neill, P., Tseng, E., Ahmed, Z., Collier, D. A., Jeffery, E. D., Prabhakar, S., Schalkwyk, L., Jops, C., Gandal, M. J., Sheynkman, G. M., Hannon, E. and Mill, J., Full-length transcript sequencing of human and mouse cerebral cortex identifies widespread isoform diversity and alternative splicing. *Cell Rep* **37**, 110022 (2021).

34. Veiga, D. F. T., Nesta, A., Zhao, Y., Deslattes Mays, A., Huynh, R., Rossi, R., Wu, T. C., Palucka, K., Anczukow, O., Beck, C. R. and Banchereau, J., A comprehensive long-read isoform analysis platform and sequencing resource for breast cancer. *Sci Adv* **8**, eabg6711 (2022).

35. Sharon, D., Tilgner, H., Grubert, F. and Snyder, M., A single-molecule long-read survey of the human transcriptome. *Nat Biotechnol* **31**, 1009-1014 (2013).

36. Palmer, C. R., Liu, C. S., Romanow, W. J., Lee, M. H. and Chun, J., Altered cell and RNA isoform diversity in aging Down syndrome brains. *Proc Natl Acad Sci U S A* **118**, (2021).

37. PacBio (PacBio, 2022), https://www.pacb.com/products-and-services/applications/rna-sequencing/single-cell-rna-sequencing/.

38. Al'Khafaji, Aziz M., Smith, Jonathan T., Garimella, Kiran V, Babadi, Mehrtash, Sade-Feldman, Moshe, Gatzen, Michael, Sarkizova, Siranush, Schwartz, Marc A., Popic, Victoria, Blaum, Emily M., Day, Allyson, Costello, Maura, Bowers, Tera, Gabriel, Stacey, Banks, Eric, Philippakis, Anthony A., Boland, Genevieve M., Blainey, Paul C. and Hacohen, Nir, High-throughput RNA isoform sequencing using programmable cDNA concatenation. *bioRxiv*, 2021.2010.2001.462818 (2021).

CHAPTER 2

REPLY: *APP* GENE COPY NUMBER CHANGES REFLECT EXOGENOUS CONTAMINATION

Ming-Hsiang Lee[1,3], Christine S. Liu[1,2,3], Yunjiao Zhu[1], Gwendolyn E. Kaeser[1], Richard Rivera[1], William J. Romanow[1], Yasuyuki Kihara[1] & Jerold Chun[1†]

[1]Sanford Burnham Prebys Medical Discovery Institute, La Jolla, CA, USA.

[2]Biomedical Sciences Program, School of Medicine, University of California San Diego, La Jolla, CA, USA.

[3]These authors contributed equally: Ming-Hsiang Lee, Christine S. Liu.

[†]e-mail: jchun@sbpdiscovery.org

replying to J. Kim et al. Nature http://doi.org/10.1038/s41586-020-2522-3 (2020)

Published in *Nature*

August 2020

In the accompanying comment[1], Kim et al. conclude that somatic gene recombination (SGR) and amyloid precursor protein (*APP*) genomic complementary DNAs (gencDNAs) in the brain are contamination artefacts and do not naturally exist. We disagree. Here we address the three types of analyses used by Kim et al. to reach their conclusions: informatic contaminant identification, plasmid PCR, and single-cell sequencing. Additionally, Kim et al. requested "reads supporting novel *APP* insertion breakpoints," and we now provide ten different examples that support *APP* gencDNA insertion within eight chromosomes beyond wild-type *APP* on chromosome 21 from patients with Alzheimer's disease. If SGR exists, as experimentally supported here and previously[2,3], contamination scenarios become moot.

Our informatic analyses of data generated by an independent laboratory (Park et al.)[4] complement, and are entirely consistent with, what Lee et al.[2] presented via nine distinct lines of evidence, in addition to three from a prior publication[3]. Plasmid contamination was identified in a single pull-down dataset after publication of Lee et al.[2]; however, subsequent analyses did not alter any of our conclusions, including those of our prior publications[3,5], and plasmid contamination-free replication of this approach by ourselves and others supported the original conclusions. Novel retro-insertion sites, alterations of *APP* gencDNA number and form within cell types from the same brain, and pathogenic SNVs that occur only in samples from patients with AD, all support the existence of *APP* gencDNAs produced by SGR.

One predicted outcome of SGR is the generation of novel retro-insertion sites distinct from the wild-type locus, as we demonstrated using DNA in situ hybridization (DISH; Fig. 2n in Lee et al.). Analyses of independently published data sets[4] produced by whole-exome pull-down of DNA from laser-captured human hippocampus or blood revealed ten different *APP* insertion sites within eight different chromosomes (Fig. 1, Supplementary Table 1). We identified clipped reads spanning *APP* untranslated regions (UTRs) and new genomic insertion sites on chromosomes 1, 3, 9, 10, and 12 (Fig. 1a; wild-type *APP* is located on chromosome 21). The corresponding paired-end reads mapped to the same inserted chromosome. We also identified reads spanning *APP* exon–exon junctions of gencDNAs that had mate-reads mapping to other

genomic sites on chromosomes 1, 3, 5, 6, and 13 (Fig. 1b). We are unaware of contamination sources that could produce these results that are entirely consistent with our DISH data showing *APP* gencDNA locations distinct from wild-type *APP*. These new *APP* gencDNA insertion sites strongly support the natural occurrence of *APP* gencDNAs.

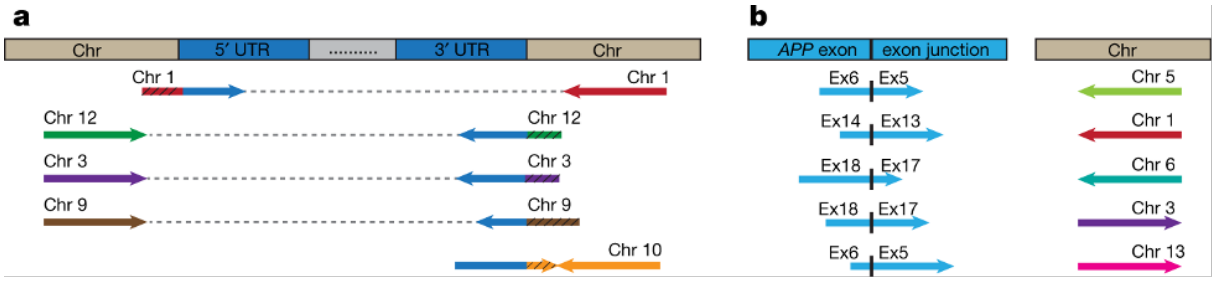

Figure 2.1. Identification of novel *APP* insertion sites in the human genome.
a, Clipped reads spanning *APP* UTRs and novel chromosomal insertion sites were identified. The paired mate-reads of the clipped reads (black hatching) uniquely mapped to the same chromosomes. **b**, Discordant read-pairs were identified where one read spanned an *APP* exon–exon junction and the corresponding mate-read mapped to a novel chromosome. Each chromosome has a unique colour. Arrowhead direction represents the read orientation after mapping to the human reference genome. Arrows oriented in the same direction support sequence inversions. See detailed sequence and alignment information in Supplementary Table 1.

An *APP* plasmid contaminant (pGEM-T Easy *APP*) was found in our single pull-down dataset; however, we could not definitively determine which *APP* exon–exon reads resulted from gencDNAs as opposed to plasmid contamination, especially in view of the 11 other distinct and uncontaminated approaches that had independently supported and/or identified *APP* gencDNAs. Three other pull-down datasets from our laboratory were informatically analysed and found to contain *APP* gencDNA reads while being free from *APP* plasmid contamination by both VecScreen[6] and subsequent use of the Vecuum script[7] (Fig. 2a, b). Possible external source contamination noted by Kim et al. in two of three data sets could not definitively account for all *APP* exon–exon junctions.

The recent availability of independently generated datasets derived from patients with AD[4] provided a test for the independent reproducibility of *APP* gencDNA identification. Five brain and two blood samples from individuals with sporadic AD (SAD) contained *APP* gencDNA

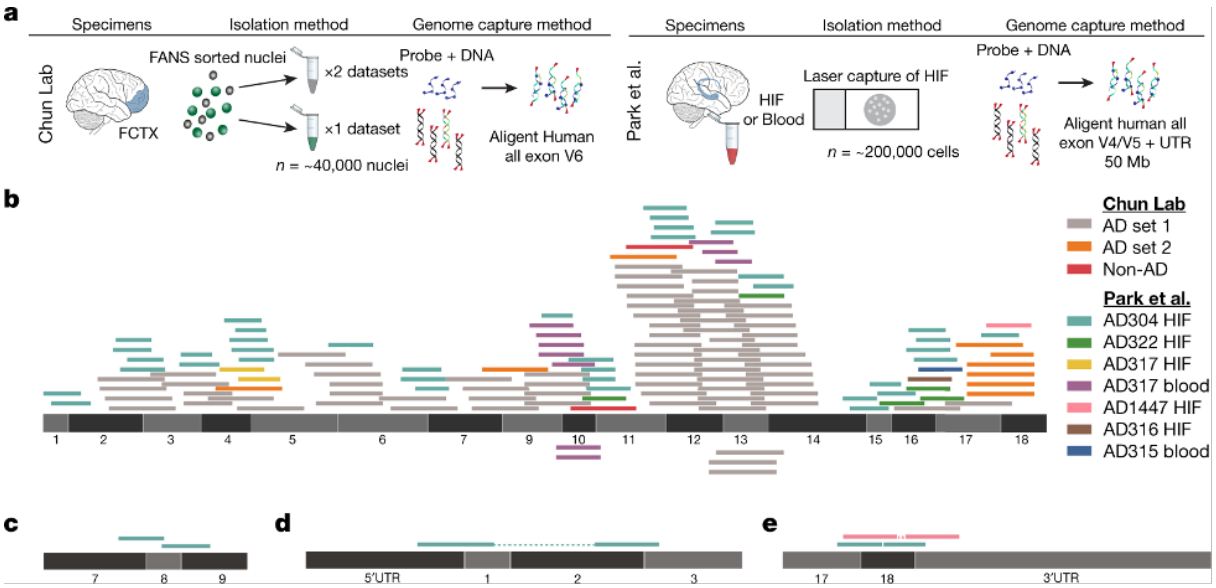

Figure 2.2. Identification of *APP* gencDNA sequences in ten new whole-exome pull-down datasets from two independent laboratories.
a, Method schematic depicting the standard protocol for whole-exome pull-downs and highlighted methodological differences between the independent laboratories (our lab and Park et al.[4]). **b**, *APP-751* sequence with non-duplicate gencDNA reads from the ten new datasets; colour key indicates the source reads for all panels. **c**, Reads that map to junctions between *APP* exons 7, 8, and 9 that are absent from *APP-751*. **d,e**, Paired reads that represent a DNA fragment containing both an exon–exon junction and an *APP* 3' or 5' UTR.

sequences and were shown to be plasmid-free by Vecuum[7] screening (Fig. 2a–e). In addition to exon–exon junction reads and novel insertion sites, we also identified *APP* UTR sequences paired with reads containing *APP* gencDNA exon–exon junctions (Fig. 2d, e). This may be explained by a key experimental design factor: the pull-down probes used by Park et al. contain sequences corresponding to the 5'and 3'UTRs of *APP*.

In addition to *APP* plasmid and amplicon contaminants, Kim et al. invoked genome-wide mouse and human mRNA contamination in the Park et al. data set. We cannot address conditions in the Park et al. laboratory but note that it is completely independent of our own. Kim et al. explain this by implicating the generation of DNA from mRNA, which requires reverse transcriptase activity. The Agilent SureSelect pull-down used by Park et al. and in our experiments do not use reverse transcriptase (Fig. 2a and Supplementary Methods), and we

are unaware of any mechanism that would generate DNA from RNA in the absence of reverse transcriptase activity under the conditions used. An alternative explanation is the existence of gencDNAs that affect other genes, as we previously detected in non-*APP* intra-exonic junctions (IEJs) found in commercial cDNA Iso-Seq data sets (Extended Data Fig. 1). Additional validation would be required for new genes, but we note that an average of 450 Mb of extra DNA exists within cortical neurons from individuals with AD[3] that could accommodate new gencDNA sequences. Kim et al. further invoked genome-wide mouse mRNA contamination in the Park et al. data set to account for *APP* gencDNAs, but this explanation conflicts with the available data. Mouse-specific single nucleotide polymorphisms (SNPs) in the Park et al. data set cannot account for all *APP* gencDNA-supporting reads: five of seven *APP* exon–exon junction sequences do not contain putative mouse-specific SNPs at the specific region reported by Kim et al. (Fig. 3; Kim et al. Fig. 2d). Most critically, the novel *APP* gencDNA insertion sites identified here cannot be explained by genome-wide mRNA contamination.

Figure 2.3. Five *APP* gencDNA-supporting reads that span exon-exon junctions and do not contain mouse-specific SNPs.
APP gencDNA reads were identified that span the *APP* exon10–exon11 junction from the Park et al. datasets[4]. The reference sequences of human and mouse exons are indicated and the positions at which the nucleotides differ are highlighted. Five of the seven exon–exon junction-spanning reads do not contain mouse-specific SNPs.

Kim et al. used PCR of *APP* splice variant plasmids, which generated sequences containing IEJs. However, there are multiple discrepancies between this approach and our biological IEJs and gencDNAs. 1) The experimental conditions, beyond our primer sequences, were different: Kim et al. used twice the concentration of primers and more than one million times more template (250 pg *APP* plasmid is 4.6×10^7 copies versus about 40 gencDNA copies in our

PCR of 20 nuclei; based on Lee et al.[2] Fig. 5: DISH 16/17 averaged about 1.8 copies per SAD nucleus). 2) Both gencDNA and IEJ sequences can be detected with as few as 30 cycles of PCR, as we used in single molecule real-time sequencing (SMRT-seq) (Lee et al.[2] Fig. 3) versus 40 cycles used by Kim et al. 3) The agarose gels in Kim et al. are uniformly and unambiguously dominated by a vastly over-amplified about 2-kb band (Kim et al. Fig. 1c and Extended Data Fig. 3a) that is never seen in human neurons despite our routine identification of myriad smaller bands (compare with Lee et al.[2] Fig. 2b). We did observe an over-amplified about 2-kb band in our purposeful plasmid transfection experiments, which also used PCR; however, the formation of gencDNA and IEJs was comparatively limited, of sequences distinct from brain and critically, required both reverse transcriptase activity and DNA strand breakage (Lee et al.[2], Fig. 4). 4) Finally, only 45 unique IEJs from individual brains with AD and 20 from the brains of healthy controls were identified (Lee et al.[2] Fig. 3 with some overlap, fewer than 65 total) compared to the 12,426 identified by Kim et al. (an approximately 200-fold increase over biological IEJs; Kim et al. Supplementary Table 1). We wish to note that microhomology regions within *APP* exons are intrinsic to the *APP* DNA sequence and that microhomology-mediated repair mechanisms involve DNA polymerases[8,9]. The PCR results of Kim et al. differ from our biological data but might inadvertently support the endogenous formation of at least some IEJs within DNA rather than requiring RNA.

Despite these differences between the non-biological plasmid PCR data generated by Kim et al. and our data, Kim et al. conclude that IEJs from our original study[2] might have originated from contaminants. To eliminate this possibility, Lee et al.[2] presented four lines of evidence for *APP* gencDNAs containing IEJs that are independent of *APP* PCR: two different commercially produced cDNA SMRT-seq libraries, DISH, and RNA in situ hybridization (RISH). The SMRT-seq libraries revealed IEJs within *APP* (Lee et al.[2] Extended Data Fig. 1e) as well as other genes (Extended Data Fig. 1), which cannot be attributed to plasmid contamination or PCR amplification. The DISH and RISH results support the existence of *APP* gencDNAs and IEJs (see Supplementary Discussion and Lee et al.[2] Fig. 2, Extended Data Figs. 1, 2) by

using custom-designed and validated commercial probe technology (Advanced Cell Diagnostics, ACD), which was independently shown to detect exon–exon junctions[10] and single-nucleotide mutations[11]. Thus, gencDNAs and IEJs can be detected in the absence of targeted PCR. Notably, the contamination proposed by Kim et al. cannot account for the marked change in the number and forms of *APP* gencDNAs that occurs with disease state. The change is also apparent when comparing cell types; signals are vastly more prevalent in neurons than in non-neuronal cells from the same brains of individuals with SAD when the samples are processed at the same time by DISH (Lee et al.[2] Fig. 5). Independent peptide nucleic acid fluorescence in situ hybridization (PNA-FISH) and dual-point-paint experiments from our previous work further support *APP* gencDNAs[3] (Table 1). Critically, SMRT-seq identified 11 single-nucleotide variations that are considered pathogenic in familial AD and that were present only in our samples from individuals with SAD; none of them exist as plasmids in our laboratory.

Table 2.1. Summary of targeted and non-targeted *APP* PCR methods and lines of evidence that support *APP* gencDNAs and IEJs

Method	Targeted *APP* PCR	Support for the existence of IEJs and gencDNAs	Reference
Approaches without targeted *APP* PCR			
RISH on IEJ 3/16	None	IEJ 3/16 RNA signal is present in human SAD brain tissue	Lee et al.[2]
Whole-transcriptome SMRT-seq	None	An independent commercial source identified IEJs in *APP* and other genes	Public dataset[a], Lee et al.[2], this Reply
Targeted RNA SMRT-seq	None	RNA pull-down that identified *APP* IEJs	Public dataset[a], Lee et al.[2]

Table 2.1. Summary of targeted and non-targeted *APP* PCR methods and lines of evidence that support *APP* gencDNAs and IEJs (cont'd)

Method	Targeted *APP* PCR	Support for the existence of IEJs and gencDNAs	Reference
DISH of gencDNAs	None	IEJ 3/16 and exon–exon junction 16/17 showed increases in neurons compared to non-neurons from the same brain from an individual with SAD and to non-diseased neurons; J20 mice containing the *APP* transgene under a PDGF-β-promoter showed increased number and size of signal compared to non-neurons and wild-type mice	Lee et al.[2]
Dual point-paint FISH	None	Identified *APP* CNVs of variable puncta size that were not always associated with Chr21	Bushman et al.[3]
PNA-FISH	None	*APP* exon copy number increases show variable signal size and shape with semiquantitative exonic probes	Bushman et al.[3]
Agilent SureSelect targeted pull-down	None	Identified *APP* gencDNAs in brains from individuals with SAD; contains plasmid sequence contamination	Lee et al.[2], this Reply
Agilent all-exon pull-down	None	All-exon pull-downs, with no plasmid contamination by both Vecscreen and Vecuum, contain *APP* gencDNA sequences and evidence of gencDNA UTRs and novel insertion sites	Park et al.[4], this Reply
Approaches with targeted *APP* PCR			
RT-PCR and Sanger sequencing	Oligo-dT primed and targeted *APP* primers	Novel *APP* RNA variants with IEJs; predominantly in neurons from individuals with SAD	Lee et al.[2]
Genomic DNA PCR and Sanger sequencing	Yes	Identified *APP* gencDNAs with IEJs; predominantly in neurons from individuals with SAD	Lee et al.[2]

Table 2.1. Summary of targeted and non-targeted *APP* PCR methods and lines of evidence that support *APP* gencDNAs and IEJs (cont'd)

Method	Targeted *APP* PCR	Support for the existence of IEJs and gencDNAs	Reference
Genomic DNA PCR and SMRT-seq	Yes	IEJ/gencDNAs were more prevalent in number and form in neurons from individuals with SAD compared to non-diseased neurons; identified 11 pathogenic SNVs that were present only in SAD samples	Lee et al.[2]
APP-751 overexpression in CHO cells	Yes	IEJ and gencDNA formation required DNA strand breakage and reverse transcriptase	Lee et al.[2]
Single-cell qPCR	Yes; individual exon	Intragenic exon 14 single-cell qPCR showed copy number increases in prefrontal cortical neurons over cerebellar neurons from the same brain of an individual with SAD	Bushman et al.[3]

CNV, copy number variation.
[a]The Alzheimer brain Iso-Seq dataset was generated by Pacific Biosciences, Menlo Park, California.
Additional sequencing information and analysis is provided at https://downloads.pacbcloud.com/public/dataset/Alzheimer_IsoSeq_2016/.

Kim et al. compared *APP* gencDNA copy number estimates from pull-down sequencing and DISH. However, a direct comparison is not possible since the two methodologies are fundamentally different. For example, pull-downs use solution hybridization on isolated DNA, whereas DISH uses solid-phase hybridization on fixed and sorted single nuclei. Moreover, the sequences targeted are not the same. Pull-down probes target wild-type sequences for endogenous and gencDNA loci, resulting in pull-down competition. By contrast, DISH probes target only gencDNA sequences to provide greater sensitivity. Competition by wild-type loci reduces the efficiency of capture, which is underscored by 32% to 40% of nuclei that do not contain gencDNAs and would contribute only wild-type sequences (Lee et al., Fig. 5c, f). Moreover, a majority of gencDNA positive nuclei (62% to 73%) showed two or fewer signals (Lee et al., Fig. 5c, f) which reduced the relative representation of gencDNA loci. As IEJs

do not contain the full exon sequence, there is inefficient hybridization and a lack of sequence capture and detection. This limitation is overcome by SMRT-seq (Table 1). Lastly, multiple other protocol variations exist, including tissue preparation, fixation, and hybridization conditions, which explain the hypothesized discrepancies.

Kim et al.'s third type of analysis yielded a negative result via interrogation of their own single-cell whole-genome sequencing (scWGS) data, which cannot disprove the existence of *APP* gencDNAs. An average of nine neurons from the brains of seven individuals with SAD were examined, raising immediate sampling issues required to detect mosaic *APP* gencDNAs. Kim et al. identified "uneven genome amplification"[1,12-14] that resulted in about 20% of their single-cell genomes having less than 10× depth of coverage[14] with potential amplification failure at one (∼9% allelic dropout rate) or both alleles (∼2.3% locus dropout rate)[12,14]. These limitations are compounded by potential amplification biases reflected by whole-genome amplification failure rates that may miss neuronal subtypes and/or disease states, which is especially relevant to single copies of *APP* gencDNAs that are as small as about 0.15 kb (but still detectable by DISH). Kim et al. state that the increased exonic read depth relative to introns reliably detects germline retrogene insertions in single cells from affected individuals (Kim et al., Fig. 3b); however, these data also demonstrate that increased exonic read depth is not observed in all cells—or even a majority in some cases—from the same individuals carrying the germline insertions of *SKA3* (AD3 and AD4) and *ZNF100* (AD2). These results demonstrate inherent technical limitations in the work by Kim et al. that prevent the accurate detection of even germline pseudogenes present in all cells, thus explaining an inability to detect the rarer mosaic gencDNAs produced by SGR. Kim et al.'s informatic analysis is also based on the unproven assumption that the structural features of gencDNA are shared with processed pseudogenes and LINE1 elements (Kim et al. Fig. 3a and Extended Data Fig. 1a), and possible differences could prevent straightforward detection under even ideal conditions as has been documented for LINE1[15]. These issues could explain Kim et al.'s negative results.

Considering these points, we believe that our data and conclusions supporting SGR and

APP gencDNAs remain intact and warrant their continued study in the normal and diseased brain.

References

1. Kim, J., Zhao, B., Huang, A. Y., Miller, M. B., Lodato, M. A., Walsh, C. A. and Lee, E. A. APP gene copy number changes reflect exogenous contamination. *Nature* **584**, E20-E28, doi:10.1038/s41586-020-2522-3 (2020).

2. Lee, M. H., Siddoway, B., Kaeser, G. E., Segota, I., Rivera, R., Romanow, W. J., Liu, C. S., Park, C., Kennedy, G., Long, T. and Chun, J. Somatic APP gene recombination in Alzheimer's disease and normal neurons. *Nature* **563**, 639-645, doi:10.1038/s41586-018-0718-6 (2018).

3. Bushman, D. M., Kaeser, G. E., Siddoway, B., Westra, J. W., Rivera, R. R., Rehen, S. K., Yung, Y. C. and Chun, J. Genomic mosaicism with increased amyloid precursor protein (APP) gene copy number in single neurons from sporadic Alzheimer's disease brains. *Elife* **4**, doi:10.7554/eLife.05116 (2015).

4. Park, J. S., Lee, J., Jung, E. S., Kim, M. H., Kim, I. B., Son, H., Kim, S., Kim, S., Park, Y. M., Mook-Jung, I., Yu, S. J. and Lee, J. H. Brain somatic mutations observed in Alzheimer's disease associated with aging and dysregulation of tau phosphorylation. *Nat Commun* **10**, 3090, doi:10.1038/s41467-019-11000-7 (2019).

5. Rohrback, S., April, C., Kaper, F., Rivera, R. R., Liu, C. S., Siddoway, B. and Chun, J. Submegabase copy number variations arise during cerebral cortical neurogenesis as revealed by single-cell whole-genome sequencing. *Proc Natl Acad Sci USA* **115**, 10804-10809, doi:10.1073/pnas.1812702115 (2018).

6. Cummings, J. L., Morstorf, T. and Zhong, K. Alzheimer's disease drug-development pipeline: few candidates, frequent failures. *Alzheimers Res Ther* **6**, 37, doi:10.1186/alzrt269 (2014).

7. Kim, J., Maeng, J. H., Lim, J. S., Son, H., Lee, J., Lee, J. H. and Kim, S. Vecuum: identification and filtration of false somatic variants caused by recombinant vector contamination. *Bioinformatics* **32**, 3072-3080, doi:10.1093/bioinformatics/btw-383 (2016).

8. van Schendel, R., van Heteren, J., Welten, R. and Tijsterman, M. Genomic Scars Generated by Polymerase Theta Reveal the Versatile Mechanism of Alternative End-Joining. *PLoS Genet* **12**, e1006368, doi:10.1371/journal.pgen.1006368 (2016).

9. Sfeir, A. and Symington, L. S. Microhomology-Mediated End Joining: A Back-up Survival Mechanism or Dedicated Pathway? *Trends Biochem Sci* **40**, 701-714, doi:10.1016/j.tibs.2015.08.-006 (2015).

10. ACD. splice variant case study: EGFRvIII detection in glioblastoma, <https://acdbio.com/science/applications/research-areas/egfrviii>(2019).

11. Baker, A. M., Huang, W., Wang, X. M., Jansen, M., Ma, X. J., Kim, J., Anderson, C. M., Wu, X., Pan, L., Su, N., Luo, Y., Domingo, E., Heide, T., Sottoriva, A., Lewis, A.,

Beggs, A. D., Wright, N. A., Rodriguez-Justo, M., Park, E., Tomlinson, I. and Graham, T. A. Robust RNA-based in situ mutation detection delineates colorectal cancer subclonal evolution. *Nat Commun* **8**, 1998, doi:10.1038/s41467-017-02295-5 (2017).

12. Evrony, G. D., Cai, X., Lee, E., Hills, L. B., Elhosary, P. C., Lehmann, H. S., Parker, J. J., Atabay, K. D., Gilmore, E. C., Poduri, A., Park, P. J. and Walsh, C. A. Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell* **151**, 483-496, doi:10.1016/j.cell.2012.09.035 (2012).

13. Cai, X., Evrony, G. D., Lehmann, H. S., Elhosary, P. C., Mehta, B. K., Poduri, A. and Walsh, C. A. Single-cell, genome-wide sequencing identifies clonal somatic copy-number variation in the human brain. *Cell Rep* **8**, 1280-1289, doi:10.1016/j.celrep.2014.07.043 (2014).

14. Evrony, G. D., Lee, E., Mehta, B. K., Benjamini, Y., Johnson, R. M., Cai, X., Yang, L., Haseley, P., Lehmann, H. S., Park, P. J. and Walsh, C. A. Cell lineage analysis in human brain using endogenous retroelements. *Neuron* **85**, 49-59, doi:10.1016/j.neuron.2014.12.028 (2015).

15. Rohrback, S., Siddoway, B., Liu, C. S. and Chun, J. Genomic mosaicism in the developing and adult brain. *Dev Neurobiol* **78**, 1026-1048, doi:10.1002/dneu.22626 (2018).

Data availability

Data from Park et al. were deposited in the National Center for Biotechnology Information Sequence Read Archive database under accession number PRJNA532465. Data from the newly reported full exome pull-down data sets will be provided for the *APP* locus upon request.

Code availability

The source codes of the customized algorithms are available on GitHub at https://github.com/christine-liu/exonjunction.

Acknowledgements We thank L. Wolszon and D. Jones for manuscript editing. Research reported in this publication was supported by the NIA of the National Institutes of Health under award numbers R56AG067489 and P50AG005131 (J.C.) and NINDS R01NS103940 (Y.K.). This work was supported by non-federal funds from The Shaffer Family Foundation, The Bruce Ford & Anne Smith Bundy Foundation, and Sanford Burnham Prebys Medical Discovery Institute funds (J.C.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions M.-H.L., Y.K., W.J.R. and R.R. conducted laboratory experiments; C.S.L. and Y.Z. analysed sequencing data; and J.C. conceived and oversaw the experiments. G.E.K, C.S.L, and Y.Z. created figures. All authors wrote and edited the manuscript. This Reply was the work of current laboratory members.

Competing interests Sanford Burnham Prebys Medical Discovery Institute has filed the following patent applications on the subject matter of this publication: (1) PCT application number PCT/US2018/030520 entitled, 'Methods of diagnosing and treating Alzheimer's disease' filed 1 May 2018, which claims priority to US provisional application 62/500,270 filed 2 May 2017; and (2) US provisional application number 62/687,428 entitled, 'Anti-retroviral therapies and reverse transcriptase inhibitors for treatment of Alzheimer's disease' filed 20 June 2018. J.C. is a co-founder of Mosaic Pharmaceuticals.

Chapter 2, in full, is a reprint of the material as it appears in *Nature* 2020. Lee, M-H.*, Liu, C.S.*, Zhu, Y., Kaeser, G.E., Rivera, R., Romanow, W.J., Kihara, Y., Chun, J., Reply: Evidence that *APP* gene copy number changes reflect recombinant vector contamination. The dissertation author was a co-primary researcher and author of this paper.

CHAPTER 3

NOVEL BIOINFORMATIC PIPELINE FOR IDENTIFYING GENCDNAS IN SHORT-READ SEQUENCING

Introduction

The diversity of the brain can be examined at many levels - cellular, genomic, transcriptomic, etc. Genomic mosaicism, variation in individual cellular genomes, has been well documented in the human brain. Features such as aneuploidies, copy number variations (CNVs), and single-nucleotide variations (SNVs) can contribute to mosaicism (*1-20*). Initial studies identified DNA content variation (DCV) between neurons and non-neurons in the frontal cortex of the brain. These neurons showed a substantial increase of ∼200Mb DNA in comparison to lymphocytes and neurons from the cerebellum (*3, 6, 19*). More recent work has identified a novel form of genomic mosaicism in the brain: genomic cDNAs (gencDNAs) (*21*). gencDNAs are genomic sequences that resemble spliced mRNAs but are found in the genome. gencDNA sequences often resemble prominently expressed mRNA isoforms, but can also contain novel combinations of exons or junctions. Novel splice junctions called intra-exonic junctions (IEJs) are sometimes observed in gencDNAs, joining together the middle of one exon to the middle of another (not necessarily adjacent) exon with all intervening sequencing spliced out. Three requirements for gencDNA formation were identified: 1) gene expression, 2) reverse transcriptase activity, and 3) DNA strand breaks. gencDNAs are thought to form through reverse transcription of the expressed mRNA and reinsertion into a random location in the genome. The first gencDNA identified originated from the amyloid precursor protein gene, *APP*, and the prevalence of the gencDNA was increased in Alzheimer's disease.

We hypothesize that many other genes besides *APP* could generate gencDNAs and that these gencDNAs may partially account for the increased DNA content observed in neurons. The frequency with which *APP* gencDNAs were observed suggested that next-generation sequencing approaches could be used to identify additional genes that became gencDNAs. Currently, no other published studies have identified additional gencDNAs and their relationship to disease, but many sequencing datasets from brain have been produced to examine other genomic features. We present a comprehensive survey of 3,646 publicly available sequencing datasets from various diseases, brain regions, and sequencing strategies that were analyzed with a novel pipeline to

detect evidence of gencDNAs.

Results

NGS sequencing uses read lengths that are typically 50-250 nucleotides (nt) in length. These reads are long enough to cover the junction between two exons (Fig 1A). These types of reads are commonplace in RNA-seq, but should not be observed often in genomic DNA. Exon-exon spanning reads detected in genomic DNA can serve as potential indicators of the presence of gencDNAs. To identify these reads in genomic DNA sequencing, we started by mapping the reads with STAR, a short-read aligner typically used for RNA-seq, and followed with custom bash and Python scripts to parse through the alignments to identify exon junction-spanning reads (Fig 1B)(*22*). The custom scripts specifically pulled out these reads of interest and validated their mapping with an additional aligner, blat (*23*). STAR was chosen as the initial aligner for its ability to map across large "gaps" where the read does not contain large stretches of the genome, typically corresponding to introns in RNA-seq. Because the structure of a gencDNA resembles processed mRNA, reads originating from them would resemble RNA-seq reads. Other alignment tools would instead split the read and align the two parts separately, making it more difficult to find the individual reads that map across an exon-exon junction (*24, 25*).

The custom scripts rely on STAR's ability to align these "split reads" accurately. In order to determine how much of the read needs to align to a single exon while the rest of the read maps to another exon, we randomly generated one million 150nt reads with varying split lengths and counted which ones were mapped accurately (Fig 1C). Reads with at least 15nt or more mapped to a single exon were mapped correctly $\geq$95% of the time. This 15nt limit was used as a cutoff for identifying split reads. The two "blocks" that made up the split read were then compared to a reference annotation file that contained all the coordinates of annotated exons in the reference genome. These reference files are generated for the reference genome of interest. The comparisons are then used to mark whether the split reads map to precise splice sites or if they indicate that the two exons were joined together at an intra-exonic splice junction. Reads that map to known pseudogenes are removed from further validation steps as known pseudogenes

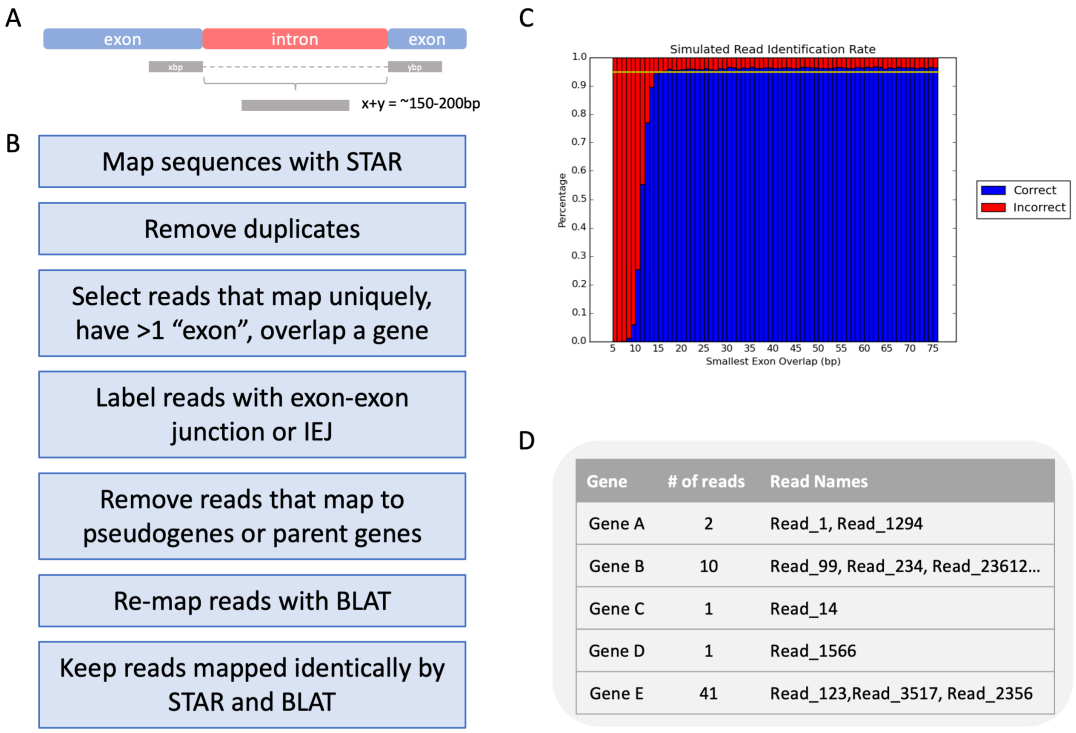

Figure 3.1. Bioinformatic pipeline for identifying gencDNAs.
(**A**) Schematic of exon-exon junction-spanning reads that are identified as evidence for gencDNAs. (**B**) General workflow for processing raw sequencing data to generate a list of potential genes that have become gencDNAs and evidence reads (**C**) Simulated reads with exon-exon junctions used to determine the smallest number of nucleotides necessary to map to an exon accurately. (**D**) Example of information provided in the output file.

may contain exon-exon junctions. As a final confirmation step, these reads are realigned with blat, another aligner that accommodates large gaps in the mapping (*23*). Reads with alignments identical between STAR and blat are output as gencDNA evidence. The output file contains a list of genes, the number of gencDNA-containing reads that were detected for each gene, and the corresponding read names (Fig 1D).

The first gencDNA identified, *APP*, was proposed to play a role in the development of sporadic Alzheimer's disease (*21*). The original publication also suggested that *APP* may not be the only gene to become a gencDNA and contribute to a brain disorder. To investigate whether other genes could be found in gencDNA form in the brain, we examined over 3,000 sequencing datasets. Many sequencing studies have been carried out using brain tissue, but none

have searched for the presence of gencDNAs in the DNA. These datasets are readily available from various data repositories. The only criteria for inclusion in this study were that it originated from brain tissue and that DNA was sequenced. Brain tissue is not easily obtainable, and we wanted to be inclusive of all regions, diseases, sequencing strategies, etc. A total of 3,646 datasets were downloaded: 3,547 from brain tissue, while the rest were composed of other patient-matched tissues. A variety of sequencing library preparation strategies were represented, including whole-genome sequencing (WGS), whole-exome sequencing (WES), targeted gene panels, and single-cell WGS (*4, 7, 10, 12, 26-39*). All of these modalities generate data that can be analyzed using the gencDNA pipeline, however, targeted strategies may be more limited in which genes could be detected as gencDNAs. Each of these datasets was first analyzed with Vecuum, a bioinformatics tool for identifying plasmid contamination in sequencing data, to eliminate the possibility that these exon-exon spanning reads originated from contamination (*40*). "Clean" datasets with no plasmid contamination were then analyzed using the gencDNA pipeline. 27 samples were removed after Vecuum screening.

To account for potential concerns that cDNA library contamination would produce misleading results that were not detected by Vecuum due to the lack of plasmid vector backbone sequence, samples with an outlier number of genes ($>$10) identified as gencDNAs were removed from the sample pool. cDNA library contamination would result in reads spanning exon-exon junctions from a large number of genes. Out of 3619 samples (after removing samples with Vecuum-detected contamination), 359 were removed using this criterion. Notably, all samples from a single study known to have additionally generated matching single-cell RNA-seq libraries were removed (*38*).

2451 samples out of 3260 total (75.18%) that fit our criteria had zero gencDNAs. Using a cutoff of two reads necessary to support each potential gencDNA gene, 42.77% of samples had only one gencDNA gene (Fig 2A). Very few genes were repeatedly detected as potential gencDNAs, and gencDNA reads appeared to occur at a very low frequency, $\sim$1/500,000,000 reads. These results are consistent with the idea that they may occur in an individual cell that

may not be propagated through continued division. gencDNAs that occur in neurons would not be replicated in another cell because neurons are postmitotic.

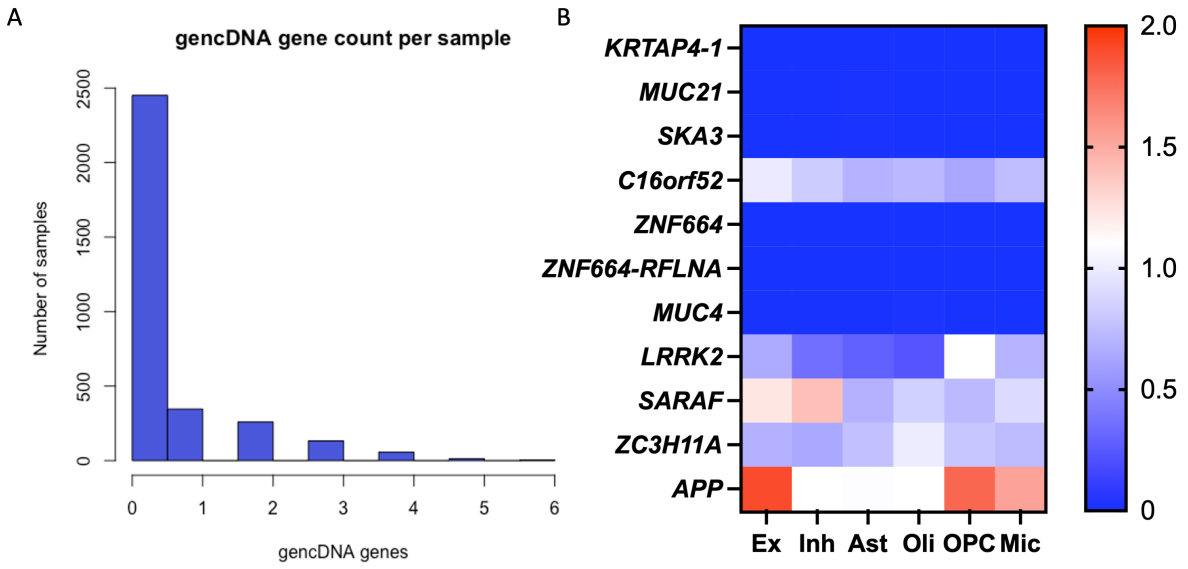

Figure 3.2. Results of gencDNA identification in publicly available datasets.
(**A**) Histogram of the number of samples with a certain number of gencDNAs (**B**) Heat map of expression of top 10 most detected gencDNA genes in different cell types of the human brain (*41*).

Surprisingly, there did not appear to be any enrichment of gencDNAs in any particular disease or brain region. Gene expression also did not serve as a predictor of genes that were more likely to be detected as gencDNAs. Several of the top ten most detected gencDNA genes (detected in the most samples) did not appear to have high expression in various cell types of the brain (Fig 2B, Table 1)(*41*). Contrasting with the original *APP* gencDNA report, no *APP* gencDNA reads were detected in Alzheimer's disease (AD) samples. 1215 AD samples were included in this study, and zero *APP* gencDNA-supporting reads were detected in any of them. Given the estimate that >60% of neurons in AD contained at least one *APP* gencDNA, we would have expected to identify these gencDNAs readily (*21*). It could be argued that many samples included in this study were unsorted, bulk samples where the glial cells could easily outnumber the neurons, however, 945 of the 1215 AD datasets were single cells sorted for NeuN+. If the estimate held true, we'd expect that about 567 of the datasets would have an *APP* gencDNA.

Table 3.1. Top 10 most detected gencDNAs by number of samples

Gene	Sample Count	Total Read Count
KRTAP4-1	665	4764
MUC21	145	482
SKA3	114	11830
C16orf52 (*MOSMO*)	100	477
ZNF664	99	5095
ZNF664-RFLNA	74	1264
MUC4	63	146
LRRK2	25	139
SARAF	23	2076
ZC3H11A	20	43

Discussion

The original report identifying *APP* gencDNAs described sequences that occurred with relatively high frequency in neurons. This suggested that these sequences could easily be identified from next-generation technologies that assess entire genomes in an unbiased manner or even through exon/gene-targeted strategies. We hypothesized that examining short-read sequencing datasets from the brain would uncover several other genes that could form gencDNAs and be linked to other neurological diseases and disorders. We developed a bioinformatic pipeline combining established tools and custom scripts with the goal of identifying exon-exon junction-spanning reads that would be indicative of cDNA-like sequences in the genome. This pipeline was then used to analyze 3,646 publicly available short-read sequencing datasets, a majority of which were from post-mortem human brain tissue.

We had hypothesized that similar to *APP*, the presence of other disease-related genes as gencDNAs would be linked to their respective diseases and affected brain regions. We did not observe an enrichment of gencDNA-generating genes that was linked to disease or brain region. More surprisingly, we also did not detect *APP* gencDNAs in any of our AD samples. Several of these datasets were whole-genome amplifications of single neurons, and if the previously observed frequency had held true, we would have expected a majority of these datasets to have detectable reads spanning *APP* exon-exon junctions.

These negative data could reflect the limitations of this approach to identifying gencDNA sequences. If we assume that gencDNAs are forming mostly in post-mitotic neurons, then these gencDNAs will not be propagated to additional cells through mitotic divisions. Bulk sequencing of the genome does not fully capture the entire genome of each individual cell that is sampled. The odds of capturing a DNA fragment that spans an exon-exon junction of a gencDNA that only occurs in a single cell out of thousands that are sampled are low. Targeting specific genes or only exonic sequences can improve the odds of capturing a gencDNA, but these approaches are still limited by the inability to thoroughly examine the entire genome of each individual cell. This suggests that single-cell whole-genome amplification (WGA) may provide the best estimate of how frequently gencDNAs are formed and which genes become gencDNAs. However, this method is limited too, and several studies have reported "uneven genome amplification" resulting in allelic dropout that may prevent the capture of gencDNA insertions (*28, 29*). This method also relies on the ability to sort out neurons or other cell types of interest which may be subject to its own selection biases.

Another potential explanation for these negative data is that these exon-exon junction reads are an artifact of the sample preparation process or the result of contamination. There are no mechanisms that would explain how these reads would be spliced together exactly at a known mRNA junction, however, further analysis needs to be done to determine the prevalence of similar reads that cover a large "splice junction" but do not join exons together (potentially in an intergenic region). The frequency with which we observed exon-exon junction spanning reads is rare enough that it could be randomly occurring. It's nearly impossible to completely rule out the potential for cDNA contamination. Unpublished calculations from our lab indicate that on average $\sim$10% of reads from RNA-sequencing span exon-exon junctions. A small amount of cDNA contamination would be indistinguishable from gencDNA sequence, and the resulting sequencing reads would exhibit the same hallmark exon-exon junction spanning structure.

One important factor that this approach does not address is the gencDNA's site of integration. 150-200bp is not long enough to span more than an exon-exon junction or two, so

we can only make inferences about the rest of the gencDNA structure and have no information about where it was inserted into the genome. Long-read genomic sequencing is one way to obtain this information. Long sequencing reads will easily cover an entire gencDNA and can potentially include both 5' and 3' flanking sequences. This method can provide both confirmation of the intron-less structure and placement in the genome. The considerably lower throughput and consequently increased limitations in covering the entire genome of an individual cell are challenges of long-read sequencing for identifying gencDNAs.

The difficulty in identifying additional gencDNAs reflects their rarity but also the fact that the current technology does not quite have the resolution to detect anything at such low levels with confidence. This study hints at the possible presence of gencDNA sequences in the genome but was unable to detect any at a meaningful level. As sequencing technologies continue to advance, it may one day be possible to sequence the entire individual genomes of thousands of cells, providing more clarity regarding the frequency with which gencDNAs are integrated into the genome and what role they may play in disease.

References

1. Bae, T., Fasching, L., Wang, Y., Shin, J. H., Suvakov, M., Jang, Y., Norton, S., Dias, C., Mariani, J., Jourdon, A., Wu, F., Panda, A., Pattni, R., Chahine, Y., Yeh, R., Roberts, R. C., Huttner, A., Kleinman, J. E., Hyde, T. M., Straub, R. E., Walsh, C. A., Brain Somatic Mosaicism Network section, sign, Urban, A. E., Leckman, J. F., Weinberger, D. R., Vaccarino, F. M. and Abyzov, A., Analysis of somatic mutations in 131 human brains reveals aging-associated hypermutability. *Science* **377**, 511-517 (2022).

2. Bizzotto, S., Dou, Y., Ganz, J., Doan, R. N., Kwon, M., Bohrson, C. L., Kim, S. N., Bae, T., Abyzov, A., Network, Nimh Brain Somatic Mosaicism, Park, P. J. and Walsh, C. A., Landmarks of human embryonic development inscribed in somatic mutations. *Science* **371**, 1249-1253 (2021).

3. Bushman, D. M., Kaeser, G. E., Siddoway, B., Westra, J. W., Rivera, R. R., Rehen, S. K., Yung, Y. C. and Chun, J., Genomic mosaicism with increased amyloid precursor protein (APP) gene copy number in single neurons from sporadic Alzheimer's disease brains. *Elife* **4**, (2015).

4. Cai, X., Evrony, G. D., Lehmann, H. S., Elhosary, P. C., Mehta, B. K., Poduri, A. and Walsh, C. A., Single-cell, genome-wide sequencing identifies clonal somatic copy-number variation in the human brain. *Cell Rep* **8**, 1280-1289 (2014).

5. Costantino, I., Nicodemus, J. and Chun, J., Genomic Mosaicism Formed by Somatic Variation in the Aging and Diseased Brain. *Genes (Basel)* **12**, (2021).

6. Fischer, H. G., Morawski, M., Bruckner, M. K., Mittag, A., Tarnok, A. and Arendt, T.., Changes in neuronal DNA content variation in the human brain during aging. *Aging Cell* **11**, 628-633 (2012).

7. Knouse, K. A., Wu, J., Whittaker, C. A. and Amon, A., Single cell sequencing reveals low levels of aneuploidy across mammalian tissues. *Proc Natl Acad Sci U S A* **111**, 13409-13414 (2014).

8. Lodato, M. A., Rodin, R. E., Bohrson, C. L., Coulter, M. E., Barton, A. R., Kwon, M., Sherman, M. A., Vitzthum, C. M., Luquette, L. J., Yandava, C. N., Yang, P., Chittenden, T. W., Hatem, N. E., Ryu, S. C., Woodworth, M. B., Park, P. J. and Walsh, C. A., Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* **359**, 555-559 (2018).

9. Lodato, M. A. and Walsh, C. A., Genome aging: somatic mutation in the brain links age-related decline with disease and nominates pathogenic mechanisms. *Hum Mol Genet* **28**, R197-R206 (2019).

10. Lodato, M. A., Woodworth, M. B., Lee, S., Evrony, G. D., Mehta, B. K., Karger, A., Lee, S., Chittenden, T. W., D'Gama, A. M., Cai, X., Luquette, L. J., Lee, E., Park, P. J. and Walsh, C. A., Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science* **350**, 94-98 (2015).

11. Luquette, L. J., Miller, M. B., Zhou, Z., Bohrson, C. L., Zhao, Y., Jin, H., Gulhan, D., Ganz, J., Bizzotto, S., Kirkham, S., Hochepied, T., Libert, C., Galor, A., Kim, J., Lodato, M. A., Garaycoechea, J. I., Gawad, C., West, J., Walsh, C. A. and Park, P. J., Single-cell genome sequencing of human neurons identifies somatic point mutation and indel enrichment in regulatory elements. *Nat Genet* **54**, 1564-1571 (2022).

12. McConnell, M. J., Lindberg, M. R., Brennand, K. J., Piper, J. C., Voet, T., Cowing-Zitron, C., Shumilina, S., Lasken, R. S., Vermeesch, J. R., Hall, I. M. and Gage, F. H., Mosaic copy number variation in human neurons. *Science* **342**, 632-637 (2013).

13. McConnell, M. J., Moran, J. V., Abyzov, A., Akbarian, S., Bae, T., Cortes-Ciriano, I., Erwin, J. A., Fasching, L., Flasch, D. A., Freed, D., Ganz, J., Jaffe, A. E., Kwan, K. Y., Kwon, M., Lodato, M. A., Mills, R. E., Paquola, A. C. M., Rodin, R. E., Rosenbluh, C., Sestan, N., Sherman, M. A., Shin, J. H., Song, S., Straub, R. E., Thorpe, J., Weinberger, D. R., Urban, A. E., Zhou, B., Gage, F. H., Lehner, T., Senthil, G., Walsh, C. A., Chess, A., Courchesne, E., Gleeson, J. G., Kidd, J. M., Park, P. J., Pevsner, J., Vaccarino, F. M. and Brain Somatic Mosaicism, Network, Intersection of diverse neuronal genomes and neuropsychiatric disease: The Brain Somatic Mosaicism Network. *Science* **356**, (2017).

14. Peterson, S. E., Westra, J. W., Rehen, S. K., Young, H., Bushman, D. M., Paczkowski, C. M., Yung, Y. C., Lynch, C. L., Tran, H. T., Nickey, K. S., Wang, Y. C., Laurent, L. C., Loring, J. F., Carpenter, M. K. and Chun, J., Normal human pluripotent stem cell lines exhibit pervasive mosaic aneuploidy. *PLoS One* **6**, e23018 (2011).

15. Rehen, S. K., Yung, Y. C., McCreight, M. P., Kaushal, D., Yang, A. H., Almeida, B. S., Kingsbury, M. A., Cabral, K. M., McConnell, M. J., Anliker, B., Fontanoz, M. and Chun, J., Constitutional aneuploidy in the normal human brain. *J Neurosci* **25**, 2176-2180 (2005).

16. Rohrback, S., April, C., Kaper, F., Rivera, R. R., Liu, C. S., Siddoway, B. and Chun, J., Submegabase copy number variations arise during cerebral cortical neurogenesis as revealed by single-cell whole-genome sequencing. *Proc Natl Acad Sci U S A* **115**, 10804-10809 (2018).

17. Rohrback, S., Siddoway, B., Liu, C. S. and Chun, J., Genomic mosaicism in the developing and adult brain. *Dev Neurobiol* **78**, 1026-1048 (2018).

18. Wang, Y., Bae, T., Thorpe, J., Sherman, M. A., Jones, A. G., Cho, S., Daily, K., Dou, Y., Ganz, J., Galor, A., Lobon, I., Pattni, R., Rosenbluh, C., Tomasi, S., Tomasini, L., Yang, X., Zhou, B., Akbarian, S., Ball, L. L., Bizzotto, S., Emery, S. B., Doan, R., Fasching, L., Jang, Y., Juan, D., Lizano, E., Luquette, L. J., Moldovan, J. B., Narurkar, R., Oetjens, M. T., Rodin, R. E., Sekar, S., Shin, J. H., Soriano, E., Straub, R. E., Zhou, W., Chess, A., Gleeson, J. G., Marques-Bonet, T., Park, P. J., Peters, M. A., Pevsner, J., Walsh, C. A., Weinberger, D. R., Brain Somatic Mosaicism, Network, Vaccarino, F. M., Moran, J. V., Urban, A. E., Kidd, J. M., Mills, R. E. and Abyzov, A., Comprehensive identification of somatic nucleotide variants in human brain tissue. *Genome Biol* **22**, 92 (2021).

19. Westra, J. W., Rivera, R. R., Bushman, D. M., Yung, Y. C., Peterson, S. E., Barral, S. and Chun, J., Neuronal DNA content variation (DCV) with regional and individual differences in the human brain. *J Comp Neurol* **518**, 3981-4000 (2010).

20. Yang, A. H., Kaushal, D., Rehen, S. K., Kriedt, K., Kingsbury, M. A., McConnell, M. J. and Chun, J., Chromosome segregation defects contribute to aneuploidy in normal neural progenitor cells. *J Neurosci* **23**, 10454-10462 (2003).

21. Lee, M. H., Siddoway, B., Kaeser, G. E., Segota, I., Rivera, R., Romanow, W. J., Liu, C. S., Park, C., Kennedy, G., Long, T. and Chun, J., Somatic APP gene recombination in Alzheimer's disease and normal neurons. *Nature* **563**, 639-645 (2018).

22. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T. R., STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).

23. Kent, W. J., BLAT–the BLAST-like alignment tool. *Genome Res* **12**, 656-664 (2002).

24. Langmead, B., Trapnell, C., Pop, M. and Salzberg, S. L., Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009).

25. Li, H. and Durbin, R., Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).

26. Cookson, M. NABEC: Exome Sequencing of North American Brain Expression Consortium (NABEC) Subjects (dbGaP). phs001301.v3.p1

27. Erwin, J. A., Paquola, A. C., Singer, T., Gallina, I., Novotny, M., Quayle, C., Bedrosian, T. A., Alves, F. I., Butcher, C. R., Herdy, J. R., Sarkar, A., Lasken, R. S., Muotri, A. R. and Gage, F. H., L1-associated genomic regions are deleted in somatic cells of the healthy human brain. *Nat Neurosci* **19**, 1583-1591 (2016).

28. Evrony, G. D., Cai, X., Lee, E., Hills, L. B., Elhosary, P. C., Lehmann, H. S., Parker, J. J., Atabay, K. D., Gilmore, E. C., Poduri, A., Park, P. J. and Walsh, C. A., Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell* **151**, 483-496 (2012).

29. Evrony, G. D., Lee, E., Mehta, B. K., Benjamini, Y., Johnson, R. M., Cai, X., Yang, L., Haseley, P., Lehmann, H. S., Park, P. J. and Walsh, C. A., Cell lineage analysis in human brain using endogenous retroelements. *Neuron* **85**, 49-59 (2015).

30. Keogh, M. J., Wei, W., Aryaman, J., Walker, L., van den Ameele, J., Coxhead, J., Wilson, I., Bashton, M., Beck, J., West, J., Chen, R., Haudenschild, C., Bartha, G., Luo, S., Morris, C. M., Jones, N. S., Attems, J. and Chinnery, P. F., High prevalence of focal and multi-focal somatic genetic variants in the human brain. *Nat Commun* **9**, 4257 (2018).

31. Leija-Salazar, M., Pittman, A., Mokretar, K., Morris, H., Schapira, A. H. and Proukakis, C., Investigation of Somatic Mutations in Human Brains Targeting Genes Associated With Parkinson's Disease. *Front Neurol* **11**, 570424 (2020).

32. Li, J., Shi, M., Ma, Z., Zhao, S., Euskirchen, G., Ziskin, J., Urban, A., Hallmayer, J. and Snyder, M., Integrated systems analysis reveals a molecular network underlying autism spectrum disorders. *Mol Syst Biol* **10**, 774 (2014).

33. Park, J. S., Lee, J., Jung, E. S., Kim, M. H., Kim, I. B., Son, H., Kim, S., Kim, S., Park, Y. M., Mook-Jung, I., Yu, S. J. and Lee, J. H., Brain somatic mutations observed in Alzheimer's disease associated with aging and dysregulation of tau phosphorylation. *Nat Commun* **10**, 3090 (2019).

34. Sanchez-Luque, F. J., Kempen, M. H. C., Gerdes, P., Vargas-Landin, D. B., Richardson, S. R., Troskie, R. L., Jesuadian, J. S., Cheetham, S. W., Carreira, P. E., Salvador-Palomeque, C., Garcia-Canadas, M., Munoz-Lopez, M., Sanchez, L., Lundberg, M., Macia, A., Heras, S. R., Brennan, P. M., Lister, R., Garcia-Perez, J. L., Ewing, A. D. and Faulkner, G. J., LINE-1 Evasion of Epigenetic Repression in Humans. *Mol Cell* **75**, 590-604 e512 (2019).

35. Tombacz, D., Maroti, Z., Kalmar, T., Csabai, Z., Balazs, Z., Takahashi, S., Palkovits, M., Snyder, M. and Boldogkoi, Z., High-Coverage Whole-Exome Sequencing Identifies Candidate Genes for Suicide in Victims with Major Depressive Disorder. *Sci Rep* **7**, 7106 (2017).

36. Upton, K. R., Gerhardt, D. J., Jesuadian, J. S., Richardson, S. R., Sanchez-Luque, F. J., Bodea, G. O., Ewing, A. D., Salvador-Palomeque, C., van der Knaap, M. S., Brennan, P. M., Vanderver, A. and Faulkner, G. J., Ubiquitous L1 mosaicism in hippocampal neurons. *Cell* **161**, 228-239 (2015).

37. van den Bos, H., Spierings, D. C., Taudt, A. S., Bakker, B., Porubsky, D., Falconer, E., Novoa, C., Halsema, N., Kazemier, H. G., Hoekstra-Wakker, K., Guryev, V., den Dunnen, W. F., Foijer, F., Tatche, M. C., Boddeke, H. W. and Lansdorp, P. M., Single-cell whole genome sequencing reveals no evidence for common aneuploidy in normal and Alzheimer's disease neurons. *Genome Biol* **17**, 116 (2016).

38. Velmeshev, D., Schirmer, L., Jung, D., Haeussler, M., Perez, Y., Mayer, S., Bhaduri, A., Goyal, N., Rowitch, D. H. and Kriegstein, A. R., Single-cell genomics identifies cell type-specific molecular changes in autism. *Science* **364**, 685-689 (2019).

39. Wang, M., Beckmann, N. D., Roussos, P., Wang, E., Zhou, X., Wang, Q., Ming, C., Neff, R., Ma, W., Fullard, J. F., Hauberg, M. E., Bendl, J., Peters, M. A., Logsdon, B., Wang, P., Mahajan, M., Mangravite, L. M., Dammer, E. B., Duong, D. M., Lah, J. J., Seyfried, N. T., Levey, A. I., Buxbaum, J. D., Ehrlich, M., Gandy, S., Katsel, P., Haroutunian, V., Schadt, E. and Zhang, B., The Mount Sinai cohort of large-scale genomic, transcriptomic and proteomic data in Alzheimer's disease. *Sci Data* **5**, 180185 (2018).

40. Kim, J., Maeng, J. H., Lim, J. S., Son, H., Lee, J., Lee, J. H. and Kim, S., Vecuum: identification and filtration of false somatic variants caused by recombinant vector contamination. *Bioinformatics* **32**, 3072-3080 (2016).

41. Lake, B. B., Chen, S., Sos, B. C., Fan, J., Kaeser, G. E., Yung, Y. C., Duong, T. E., Gao, D., Chun, J., Kharchenko, P. V. and Zhang, K., Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat Biotechnol* **36**, 70-80 (2018).

Chapter 3, in part, is currently being prepared for submission for publication of the material. Liu, C.S., Zhu, Y., Chun, J. The dissertation author was the primary researcher and author of this material.

CHAPTER 4

ALTERED CELL AND RNA ISOFORM DIVERSITY IN AGING DOWN SYNDROME BRAINS

Carter R. Palmer[a,b,1], Christine S. Liu[a,b,1], William J. Romanow[a], Ming-Hsiang Lee[a], Jerold Chun[a,*]

[a]Translational Neuroscience Initiative, Sanford Burnham Prebys Medical Discovery Institute, La Jolla, CA 92037, USA.

[b]Biomedical Sciences Program, School of Medicine, University of California San Diego, La Jolla, CA 92093, USA.

[1]C. Palmer and C. Liu contributed equally to this work.

[*] Corresponding author, jchun@sbpdiscovery.org

Published in *PNAS*

November 2021

Abstract

Down syndrome (DS) – trisomy of human chromosome 21 (HSA21) – is characterized by lifelong cognitive impairments and development of neuropathological hallmarks of Alzheimer's disease (AD). The cellular and molecular modifications responsible for these effects are not understood. Here we performed single-nucleus RNA-sequencing (snRNA-seq) employing both short- (Illumina) and long-read (Pacific Biosciences) sequencing technologies on a total of 29 DS and non-DS control prefrontal cortex samples. In DS, the ratio of inhibitory-to-excitatory neurons was significantly increased, which was not observed in previous reports examining sporadic AD. DS microglial transcriptomes displayed AD-related aging and activation signatures in advance of AD neuropathology, with increased microglial expression of C1q complement genes (associated with dendritic pruning) and the HSA21 transcription factor gene *RUNX1*. Long-read sequencing detected vast RNA isoform diversity within and among specific cell types including numerous novel sequences that differed between DS and normal brains. Notably, over 8,000 genes produced RNAs containing intra-exonic junctions, including amyloid precursor protein (*APP*) that had previously been associated with somatic gene recombination. These and related results illuminate large-scale cellular and transcriptomic alterations as novel features of the aging DS brain.

Significance Statement

Down syndrome (DS) neurocognitive disabilities associated with trisomy 21 are known; however, gene changes within individual brain cells occurring with age are unknown. Here, we interrogated >170,000 cells from 29 aging DS and control brains using single-nucleus RNA-sequencing. We observed increases in inhibitory-over-excitatory neurons, microglial activation in the youngest DS brains coinciding with overexpression of genes associated with microglial-mediated synaptic pruning, and overexpression of the chromosome 21 gene *RUNX1* that may be a potential driving factor in microglial activation. Single-nucleus long-read sequencing revealed hundreds of thousands of novel RNA transcripts. These included diverse species for

the Alzheimer's disease gene - amyloid precursor protein – that contained intra-exonic junctions (IEJs) previously associated with somatic gene recombination, also identified in ∼8,000 other genes.

MAIN TEXT

Introduction

Down Syndrome (DS) is a common genetic disorder affecting ∼1 in 700 live births (*1*). It is caused by the triplication of human chromosome 21 (HSA21) and results in numerous impairments. Brain abnormalities produce deficits in cognitive performance, learning, and language acquisition, as well as short and long-term memory impairment (*2*). As DS individuals age, they show increased incidence of dementia and neuropathological hallmarks of Alzheimer's disease (AD) by their 40's (*3*). The mechanistic etiology of the complex DS phenotype is known only in part. It includes defects in neuronal development (*4*) and GABA signaling (*5*). Alterations in dendritic spine dynamics have also been reported in DS models (*6-8*). It is hypothesized that the early onset of AD neuropathology and dementia in DS is driven by overexpression of genes located on HSA21 such as the kinase *DYRK1A* and especially amyloid precursor protein (*APP*) (*9*). Notably, increased brain transcription (*9*) and increased copy numbers of the *APP* gene have been linked to *APP* somatic gene recombination associated with sporadic AD. This form of gene recombination produced internally truncated RNA sequences containing intra-exonic junctions (IEJs) (*10*).

Single-cell sequencing technologies have opened new avenues to understanding cellular transcriptomics, particularly through the use of single-nucleus RNA-seq (snRNA-seq), which has been applied to normal (*11, 12*) and diseased (*13-15*) human brains, but has not been reported for postnatal or aging DS brains. Bulk RNA-sequencing studies of DS brains identified global alterations in gene expression (*16, 17*), but how specific cell types or RNA isoforms are impacted remains unknown. In addition, nearly all single-cell or single-nucleus transcriptomic studies lack unbiased RNA isoform information, which may have important biological consequences for

cellular function (*18, 19*). Here, we report single-nucleus analyses using both short and long-read snRNA-seq on aging DS brains versus normal controls. These results revealed significant differences between aging DS and normal brains in regard to their cellular composition and isoform-specific transcriptomes, including novel truncated RNAs containing IEJs and involving not only *APP* but thousands of other genes.

Results

DS and control brain sample characteristics

snRNA-seq was performed on DS and control samples (Fig. 1A). Human cerebral cortical Brodmann Areas 8/9 (BAs 8/9) from 56 DS and control brains were sectioned and assessed for cortical layers, RNA integrity number (RIN), and defined neuropathological signs of AD (Fig. S1A and Dataset S1). The prefrontal cortex was profiled because it is essential to memory and behavior, and gene expression differences could offer key insights into the DS brain (*20*). Samples with the confirmed presence of all six cortical layers and a RIN ≥ 6 were included in subsequent sequencing experiments. Samples with a RIN <6 were excluded from analysis because of negative trends in key single-cell output characteristics including the number of captured nuclei, total genes detected, and median genes per nucleus (Fig. S1B). Twenty-nine samples met the inclusion criteria (Fig. 1B) and were processed for snRNA-seq. Nine DS and 14 control samples were matched for age, sex, and RIN and were used for primary analyses (Fig. S1C and Dataset S1, Mann-Whitney U-test, P >0.1). Brains were categorized as "young" if they were ≤ 36 years of age, and thioflavin-S staining confirmed a lack of neuropathological hallmarks of AD (Fig. S1D). Another six control brains older than 70 years ("Ctrl-old") were processed for snRNA-seq to profile aging in control brains. In addition to matching samples for age, sex, and RIN, potential batch effects were accounted for by randomizing samples during processing and utilizing Seurat version 3 for analysis (see Methods)(*21*). A total of 172,237 filtered transcriptomic cell profiles were generated from snRNA-seq cDNA libraries (using 10X Genomics Single Cell 3' v3 system) and clustered using Seurat version 3 (*22*). Clusters were

identified with marker genes previously established in the human prefrontal cortex and labeled as: astrocytes, Ast; endothelial cells, End; excitatory neurons, Ex1-8; inhibitory neurons, In1-8; microglia, Mic; oligodendrocytes, Oli; oligodendrocyte precursor cells, OPC; and pericytes, Per (*11*)(Fig. 1C). This approach labeled 20 known neuronal subclusters, while only 11 were resolved by unbiased techniques (Fig. S1E). Clusters displayed a gene expression pattern similar to previous snRNA-seq classifications within the human brain (Fig. S1F, Dataset S2). All major cell types were present in each cohort and not significantly altered by sex or processing batch (Fig. S1G-I, Dataset S1). Pre-fragmented samples from the same cDNA libraries used for short-read sequencing enabled generation of approximately 98 million long reads that revealed 434,201 unique RNA isoforms with cellular barcodes.

Increased inhibitory:excitatory neuron ratios and neuronal subtype alterations in DS prefrontal cortex

An imbalance in inhibitory vs. excitatory neuronal firing has been reported in mouse models of DS (*23*). However, it is unclear if such an imbalance exists in human DS. At all examined ages, the proportion of inhibitory to total neurons was significantly increased in DS brains compared to controls (Fig. 1D, unpaired t-test, $p = 0.04$). A multiple linear regression analysis accounting for sex, RIN, age, and DS vs. control status was also performed. DS was the only variable with a significant effect ($p = 0.03$). Immunolabelling for inhibitory neurons supported the snRNA-seq data (Fig. S2A-B). An inhibitory:excitatory imbalance was not observed in published AD datasets (*13*)(Fig. S2C), potentially indicating that this is a DS-specific phenomenon. Notably, this imbalance was previously observed in single-cell ATAC-seq analysis of the Ts65Dn mouse model of DS (*24*)(Fig. S2D), further supporting the imbalance as a feature of the DS brain. Focused analyses on DS-young vs. Ctrl-young samples identified similar results (Fig. S2E), supporting early changes in DS. The proportions of inhibitory neurons from the In1a, In1b, In3, and In4b clusters were increased compared to controls (Fig. 2E). By comparison, proportions of excitatory neuronal clusters were relatively unchanged or slightly

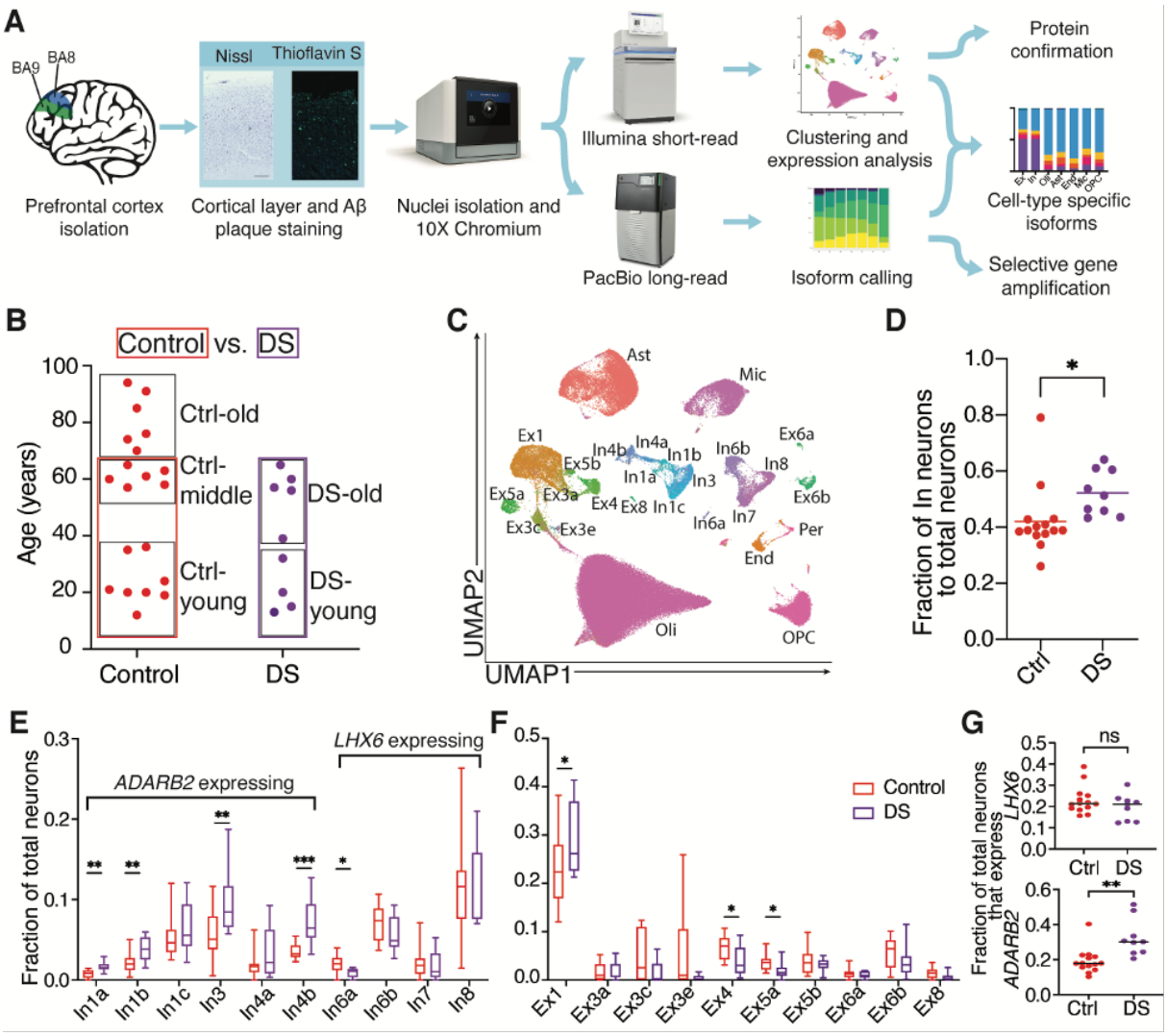

Figure 4.1. Experimental approach for cell clustering and altered neuronal fractions in DS. (**A**) Experimental outline for selecting samples and processing short and long-read sequencing of snRNA-seq data. (**B**) Ages for all samples analyzed by snRNA-seq and sample groups used in subsequent analyses (boxed). (**C**) UMAP and cell-type assignments of nuclei from DS and control age-matched brains. (**D**) Fraction of total neurons identified as inhibitory (In) in control and DS brains. (**E** and **F**) Fraction of inhibitory (E) and excitatory (Ex) (F) neuronal subtypes in control and DS brains. For (E and F) boxes extend from the 25th to 75th percentiles and whiskers extend from minimum to maximum values. (**G**) Fraction of inhibitory neurons that expressed *LHX6* or *ADARB2*. For (D-G) asterisks denote statistical significance in unpaired t-test (* p <0.05, ** p <0.01, *** p <0.001).

reduced at all ages with the exception of the Ex1 subcluster (Fig. 2F)(*11*).

Inhibitory neurons within the cerebral cortex arise developmentally from defined portions

of the embryonic ganglionic eminence (GE). In mice, interneurons expressing *LHX6* (In6-In8) originate from the medial ganglionic eminence (MGE) while interneurons expressing *ADARB2* (In1-In4) are derived from the caudal ganglionic eminence (CGE)(*25*). A distinction between *LHX6* and *ADARB2*-expressing neurons is also observed in humans (Fig. S1F). *LHX6*-expressing neurons were present in the human DS brain at similar proportions to controls, while *ADARB2*-expressing neurons were overrepresented in DS (Fig. 1G; unpaired t-test, $p < 0.01$), supporting a CGE origin of neuronal imbalance. These cellular subtype proportional changes were also observed in the DS-young cohort (Fig. S2F-H).

Cell-type-specific changes in HSA21 genes in the DS brain

HSA21 trisomy alters expression of HSA21 genes (*26*). However, direct proportionality between gene copy number and transcription is not expected (*16, 17, 27, 28*). Comparison of DS to control brains across genes on HSA21 in each individual cell type identified 308 of 4,008 genes with an expression fold change >1.1 in DS, and only 9 showed a >1.5-fold increase (Fig. S3A), signifying limited overexpression of HSA21 genes.

To further investigate HSA21 gene expression in DS brain cell types, data were filtered to focus on differentially expressed genes (DEGs) having a $\log_2$-fold change >0.25, a Bonferroni adjusted p-value <0.05, and expression in at least 10% of cells within analyzed cell types. A limited number of genes on HSA21 were observed as DEGs, but as expected, significantly more of these DEGs had increased expression in DS (Fig. 2A-B). Microglia had the greatest number of HSA21 DEGs (DEGs=25), with endothelial cells (DEGs=20) and neurons (Ex DEGs=17; In DEGs=15) also showing extensive changes in HSA21 gene expression (Fig. S3A). Differential expression analysis of all genes also identified the greatest expression changes occurring in microglia (Fig. S3B). Numerous genes on HSA21 were differentially expressed in multiple cell types, including the cell adhesion molecules, *NCAM2* and *DSCAM*, the splicing regulator, *SON*, and the kinase, *DYRK1A*. In neurons, *DSCAM* (*29*), *CXADR* (*30*), *APP* (*31*), and *NCAM2* (*32*) were altered in both excitatory and inhibitory neuronal populations in the DS brain; these genes

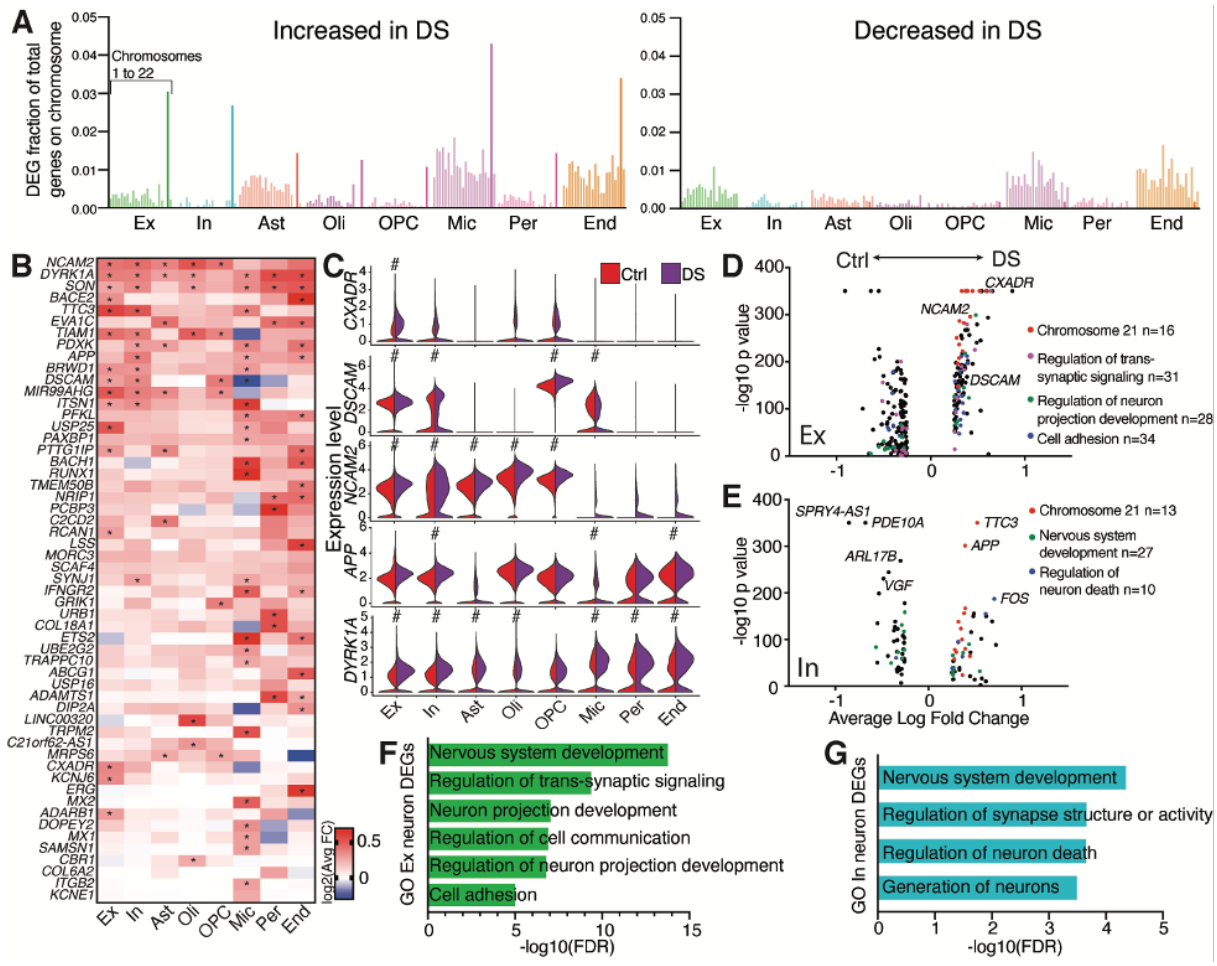

Figure 4.2. Gene expression changes in DS.
(**A**) Bar graphs displaying the fraction of annotated genes on each chromosome detected as an upregulated or downregulated differentially expressed gene (DEG) in DS compared to control brains for each specific cell type. Chr21 is bolded. (**B**) Heatmap displaying the $\log_2$-fold change of select HSA21 genes. Asterisks denote genes that meet the criteria as DEGs ($\log_2$-fold change >0.25, a Bonferroni adjusted p-value <0.05, and expression in at least 10% of cells within analyzed cell types). (**C**) Violin plots of five HSA21 genes hypothesized to play central roles in DS; # symbol denotes genes that meets DEG criteria. (**D** and **E**) Volcano plots displaying DEGs for each cell type, color coded by Gene Ontology (GO) biological processes classification. (**F** and **G**) Key biological processes determined by GO analyses for excitatory neuron (F) and inhibitory neuron (G) full transcriptome DEGs.

are directly involved in neuronal cell-cell interactions and neurite outgrowth (Fig. 2C). Notably, *DSCAM* was specifically expressed in *ADARB2*-expressing interneurons, potentially playing a key role in the observed overrepresentation of these cells. In contrast, *DSCAM* was very lowly expressed in *LHX6*-expressing interneurons (Fig. S3C), while its expression was significantly

downregulated in DS microglia compared to controls (Fig. 2B-C).

Dysregulation of key neurological pathways in DS revealed by Gene Ontology (GO)

To study functional changes in DS neurons, DEGs from the entire transcriptome were analyzed by Gene Ontology (GO)(*33-35*)(Fig. 2D-G). DS excitatory neuron DEGs were involved in the regulation of trans-synaptic signaling, the regulation of neuron projection development, and cell adhesion (Fig. 2D and F and Fig. S3D-E). Significantly upregulated genes included the ephrin receptors, *EPHA3*, *EPHA5*, and *EPHA6*, which are involved in neural development (*36*), the membrane receptors, *ROBO1* and *ROBO2*, and secreted guidance cues, *SEMA3C* and *SLIT2* (Dataset S3), which are all involved in axonal guidance and maintenance of synaptic connections (*37*).

Inhibitory neuron GO categories included nervous system development and regulation of neuron death (Fig. 2E and G and Fig. S2F-G). Intriguingly, the immediate early gene *FOS*, a marker of neuronal firing (*38*), was the most overexpressed gene not on HSA21 (Fig. 2E), supporting increased inhibitory neuron firing in addition to inhibitory neuron overrepresentation in the DS brain.

Variable expression of aging-associated genes in controls

Control brains were used to identify DEGs associated with brain age. No shift in the overall inhibitory:excitatory neuron ratio was observed with age (Fig. S4A). However, slight changes were observed in specific subtypes of both inhibitory and excitatory neurons, including a decrease in the number of *SST*-expressing interneurons (In7 and In8; Fig. S4B-C). Aging DEGs were identified from the comparisons between Ctrl-young vs. Ctrl-middle and Ctrl-middle vs. Ctrl-old, and particularly involved microglia and astrocytes (Fig. 3A and Dataset S6-8). To study the effects of aging on the transcriptome in a continuous manner, an unsupervised pseudotime-trajectory analysis using reversed graph embedding via Monocle 3 (*39*) was pursued for individual cell types. Microglia, astrocytes, and oligodendrocytes clustered in an age-dependent manner (Fig. S4D) and displayed pseudotime trajectories that clearly tracked from young to old samples

(Fig. 3B-G). Microglia clustered distinctly by age and showed widespread transcriptomic hallmarks of activation with increasing age (*15, 40*)(Fig. 3B) including increased expression of the inflammatory mediator, *SPP1*, and loss of both the chemokine receptor, *CX3CR1*, and the purinergic receptor, *P2RY13* (Fig. 3H).

Pseudotime analysis of astrocytes independently partitioned into two trajectories, each proceeding from young to old. One of the groups predominantly expressed two markers of astrocyte activation, *GFAP* and *FOS*, and is referred to as GFAP+ (Fig. S4E). Ctrl-young, Ctrl-middle, and Ctrl-old groups had 26%, 22%, and 39% of total astrocytes partition as GFAP+, respectively, signifying that, unlike microglia, there does not appear to be activation of all astrocytes during aging. However, both GFAP+ and GFAP- partitions showed transcriptomic signs of aging. Genes including *MMD2* and *ERBIN*, were significantly changed in each trajectory, while others, including the OPN receptor *CD44*, increased only in GFAP+ astrocytes (Fig. 3I).

As expected, concurrent with exhaustion of the OPC pool, the ratio of OPCs to oligodendrocytes decreased with age (Fig. S4F). This correlated with decreased expression of the AMPA receptor subunit *GRIA2*, which has been tied to oligodendrocyte survival and myelination (*41*)(Fig. 3J).

Early activation of DS microglia

Many DEGs identified in the aging brain were also differentially expressed in DS compared to age-matched control brains, with a striking signal in microglia, where 40 of 45 aging DEGs were also DS DEGs (Fig. S5A). Pseudotime analysis of microglia from all cohorts also indicated an aged microglial state in DS-young brains (Fig. S5B-C). To study microglial gene expression and different activation states, microglia from all samples were clustered separately from other cell types (Fig. 4A). Analyses were focused on clusters containing $>$2.5% of total microglia, resulting in 4 distinct microglial clusters with gene expression profiles similar to other single-nucleus microglial datasets (*15, 40*)(Fig. 4B). The largest cluster, labeled "Homeostatic," was comprised of microglia expressing homeostatic markers including *CX3CR1*, *P2RY12*, and

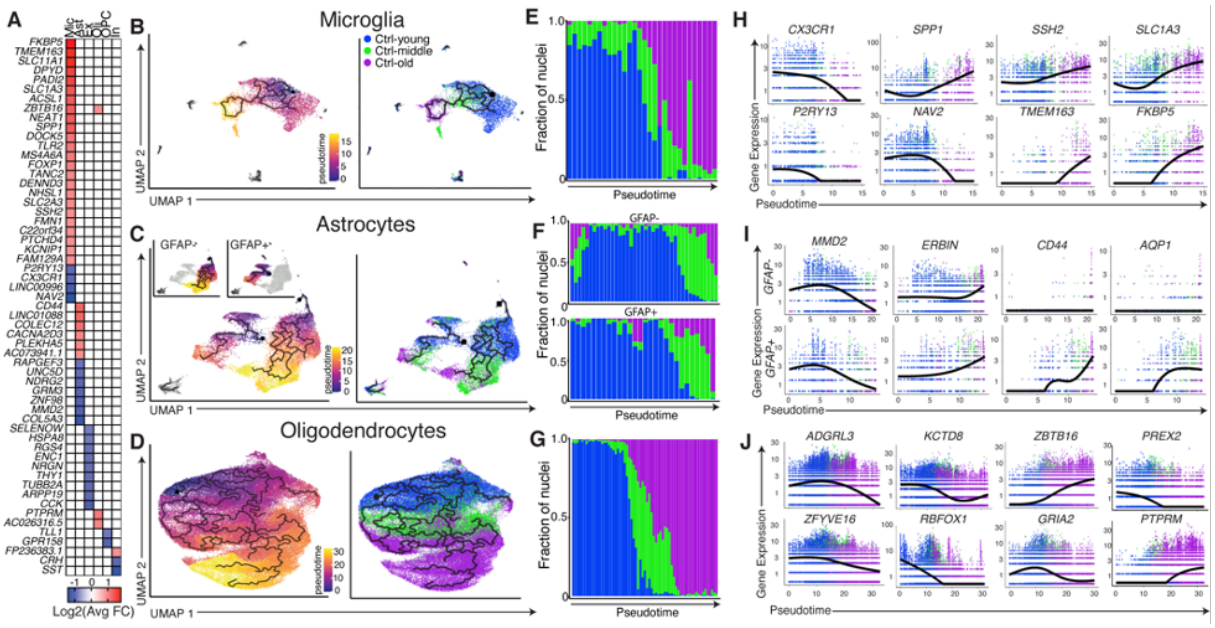

Figure 4.3. Cell-type-specific signatures of aging in control brains.
(**A**) Heatmap of most differentially expressed aging DEGs for each cell type. (**B**, **C**, and **D**) Unsupervised pseudotime trajectories with cells colored by pseudotime assignment (left) and age (right). OPCs are not included in oligodendrocyte analysis. (**H**, **I**, and **J**) Expression levels of aging DEGs of interest with respect to pseudotime. (F) includes plots for both the GFAP+ partition and GFAP- partition with respect to pseudotime. In (B-G and H-J), each dot represents a single nucleus.

P2RY13, while lacking activation markers such as *SPP1*. A second cluster, labeled "Activated," had high expression of complement components, *C1QA*, *C1QB*, and *C1QC*, as well as *CD14*, *ERC2*, and *PTPN2*. The third major cluster, "Antigen presenting," contained highly expressed genes associated with antigen presentation including *CD83*, *HLA-DRA*, *HLA-DRB1*, and *HLA-DPB1*, as well as *PADI2*, *MSR1*, and *APOC1*. Lastly, a small subset of microglia expressed transcripts typically associated with oligodendrocytes, specifically *MBP*, *PLP1*, and *ST18*; these "Phagocytosing" microglia are hypothesized to internalize oligodendrocyte transcripts while phagocytosing myelin (*15*). Strikingly, >80% of microglia from every Ctrl-young sample clustered as homeostatic, contrasting with the age-matched DS-young cohort that averaged only 28% (Fig. 4C). DS-young microglia were largely classified as activated (Fig. 4C). As expected, DS-old microglia clustered as both activated and antigen presenting, likely associated with the

AD pathology in this cohort (Fig. 4C). Broad increases in expression of microglial activation markers and a loss of homeostatic gene expression were observed in microglia in all DS samples, as well as the DS-young cohort (Fig 4D). Transcripts of *CX3CR1* and *C1QA* were observed in generally distinct cells (Fig. S5D), and *CX3CR1* transcripts were sequenced primarily in microglia (Fig. S5E). Significant decreases in CX3CR1 protein were observed in DS-young brains (Fig. 4E).

AD-associated gene upregulation in DS microglia

AD neuropathology uniformly occurs in DS individuals beyond age 40. A direct comparison of DS microglia to human AD microglia (*13*) revealed a shared increase in *APOE* and *PTPRG* expression (Fig. S5F). However, DS microglia displayed distinct profiles wherein most genes that were downregulated in AD showed upregulation in DS (Fig. S5F) and numerous DEGs in DS microglia were not identified as DEGs in AD. Microglial gene expression was also compared to previously defined disease-associated microglia (DAM) expression signatures (*42*)(Fig. S5G). Antigen presenting microglia most closely resembled DAMs, displaying increased expression of many DAM upregulated DEGs and decreased expression of DAM downregulated DEGs.

As previously noted, the DS-young cohort lacked the pathological hallmarks of AD. However, microglial genes associated with the earliest signs of AD onset (*13*) were upregulated in DS-young microglia including *VSIG4*, *ADGRG1*, *CACNA1A*, and *C1QC* (Fig. S5H), supporting overlap of microglial AD-like activation occurring in the young DS brain. With increasing age, DS microglia showed reductions in complement-associated genes and accompanying increases in antigen presentation-associated genes like the major histocompatibility complex (MHC) and *CD83* (Fig. S5I). These results support precocious and evolving microglial activation states with age, modified for differing age-dependent activities, including in response to developing AD.

Increased expression in DS of microglial genes related to synaptic function

Microglia are implicated in synapse and memory loss through both complement-mediated (*43, 44*) and *ADGRG1*-mediated (*45*) pathways. Remarkably, all three gene components of

complement C1q (*C1QA*, *C1QB*, *C1QC*) as well as *ADGRG1* were significantly overexpressed in DS microglia, particularly in the DS-young cohort (Fig. 4F, Fig. S5J), suggesting that overactive pruning by microglia may occur in DS. In addition, a decrease of *P2RY12* expression was identified (Fig. 4D); loss or inhibition of *P2RY12* has been linked to impaired synaptic function (*46*). These microglial transcriptomic alterations affecting neurons may contribute to neurocognitive changes in DS.

A significant decrease in the density of dendritic spines was reported in the Dp16 mouse model of DS, which could be reversed by the depletion or inhibition of microglia (*47*). However, these findings directly conflict with data from the Tc1 and Ts1Rhr mouse models of DS that show no changes in dendritic spine density (*7*). To discern if a HSA21 gene might be responsible for this discrepancy, we profiled the HSA21 DEGs in human microglia and cross referenced these with genes triplicated in the Dp16 model, but functionally diploid in Tc1 and Ts1Rhr mice (*48-50*). Multiple microglial DEGs were triplicated in Dp16 but not Tc1 or Ts1Rhr mice. These genes included the IFN receptor, *IFNGR2*, the splicing regulator, *SON*, and most significantly, the transcription factor, *RUNX1* (Fig. 4G). *RUNX1* overexpression was observed broadly in microglia across all DS samples (Fig. S5J) and was the most overexpressed HSA21 microglial DEG (Fig. 4G). RUNX1 is a key transcription factor in regulating microglial gene expression (*51*), and its expression typically decreases after early neurodevelopment but can be induced following brain injury in adults (*52*). *TRPM2*, a calcium channel that has been tied to microglia activation (*53*), was one of the top microglial DEGs but is not triplicated in any of the three mouse models discussed, signifying that microglia activation may be even more striking in the human brain. Furthermore, *BACH1*, triplicated in Tc1 and Dp16 mice, encodes a transcriptional repressor involved in the development of numerous antigen presenting cell subtypes including macrophages, and its diminished expression correlates with protective autoimmune effects (*54, 55*), which may be relevant to microglial activation states.

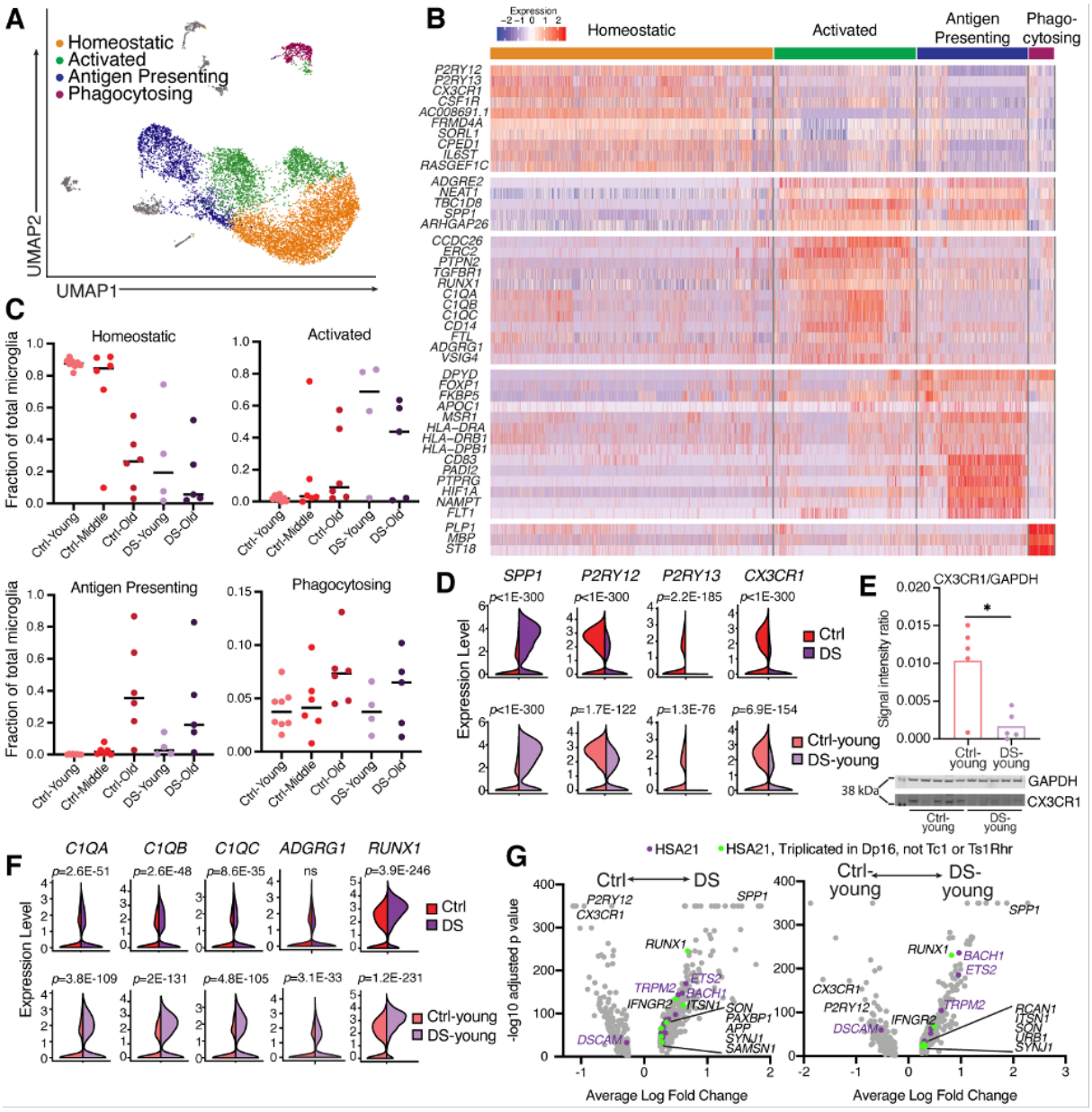

Figure 4.4. Hallmarks of microglial activation in DS microglia.
(**A**) UMAP of microglia from all processed samples, colored by microglial sub-cluster. (**B**) Heatmap displaying key differentially expressed genes used to define microglial sub-clusters. (**C**) Fraction of total microglia from each sample that clustered in each of the four major microglia sub-clusters. (**D**) Violin plots of gene expression for hallmark microglial activation genes from DS vs Ctrl and DS-young vs Ctrl-young cohorts, adjusted p-value using Bonferroni correction on Wilcoxon rank sum test. (**E**) Western blot for CX3CR1 and quantification relative to GAPDH. Asterisk denotes statistical significance in unpaired t-test (p = 0.011). (**F**) Violin plots of gene expression for C1q complement genes, *ADGRG1*, and *RUNX1*. (**G**) Volcano plots for total DEGs in microglia (gray), DEGs from HSA21 that are triplicated in the Dp16 mouse model but not Tc1 or Ts1Rhr (green) and all other HSA21 microglia DEGs (purple) (*48-50*).

Transcriptomic diversity discovered through long-read sequencing of single-nucleus cDNA libraries

Long-read single molecular real-time (SMRT) sequencing enables profiling of full-length RNA isoforms from single-cell cDNA libraries (*56, 57*). Single-nucleus cDNA libraries were sequenced with SMRT sequencing to obtain approximately 98 million long reads from 16 individual brains, eight each from control and DS cohorts (Fig. S6A). Each sample was sequenced to a depth of 5.5 - 7.5 million raw long reads. Of these, 34,988,576 total reads had both a cellular barcode and a unique molecular identifier (UMI). Reads were analyzed using cDNA_Cupcake and SQANTI2 (*58*) to identify isoforms and group them into 4 main categories defined as: Full Splice Match (FSM) isoforms that match GENCODE v28 annotations; Incomplete Splice Match (ISM) isoforms that only partially match annotations and result from 3' and/or 5' truncations; Novel In Catalog (NIC) isoforms that have not been annotated but contain known splice sites and exons; and Novel Not in Catalog (NNC) isoforms that contain at least one novel splice site. After filtering, 434,201 unique isoforms remained, supported by a total of 6,905,832 reads (Fig. S6B), and 47.7% of these isoforms were supported by at least two reads with distinct UMIs. Matching the cell barcodes back to the originating cell identified by short-read sequencing enabled cell type identification for 40.42% of the isoform reads. A majority of the reads that were not associated with a cell type had a barcode that either corresponded to a cell that did not pass QC in the Seurat analysis (64.7% of unidentified reads) or was determined to be background in the cellranger analysis (25.4%). A small percentage of these reads contained known 10X Genomics cellular barcodes that were not observed in the short-read dataset (1.6%). The remaining reads had cellular barcodes that were not among the available 10X Genomics cellular barcodes and could have resulted from error introduced during library preparation or sequencing (8.3%). Interestingly, long-read coverage was enough to identify most cell types (Supplementary Text).

Vast isoform diversity was observed in the brain. Novel isoforms (NIC and NNC) displayed greater variation than annotated forms (FSM)(Fig. 5A), but the overall proportion of

novel isoforms did not change with age or across DS and control cohorts (Fig. 5B, two-way ANOVA, NIC by cohort p = 0.96, NNC by cohort p = 0.13). However, the proportion of novel isoforms did vary with cell type. Analysis of all cells showed the greatest NIC and NNC isoform diversity in astrocytes, whereas endothelial cells and pericytes showed the least (Fig. 5C, NIC by cell type $p < 0.0001$, NNC by cell type $p < 0.0001$). Excitatory neurons, inhibitory neurons, and oligodendrocytes also showed NNC enrichment (Fig. 5C). Multiple types of NNC isoforms were observed and included features such as: novel exon junctions within an intron that created an entirely new exon that does not overlap with any previously annotated exon sequences; intra-exonic junctions (IEJs)(*10*) that were formed by joining the internal regions of two exons; and intron retention junctions that were formed by a new splice site that extends exon coordinates partially into the next intron (Fig. 6SC-E; IEJs were found in over 8,000 genes (Dataset S10)). Notably, prior studies using cap analysis gene expression (CAGE) sequencing demonstrated that 94% of NIC and 87% of NNC reads have 5' ends that contain the transcription start site (*59*), supporting the conclusion that NIC and NNC reads are indeed bona fide transcripts and not sequencing artifacts.

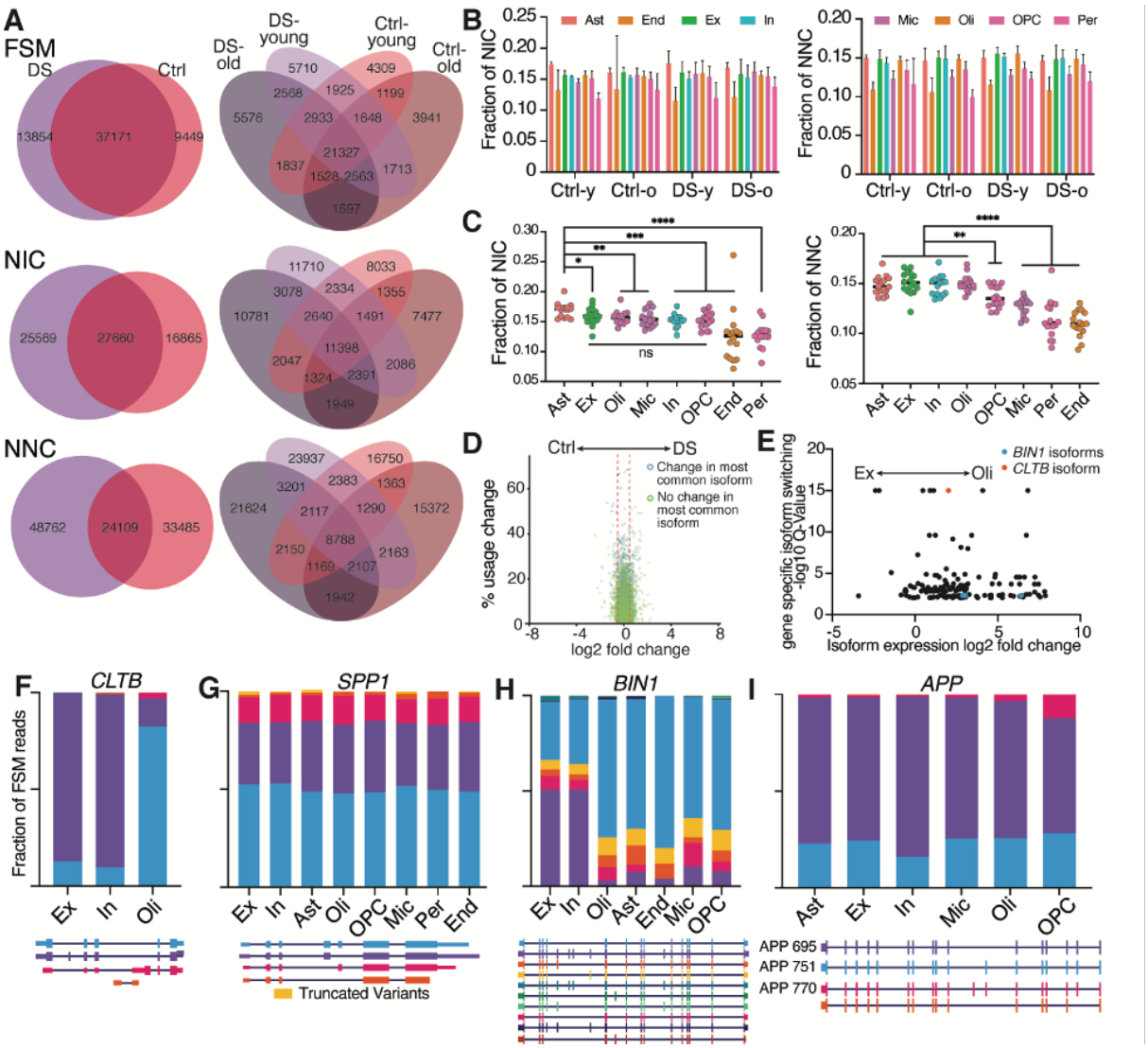

Figure 4.5. Novel isoform variants and specific isoform changes in different brain cell types. (**A**) Overlap of full splice match (FSM), novel in catalog (NIC), and novel not in catalog (NNC) isoforms across DS, aging, and control (Ctrl) cohorts. (**B**) Fraction of total isoforms called as NIC or NNC by cohort and cell type. Two-way ANOVA, NIC by cohort p=0.96, NIC by cell type p<0.0001, NNC by cohort p=0.13, NNC by cell type p<0.0001. Error bars represent one standard deviation. (**C**) Fraction of isoforms classified as NIC or NNC for each sample by cell type. Asterisks denote statistical significance in unpaired t-test (*p<0.05, **p<0.01, ***p<0.001, ****p<0.0001). (**D**) Volcano plot of isoform usage differences between DS and control cohorts. % usage change represents the redistribution of isoform proportions within a gene. Log2FC is the change in expression of an isoform between two conditions. (**E**) Plot of significantly changed isoforms between excitatory and oligodendrocyte populations. Q-value represents the false discovery rate for which at least one isoform of a particular gene displays differential proportionality across compared groups, signifying changes in isoform usage at the gene level. Only isoforms with <0.01 Q value and non-zero expression in both cell types are displayed. (**F-I**) Cell-type-specific isoform proportions for cell types with over 50 unique reads mapping to FSM isoforms for *CLTB* (F), *SPP1* (G), *BIN1* (H), and *APP* (I).

Differential isoform expression and usage are observed across cell types

Identification of specific isoforms that are differentially expressed could potentially offer targets for modern therapeutics, such as antisense oligonucleotides or gene therapies (*60, 61*). Differential isoform expression and proportional isoform usage were analyzed using tappAS (*62*). Limited differential isoform expression or differential isoform usage was identified between control and DS samples (Fig. 5D). However, pairwise comparisons of isoform usage between cell types revealed numerous genes for which a cell type preferentially utilized one isoform over others (Fig. 5E and Fig. S7A). Numerous genes switched cell-type-specific isoforms including *SEPT8*, *RPL13* (Fig. S7B-C) and *CLTB* (Fig. 5F), which plays an important role in clathrin-mediated endocytosis and utilized a different isoform in neurons compared to oligodendrocytes.

To characterize isoform diversity further, *SPP1*, *BIN1*, and *APP* were selectively amplified from single-nucleus cDNA libraries using the Read1 primer from the 10X adapter and primers designed against their 5'UTRs. Minimal changes in the proportional expression of *SPP1* FSM isoforms across cell types (Fig. 5G) or between DS and control cohorts were detected (Fig. S7D). A significantly higher proportion of reads were found originating from microglia in DS as compared to controls (Fig. S7E), which is consistent with the differential expression displayed in the short-read data. Untargeted sequencing identified a *SPP1* NNC isoform with a novel exon that was confirmed with targeted sequencing. Targeted sequencing also identified an additional 4 isoforms containing this exon; altogether, this novel exon was supported by 73 UMIs (Fig. S7F).

In genome-wide association studies, mutations in a region upstream of *BIN1* showed the second highest odds-ratio for sporadic AD (*63, 64*). *BIN1* transcripts were selectively amplified, which revealed cell-type-specific isoform switching similar to patterns reported in mice (*56*). The shortest isoform that lacks all alternatively spliced exons was predominantly sequenced in non-neuronal cell types, whereas the longest isoform that contains all alternatively spliced exons was the predominant isoform in both excitatory and inhibitory neurons (Fig. 5H).

The AD-associated gene, *APP*, expresses two major brain isoforms, encoding APP-695 and APP-751, with the literature supporting neuron-specific expression of APP-695 that lacks a

Kunitz-type serine protease inhibitory domain, implicating neurons as the source for soluble Aβ in the brain (65, 66). However, *APP-695* was observed to be the predominant RNA isoform in all cell types (Fig. 5I). Furthermore, total expression levels of *APP* are similar across neurons, oligodendrocytes, OPCs, pericytes, and endothelial cells (Fig. 2C), signifying that many cell types in the human brain contribute significantly to the RNA expression of *APP-695* rather than just neurons. Isoform diversity also included NNC species containing IEJs in *APP*, consistent with the literature (*10*)(Fig. S8A).

Discussion

Transcriptomic effects of HSA21 trisomy at the level of single cells in the postnatal and aging DS brain have not been previously reported. snRNA-seq using short and long-read sequencing, as well as targeted-gene approaches, revealed differences involving multiple transcriptomic pathways and cell types. Most transcriptomic changes affected non-HSA21 genes, supporting global effects of HSA21 trisomy on the transcriptome. However, notable exceptions included *APP*, *NCAM2*, *DYRK1A*, *SON*, *BACE2*, and *TTC3*, indicating dosage effects on select HSA21 genes within specific cell types. Increased neuronal inhibitory:excitatory ratios and increased neurodevelopmental gene expression existed at all examined ages in DS. Prominently, microglia exhibited transcriptomic states indicative of activation. Cell autonomous causes could include overexpression of the HSA21 transcription factors *RUNX1* and *BACH1*. Furthermore, increased C1q expression could directly affect neuronal process pruning. Aging signatures in the DS brain paralleled gene expression patterns reported in multiple neurological disorders, particularly AD (*13, 67-69*).

The current study represents the first single-nucleus transcriptome analysis of the post-natal human DS brain and the largest single-nucleus profiling of RNA isoforms in the human brain to date. Limitations of this study include focused analyses of BA8,9 of the prefrontal cortex, relatively limited numbers of DS brains, and assessment of $\sim$6,000 cells per brain that while standard, represents a small percentage of total brain cells. These results identified trends

that will benefit from expanded analyses in the future. At least three identified features deserve additional comment.

First, neurodevelopmental transcriptome differences are prominent in DS brain cells, including those of cell adhesion genes like *DSCAM* (*29*), *CXADR* (*30*), *APP* (*31*) and *NCAM2* (*32*), as well as genes of the Robo-Slit-Ephrin pathways that normally contribute to axonal guidance, synapse formation, and neurogenesis (*36, 37*). Abnormal neurodevelopmental programs in DS are further supported by increased *ADARB2*-expressing (CGE-derived), but not *LHX6*-expressing (MGE-derived) inhibitory neuron ratios. CGE-derived neurons migrate prenatally from the CGE to the cortex (*70, 71*), and this increase could result in enhanced neuronal inhibition as supported by DS animal models (*72, 73*). In human DS, an imbalance of inhibition and excitation may exist considering the clinical reports of elevated seizure activity (*74*). This dichotomy may be explained by variations in neuronal subsets such as the statistically significant increase in one cluster of excitatory neurons, Ex1, contrasting with other Ex clusters in DS, as well as by other possible changes affecting epileptogenic parts of the brain that were not assessed.

Second, microglia show major transcriptomic differences at all ages in DS, indicative of activation even at the youngest age examined. Activation could again reflect cell autonomous mechanisms potentially resulting from increased *RUNX1* expression. Alternatively, microglial activation could be indicative of non-cell autonomous mechanisms activated by mismatched neurodevelopmental activities involving C1q complement and other pruning genes associated with exuberant axon outgrowth and synapse formation/elimination, whereby microglia would face a chronic activating-milieu to remove surplus or mismatched process outgrowth and neuronal connections. The downregulation of genes like *P2RY12* in microglia may also contribute to the increased prevalence of seizures in DS (*46*). In later adult life, microglial activation could additionally reflect stimuli associated with incipient AD and contribute further to neurocognitive deficits. snRNA-seq was sufficient to distinguish activation states, supporting the possibility of a distinct microglial transcriptomic profile compared to AD. The combination of persistent neurodevelopmental gene expression and induced microglial activation provide a novel facet on

functional deficits within the DS brain and may distinguish it from AD signatures.

Third, isoform resolution in single-cell transcriptomic profiling is essential to generating a full understanding of transcriptional biology yet cannot be achieved by standard 3'-short-read sequencing techniques. RNA splice variants have important roles in development and disease (*19, 75*), and these new data provide an initial platform for approaching cell-type-specific isoforms in the DS and normal brain. Additionally, this study explored the potential for identifying cell types using only long reads, an approach that would eliminate the need for short-read sequencing in cell-type-specific isoform profiling, and suggests that this can be achieved with reasonable accuracy, but would require greater sequencing depth for optimal identification. Notably, isoform diversity and usage varied extensively across cell types, while being relatively stable within cell types between disease cohorts, supporting isoform functions in maintaining cell identity. The thousands of novel sequences beyond known splice variants, whose functions are unknown, provide a new reservoir of transcripts towards understanding the normal and diseased brain. These include isoforms with novel structures like IEJs that were detected on *APP* and over 8,000 thousand additional genes, which might reflect the widespread operation of somatic gene recombination mechanisms including those relevant to AD (*10*). Additional experiments are required to determine if these IEJs reflect expression of somatically recombined genes and/or if they are novel splicing variants. Overall, these snRNA-seq studies of the normal aging and DS brain implicate both intrinsic neurodevelopmental cellular processes and RNA isoform diversity, providing new understanding and novel therapeutic targets to aid DS individuals.

Materials and Methods

See Supplementary Information for more details.

Tissue sampling and preparation

Frozen tissue samples from Brodmann Area (BA) 8 or 9 of the prefrontal cortex were obtained from multiple sources and stored at -80°C. Samples were sectioned in a cryostat set at -20°C.

RNA integrity measurement

RNA was isolated using a RNeasy isolation kit from Qiagen and evaluated on an Agilent 4200 TapeStation.

Thioflavin S staining

Tissue sections (20μm) were stained using thioflavin S to visualize amyloid plaques and tau tangles as hallmarks of Alzheimer's disease (AD) pathology.

Nissl staining

Tissue sections (20μm) were stained using Cresyl Violet to visualize the cortical layers of each section.

Nuclei isolation and generation of amplified cDNA libraries

DS and control samples were randomized and processed in groups of four to negate potential batch process variation. Tissue sections (300μm) were removed from frozen storage and immediately submerged in 1 mL of nuclei isolation buffer (20mM Tris, 320mM Sucrose, 5mM CaCl2, 3mM MgAc2, 0.1mM EDTA, 0.1% Triton-X 100, 0.2% RNase Inhibitor)(*10, 11*). Extracted nuclei were washed twice in PBS + 0.25mM EGTA + 1% BSA + 0.2% RNase inhibitors (Takara Bio, Mountain View, CA). They were then suspended in PBSE + BSA + RNase inhibitors + 1.25ug/mL 4',6-diamidino-2-phenylindole (DAPI) (Sigma, St. Louis, MO). FANS was performed on a FACSAria Fusion (BD Biosciences, Franklin Lakes, NJ) gating out debris from FSC and SSC plots and selecting DAPI+ singlets. Samples were kept on ice until sorting was complete and were immediately processed after sorting. Sorted nuclei were diluted to $\sim$700-1,500 nuclei/mL, and a final concentration was determined using a fluorescent cell counter. The 10X Genomics Single Cell 3' v3 kit was then used to prepare samples targeting 10,000 single nuclei GEMs. The protocol was followed without deviation prior to fragmentation of the cDNA libraries.

cDNA preparation and long-read sequencing

Fifty percent of the pre-fragmented cDNA library was used for long-read sequencing. If the cDNA

input concentration was too low for Pacific Biosciences (PacBio) library preparation, the cDNA library was re-amplified (Supplemental Table 1) using the same reagents and concentrations as outlined in the 10X Genomics kit protocol. 100ng of cDNA was used in the PacBio procedure for sequencing. Each sample was sequenced in an individual SMRTcell. An average of 6.003 million polymerase reads were obtained per sample.

Selective amplification of cDNA libraries and subsequent long-read sequencing

Selective amplification of the genes *APP*, *SPP1*, and *BIN1* was pursued using custom designed primers and the Read 1 primer from the 10X Genomics preparation. The same cDNA libraries used for long-read analysis were linearly amplified with only the 5' UTR primer present prior to addition of the Read 1 primer. Samples were cleaned with Pronex beads and were sequenced with PacBio Sequel II as outlined above.

Short-read snRNA-seq data processing and filtering

10X Genomics CellRanger software (v3.0.2) was used to demultiplex samples, align reads, quantify unique molecular identifiers (UMIs), and generate cell count matrices. Default parameters were used, with the exception of a pre-mRNA reference file (ENSEMBL GRCh38) to capture intronic reads originating from pre-mRNA species present in the nuclei. Using Seurat (v3.0.3), sample matrices were filtered and normalized by the default global-scaling method in Seurat.

Clustering and UMAP visualization

Lake et al.'s (*11*) dataset was used as a reference with Seurat's TransferData function to label cell types in our samples. Seurat objects from the samples in the same disease/age group were merged (Seurat merge function). For comparisons between two groups, differential expression analysis and pseudotime analysis, merged samples within a group were integrated (Seurat IntegrateData function). The integrated data was then scaled and UMAP embeddings were generated.

GAD67/NeuN staining

Tissue sections (20μm) from DS-young and age matched Ctrl-young samples were co-stained

for GAD67 and NeuN and imaged. After imaging, cells were counted in three separate rows for each section spanning from white matter to the pial surface.

Multiple linear regression of inhibitory:excitatory ratios

For DS and control cohorts, data on sex, RIN, age, and DS vs control status were collected. Data was input into tables and Prism was utilized to calculate a multiple linear least squares regression for which the independent variable was ex:in ratio. Sex and DS status were assigned binary indicator variables. No weighting was utilized and no 2-way or 3-way interactions were accounted for.

Differential Gene Expression (DEG) analysis

Seurat was used to identify differentially expressed genes (DEGs) in DS compared to control samples by cell type and between age groups. Default parameters for FindMarkers were used to identify DEGs that were expressed in at least 10% of either of the populations being compared, had at least a 0.25 log fold difference, and were significant based on a Wilcoxon Rank Sum test.

Gene enrichment analysis

Gene Ontology (GO) analysis was conducted using PANTHER (*34*, *35*, *76*).

Pseudotime analysis

Count matrices and UMAP projections of specific cell types from Seurat analysis were loaded into Monocle3 (v0.2.1). Cells were partitioned, and pseudotime trajectories were learned and plotted. Endpoints that clustered with the youngest samples' cells were chosen as the roots for each graph. Differential expression analysis was completed to determine which genes had expression that varied as a function of pseudotime. Astrocytes separated into two partitions that were analyzed individually.

Processing of long reads and isoform calling

Samples from both untargeted and targeted long-read datasets were demultiplexed and barcodes

were removed using lima (v1.10.0). Following the recommendations in the cDNA_Cupcake repository (version updated 02/07/2020) for single-cell isoform analysis, CCS reads were generated using ccs (v4.2.0) with the following parameters: --minPasses 1 --min-rq 0.8 --minLength 50 --maxLength 21000. 10X Genomics R1 and TSO primer sequences and reads with improper primer orientation were removed using lima with the parameter --isoseq. UMIs and cellular barcodes were identified for each read. Isoseq3 refine (v3.2.2) was used to remove poly-A tails and artificial concatemers before mapping to the human reference genome (GRCh38). cDNA_Cupcake's (v9.0.1) collapse_isoforms_by_sam.py was used to collapse redundant isoforms. SQANTI2 (v7.3.2) was used to filter out mono-exon isoforms and artifacts of intra-priming and annotate the identified isoforms. Scripts from cDNA_Cupcake were used to assign UMIs/barcodes and isoforms back to specific reads. Original scripts were written to match specific reads back to sample and cell type, summarize which samples each isoform was detected in, and visualize the resulting isoforms in UCSC Genome Browser.

Differential isoform expression and usage analysis

The protocol for using tappAS (v0.99.15) for "Data from Long-read Sequencing Technology" was followed.

RNAscope for microglial gene markers

Sections of tissue (20μm) were cut and processed using the recommended kit protocol (2.5 HD Duplex Assay, Advanced Cell Diagnostics 322500). Probes applied were *C1QA* (485451-C2) and *CX3CR1* (411251). Slides were imaged at 40X magnification.

Western blot analysis

Sections (200 μm) of 5 control and 5 DS brains (including 1 with a RIN below the cutoff for sequencing analysis) were lysed in RIPA buffer separated on an Invitrogen Bolt 4-12% Bis-Tris protein gel and transferred to a PVDF membrane. The blot was probed with antibodies to CX3CR1 (Invitrogen #14-6093-81) and GAPDH (Invitrogen #AM4300) and visualized using a

LI-COR Biosciences CLx Imager. Bands were quantitated using the LI-COR Image Studio Lite software.

References and Notes

1. Parker, S. E., Mai, C. T., Canfield, M. A., Rickard, R., Wang, Y., Meyer, R. E., Anderson, P., Mason, C. A., Collins, J. S., Kirby, R. S., Correa, A. and National Birth Defects Prevention, Network, Updated National Birth Prevalence estimates for selected birth defects in the United States, 2004-2006. *Birth Defects Res A Clin Mol Teratol* **88**, 1008-1016 (2010).

2. Godfrey, M. and Lee, N. R., Memory profiles in Down syndrome across development: a review of memory abilities through the lifespan. *J Neurodev Disord* **10**, 5 (2018).

3. Wisniewski, K. E., Wisniewski, H. M. and Wen, G. Y., Occurrence of neuropathological changes and dementia of Alzheimer's disease in Down's syndrome. *Ann Neurol* **17**, 278-282 (1985).

4. Haydar, T. F. and Reeves, R. H., Trisomy 21 and early brain development. *Trends Neurosci* **35**, 81-91 (2012).

5. Contestabile, A., Magara, S. and Cancedda, L., The GABAergic Hypothesis for Cognitive Disabilities in Down Syndrome. *Front Cell Neurosci* **11**, 54 (2017).

6. Chiotto, A. M. A., Migliorero, M., Pallavicini, G., Bianchi, F. T., Gai, M., Di Cunto, F. and Berto, G. E., Neuronal Cell-Intrinsic Defects in Mouse Models of Down Syndrome. *Front Neurosci* **13**, 1081 (2019).

7. Haas, M. A., Bell, D., Slender, A., Lana-Elola, E., Watson-Scales, S., Fisher, E. M., Tybulewicz, V. L. and Guillemot, F., Alterations to dendritic spine morphology, but not dendrite patterning, of cortical projection neurons in Tc1 and Ts1Rhr mouse models of Down syndrome. *PLoS One* **8**, e78561 (2013).

8. Real, R., Peter, M., Trabalza, A., Khan, S., Smith, M. A., Dopp, J., Barnes, S. J., Momoh, A., Strano, A., Volpi, E., Knott, G., Livesey, F. J. and De Paola, V., In vivo modeling of human neuron dynamics and Down syndrome. *Science* **362**, (2018).

9. Wiseman, F. K., Al-Janabi, T., Hardy, J., Karmiloff-Smith, A., Nizetic, D., Tybulewicz, V. L., Fisher, E. M. and Strydom, A., A genetic cause of Alzheimer disease: mechanistic insights from Down syndrome. *Nat Rev Neurosci* **16**, 564-574 (2015).

10. Lee, M. H., Siddoway, B., Kaeser, G. E., Segota, I., Rivera, R., Romanow, W. J., Liu, C. S., Park, C., Kennedy, G., Long, T. and Chun, J., Somatic APP gene recombination in Alzheimer's disease and normal neurons. *Nature* **563**, 639-645 (2018).

11. Lake, B. B., Chen, S., Sos, B. C., Fan, J., Kaeser, G. E., Yung, Y. C., Duong, T. E., Gao, D., Chun, J., Kharchenko, P. V. and Zhang, K., Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat Biotechnol* **36**, 70-80 (2018).

12. Hodge, R. D., Bakken, T. E., Miller, J. A., Smith, K. A., Barkan, E. R., Graybuck, L. T., Close, J. L., Long, B., Johansen, N., Penn, O., Yao, Z., Eggermont, J., Höllt, T., Levi, B. P., Shehata, S. I., Aevermann, B., Beller, A., Bertagnolli, D., Brouner, K., Casper, T., Cobbs, C., Dalley, R., Dee, N., Ding, S. L., Ellenbogen, R. G., Fong, O., Garren, E., Goldy, J., Gwinn, R. P., Hirschstein, D., Keene, C. D., Keshk, M., Ko, A. L., Lathia, K., Mahfouz, A., Maltzer, Z., McGraw, M., Nguyen, T. N., Nyhus, J., Ojemann, J. G., Oldre, A., Parry, S., Reynolds, S., Rimorin, C., Shapovalova, N. V., Somasundaram, S., Szafer, A., Thomsen, E. R., Tieu, M., Quon, G., Scheuermann, R. H., Yuste, R., Sunkin, S. M., Lelieveldt, B., Feng, D., Ng, L., Bernard, A., Hawrylycz, M., Phillips, J. W., Tasic, B., Zeng, H., Jones, A. R., Koch, C. and Lein, E. S., Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**, 61-68 (2019).

13. Mathys, H., Davila-Velderrain, J., Peng, Z., Gao, F., Mohammadi, S., Young, J. Z., Menon, M., He, L., Abdurrob, F., Jiang, X., Martorell, A. J., Ransohoff, R. M., Hafler, B. P., Bennett, D. A., Kellis, M. and Tsai, L. H., Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332-337 (2019).

14. Velmeshev, D., Schirmer, L., Jung, D., Haeussler, M., Perez, Y., Mayer, S., Bhaduri, A., Goyal, N., Rowitch, D. H. and Kriegstein, A. R., Single-cell genomics identifies cell type-specific molecular changes in autism. *Science* **364**, 685-689 (2019).

15. Schirmer, L., Velmeshev, D., Holmqvist, S., Kaufmann, M., Werneburg, S., Jung, D., Vistnes, S., Stockley, J. H., Young, A., Steindel, M., Tung, B., Goyal, N., Bhaduri, A., Mayer, S., Engler, J. B., Bayraktar, O. A., Franklin, R. J. M., Haeussler, M., Reynolds, R., Schafer, D. P., Friese, M. A., Shiow, L. R., Kriegstein, A. R. and Rowitch, D. H., Neuronal vulnerability and multilineage diversity in multiple sclerosis. *Nature* **573**, 75-82 (2019).

16. Olmos-Serrano, J. L., Kang, H. J., Tyler, W. A., Silbereis, J. C., Cheng, F., Zhu, Y., Pletikos, M., Jankovic-Rapan, L., Cramer, N. P., Galdzicki, Z., Goodliffe, J., Peters, A., Sethares, C., Delalle, I., Golden, J. A., Haydar, T. F. and Sestan, N., Down Syndrome Developmental Brain Transcriptome Reveals Defective Oligodendrocyte Differentiation and Myelination. *Neuron* **89**, 1208-1222 (2016).

17. Lockstone, H. E., Harris, L. W., Swatton, J. E., Wayland, M. T., Holland, A. J. and Bahn, S., Gene expression profiling in the adult Down syndrome brain. *Genomics* **90**, 647-660 (2007).

18. Su, C. H., D, D. and Tarn, W. Y., Alternative Splicing in Neurogenesis and Brain Development. *Front Mol Biosci* **5**, 12 (2018).

19. Raj, B. and Blencowe, B. J., Alternative Splicing in the Mammalian Nervous System: Recent Insights into Mechanisms and Functional Roles. *Neuron* **87**, 14-27 (2015).

20. Fuster, J. M., The prefrontal cortex–an update: time is of the essence. *Neuron* **30**, 319-333 (2001).

21. Tran, H. T. N., Ang, K. S., Chevrier, M., Zhang, X., Lee, N. Y. S., Goh, M. and Chen, J., A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol* **21**, 12 (2020).

22. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., 3rd, Hao, Y., Stoeckius, M., Smibert, P. and Satija, R., Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902 e1821 (2019).

23. Dierssen, M., Down syndrome: the brain in trisomic mode. *Nat Rev Neurosci* **13**, 844-858 (2012).

24. Spektor, R., Yang, J. W., Lee, S. and Soloway, P. D., Single cell ATAC-seq identifies broad changes in neuronal abundance and chromatin accessibility in Down Syndrome. *bioRxiv*, 561191 (2019).

25. Tasic, B., Yao, Z., Graybuck, L. T., Smith, K. A., Nguyen, T. N., Bertagnolli, D., Goldy, J., Garren, E., Economo, M. N., Viswanathan, S., Penn, O., Bakken, T., Menon, V., Miller, J., Fong, O., Hirokawa, K. E., Lathia, K., Rimorin, C., Tieu, M., Larsen, R., Casper, T., Barkan, E., Kroll, M., Parry, S., Shapovalova, N. V., Hirschstein, D., Pendergraft, J., Sullivan, H. A., Kim, T. K., Szafer, A., Dee, N., Groblewski, P., Wickersham, I., Cetin, A., Harris, J. A., Levi, B. P., Sunkin, S. M., Madisen, L., Daigle, T. L., Looger, L., Bernard, A., Phillips, J., Lein, E., Hawrylycz, M., Svoboda, K., Jones, A. R., Koch, C. and Zeng, H., Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563**, 72-78 (2018).

26. Lana-Elola, E., Watson-Scales, S. D., Fisher, E. M. and Tybulewicz, V. L., Down syndrome: searching for the genetic culprits. *Dis Model Mech* **4**, 586-595 (2011).

27. Stamoulis, G., Garieri, M., Makrythanasis, P., Letourneau, A., Guipponi, M., Panousis, N., Sloan-Béna, F., Falconnet, E., Ribaux, P., Borel, C., Santoni, F. and Antonarakis, S. E., Single cell transcriptome in aneuploidies reveals mechanisms of gene dosage imbalance. *Nat Commun* **10**, 4495 (2019).

28. Veitia, R. A., Bottani, S. and Birchler, J. A., Gene dosage effects: nonlinearities, genetic interactions, and dosage compensation. *Trends Genet* **29**, 385-393 (2013).

29. Simmons, A. B., Bloomsburg, S. J., Sukeena, J. M., Miller, C. J., Ortega-Burgos, Y., Borghuis, B. G. and Fuerst, P. G., DSCAM-mediated control of dendritic and axonal arbor outgrowth enforces tiling and inhibits synaptic plasticity. *Proc Natl Acad Sci U S A* **114**, E10224-e10233 (2017).

30. Patzke, C., Max, K. E., Behlke, J., Schreiber, J., Schmidt, H., Dorner, A. A., Kröger, S., Henning, M., Otto, A., Heinemann, U. and Rathjen, F. G., The coxsackievirus-adenovirus receptor reveals complex homophilic and heterophilic interactions on neural cells. *J Neurosci* **30**, 2897-2910 (2010).

31. Hoe, H. S., Lee, K. J., Carney, R. S., Lee, J., Markova, A., Lee, J. Y., Howell, B. W., Hyman, B. T., Pak, D. T., Bu, G. and Rebeck, G. W., Interaction of reelin with amyloid precursor protein promotes neurite outgrowth. *J Neurosci* **29**, 7459-7473 (2009).

32. Sheng, L., Leshchyns'ka, I. and Sytnyk, V., Neural cell adhesion molecule 2 promotes the formation of filopodia and neurite branching by inducing submembrane increases in Ca2+ levels. *J Neurosci* **35**, 1739-1752 (2015).

33. The Gene Ontology Consortium, The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res* **47**, D330-D338 (2019).

34. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. and Sherlock, G., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29 (2000).

35. Mi, H., Muruganujan, A., Ebert, D., Huang, X. and Thomas, P. D., PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res* **47**, D419-d426 (2019).

36. Wilkinson, D. G., Multiple roles of EPH receptors and ephrins in neural development. *Nat Rev Neurosci* **2**, 155-164 (2001).

37. Bashaw, G. J. and Klein, R., Signaling from axon guidance receptors. *Cold Spring Harb Perspect Biol* **2**, a001941 (2010).

38. Morgan, J. I., Cohen, D. R., Hempstead, J. L. and Curran, T., Mapping patterns of c-fos expression in the central nervous system after seizure. *Science* **237**, 192-197 (1987).

39. Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H. A. and Trapnell, C., Reversed graph embedding resolves complex single-cell trajectories. *Nat Methods* **14**, 979-982 (2017).

40. Masuda, T., Sankowski, R., Staszewski, O., Böttcher, C., Amann, L., Sagar, Scheiwe, C., Nessler, S., Kunz, P., van Loo, G., Coenen, V. A., Reinacher, P. C., Michel, A., Sure, U., Gold, R., Grün, D., Priller, J., Stadelmann, C. and Prinz, M., Spatial and temporal heterogeneity of mouse and human microglia at single-cell resolution. *Nature* **566**, 388-392 (2019).

41. Kougioumtzidou, E., Shimizu, T., Hamilton, N. B., Tohyama, K., Sprengel, R., Monyer, H., Attwell, D. and Richardson, W. D., Signalling through AMPA receptors on oligodendrocyte precursors promotes myelination by enhancing oligodendrocyte survival. *Elife* **6**, (2017).

42. Keren-Shaul, H., Spinrad, A., Weiner, A., Matcovitch-Natan, O., Dvir-Szternfeld, R., Ulland, T. K., David, E., Baruch, K., Lara-Astaiso, D., Toth, B., Itzkovitz, S., Colonna, M., Schwartz, M. and Amit, I., A Unique Microglia Type Associated with Restricting Development of Alzheimer's Disease. *Cell* **169**, 1276-1290.e1217 (2017).

43. Stevens, B., Allen, N. J., Vazquez, L. E., Howell, G. R., Christopherson, K. S., Nouri, N., Micheva, K. D., Mehalow, A. K., Huberman, A. D., Stafford, B., Sher, A., Litke, A. M., Lambris, J. D., Smith, S. J., John, S. W. and Barres, B. A., The classical complement cascade mediates CNS synapse elimination. *Cell* **131**, 1164-1178 (2007).

44. Wang, C., Yue, H., Hu, Z., Shen, Y., Ma, J., Li, J., Wang, X. D., Wang, L., Sun, B., Shi, P., Wang, L. and Gu, Y., Microglia mediate forgetting via complement-dependent synaptic elimination. *Science* **367**, 688-694 (2020).

45. Li, Tao, Chiou, Brian, Gilman, Casey K, Luo, Rong, Koshi, Tatsuhiro, Yu, Diankun, Oak, Hayeon C, Giera, Stefanie, Johnson-Venkatesh, Erin, Muthukumar, Allie K., Stevens, Beth, Umemori, Hisashi and Piao, Xianhua, A splicing isoform of GPR56 mediates microglial synaptic refinement via phosphatidylserine binding. *bioRxiv*, 2020.2004.2024.059840 (2020).

46. Badimon, A., Strasburger, H. J., Ayata, P., Chen, X., Nair, A., Ikegami, A., Hwang, P., Chan, A. T., Graves, S. M., Uweru, J. O., Ledderose, C., Kutlu, M. G., Wheeler, M. A., Kahan, A., Ishikawa, M., Wang, Y. C., Loh, Y. E., Jiang, J. X., Surmeier, D. J., Robson, S. C., Junger, W. G., Sebra, R., Calipari, E. S., Kenny, P. J., Eyo, U. B., Colonna, M., Quintana, F. J., Wake, H., Gradinaru, V. and Schaefer, A., Negative feedback control of neuronal activity by microglia. *Nature* **586**, (2020).

47. Pinto, B., Morelli, G., Rastogi, M., Savardi, A., Fumagalli, A., Petretto, A., Bartolucci, M., Varea, E., Catelani, T., Contestabile, A., Perlini, L. E. and Cancedda, L., Rescuing Over-activated Microglia Restores Cognitive Performance in Juvenile Animals of the Dp(16) Mouse Model of Down Syndrome. *Neuron* **108**, (2020).

48. Gribble, S. M., Wiseman, F. K., Clayton, S., Prigmore, E., Langley, E., Yang, F., Maguire, S., Fu, B., Rajan, D., Sheppard, O., Scott, C., Hauser, H., Stephens, P. J., Stebbings, L. A., Ng, B. L., Fitzgerald, T., Quail, M. A., Banerjee, R., Rothkamm, K., Tybulewicz, V. L., Fisher, E. M. and Carter, N. P., Massively parallel sequencing reveals the complex structure of an irradiated human chromosome on a mouse background in the Tc1 model of Down syndrome. *PLoS One* **8**, e60482 (2013).

49. Li, Z., Yu, T., Morishima, M., Pao, A., LaDuca, J., Conroy, J., Nowak, N., Matsui, S., Shiraishi, I. and Yu, Y. E., Duplication of the entire 22.9 Mb human chromosome 21 syntenic region on mouse chromosome 16 causes cardiovascular and gastrointestinal abnormalities. *Hum Mol Genet* **16**, 1359-1366 (2007).

50. Olson, L. E., Richtsmeier, J. T., Leszl, J. and Reeves, R. H., A chromosome 21 critical region does not cause specific Down syndrome phenotypes. *Science* **306**, 687-690 (2004).

51. Wehrspaun, C. C., Haerty, W. and Ponting, C. P., Microglia recapitulate a hematopoietic master regulator network in the aging human frontal cortex. *Neurobiol Aging* **36**, 2443.e2449-2443.e2420 (2015).

52. Logan, T. T., Villapol, S. and Symes, A. J., TGF-β superfamily gene expression and induction of the Runx1 transcription factor in adult neurogenic regions after brain injury. *PLoS One* **8**, e59250 (2013).

53. Malko, P., Syed Mortadza, S. A., McWilliam, J. and Jiang, L. H., TRPM2 Channel in Microglia as a New Player in Neuroinflammation Associated With a Spectrum of Central Nervous System Pathologies. *Front Pharmacol* **10**, 239 (2019).

54. So, A. Y., Garcia-Flores, Y., Minisandram, A., Martin, A., Taganov, K., Boldin, M. and Baltimore, D., Regulation of APC development, immune response, and autoimmunity by Bach1/HO-1 pathway in mice. *Blood* **120**, 2428-2437 (2012).

55. Zhang, X., Guo, J., Wei, X., Niu, C., Jia, M., Li, Q. and Meng, D., Bach1: Function, Regulation, and Involvement in Disease. *Oxid Med Cell Longev* **2018**, 1347969 (2018).

56. Gupta, I., Collier, P. G., Haase, B., Mahfouz, A., Joglekar, A., Floyd, T., Koopmans, F., Barres, B., Smit, A. B., Sloan, S. A., Luo, W., Fedrigo, O., Ross, M. E. and Tilgner, H. U., Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat Biotechnol* **36**, 1197-1202 (2018).

57. Joglekar, A., Prjibelski, A., Mahfouz, A., Collier, P., Lin, S., Schlusche, A. K., Marrocco, J., Williams, S. R., Haase, B., Hayes, A., Chew, J. G., Weisenfeld, N. I., Wong, M. Y., Stein, A. N., Hardwick, S. A., Hunt, T., Wang, Q., Dieterich, C., Bent, Z., Fedrigo, O., Sloan, S. A., Risso, D., Jarvis, E. D., Flicek, P., Luo, W., Pitt, G. S., Frankish, A., Smit, A. B., Ross, M. E. and Tilgner, H. U., A spatially resolved brain region- and cell type-specific isoform atlas of the postnatal mouse brain. *Nat Commun* **12**, 463 (2021).

58. Tardaguila, M., de la Fuente, L., Marti, C., Pereira, C., Pardo-Palacios, F. J., Del Risco, H., Ferrell, M., Mellado, M., Macchietto, M., Verheggen, K., Edelmann, M., Ezkurdia, I., Vazquez, J., Tress, M., Mortazavi, A., Martens, L., Rodriguez-Navarro, S., Moreno-Manzano, V. and Conesa, A., SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res* **28**, 396-411 (2018).

59. Wyman, Dana, Balderrama-Gutierrez, Gabriela, Reese, Fairlie, Jiang, Shan, Rahmanian, Sorena, Forner, Stefania, Matheos, Dina, Zeng, Weihua, Williams, Brian, Trout, Diane, England, Whitney, Chu, Shu-Hui, Spitale, Robert C., Tenner, Andrea J., Wold, Barbara J. and Mortazavi, Ali, A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. *bioRxiv*, 672931 (2020).

60. Rinaldi, C. and Wood, M. J. A., Antisense oligonucleotides: the next frontier for treatment of neurological disorders. *Nat Rev Neurol* **14**, 9-21 (2018).

61. Deverman, B. E., Ravina, B. M., Bankiewicz, K. S., Paul, S. M. and Sah, D. W. Y., Gene therapy for neurological disorders: progress and prospects. *Nat Rev Drug Discov* **17**, 641-659 (2018).

62. de la Fuente, L., Arzalluz-Luque, A., Tardaguila, M., Del Risco, H., Marti, C., Tarazona, S., Salguero, P., Scott, R., Lerma, A., Alastrue-Agudo, A., Bonilla, P., Newman, J. R. B., Kosugi, S., McIntyre, L. M., Moreno-Manzano, V. and Conesa, A., tappAS: a comprehensive computational framework for the analysis of the functional impact of differential splicing. *Genome Biol* **21**, 119 (2020).

63. Harold, D., Abraham, R., Hollingworth, P., Sims, R., Gerrish, A., Hamshere, M. L., Pahwa, J. S., Moskvina, V., Dowzell, K., Williams, A., Jones, N., Thomas, C., Stretton, A., Morgan, A. R., Lovestone, S., Powell, J., Proitsi, P., Lupton, M. K., Brayne, C., Rubinsztein, D. C., Gill, M., Lawlor, B., Lynch, A., Morgan, K., Brown, K. S., Passmore, P. A., Craig, D., McGuinness, B., Todd, S., Holmes, C., Mann, D., Smith, A. D., Love, S., Kehoe, P. G., Hardy, J., Mead, S., Fox, N., Rossor, M., Collinge, J., Maier, W., Jessen, F., Schürmann, B., Heun, R., van den Bussche, H., Heuser, I., Kornhuber, J., Wiltfang, J., Dichgans, M., Frölich, L., Hampel, H., Hüll, M., Rujescu, D., Goate, A. M., Kauwe, J. S., Cruchaga, C., Nowotny, P., Morris, J. C., Mayo, K., Sleegers, K., Bettens, K., Engelborghs, S., De Deyn, P. P., Van Broeckhoven, C., Livingston, G., Bass, N. J., Gurling, H., McQuillin, A., Gwilliam, R., Deloukas, P., Al-Chalabi, A., Shaw, C. E., Tsolaki, M., Singleton, A. B., Guerreiro, R., Mühleisen, T. W., Nöthen, M. M., Moebus, S., Jöckel, K. H., Klopp, N., Wichmann, H. E., Carrasquillo, M. M., Pankratz, V. S., Younkin, S. G., Holmans, P. A., O'Donovan, M., Owen, M. J. and Williams, J., Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nat Genet* **41**, 1088-1093 (2009).

64. Lambert, J. C., Heath, S., Even, G., Campion, D., Sleegers, K., Hiltunen, M., Combarros, O., Zelenika, D., Bullido, M. J., Tavernier, B., Letenneur, L., Bettens, K., Berr, C., Pasquier, F., Fiévet, N., Barberger-Gateau, P., Engelborghs, S., De Deyn, P., Mateo, I., Franck, A., Helisalmi, S., Porcellini, E., Hanon, O., de Pancorbo, M. M., Lendon, C., Dufouil, C., Jaillard, C., Leveillard, T., Alvarez, V., Bosco, P., Mancuso, M., Panza, F., Nacmias, B., Bossù, P., Piccardi, P., Annoni, G., Seripa, D., Galimberti, D., Hannequin, D., Licastro, F., Soininen, H., Ritchie, K., Blanché, H., Dartigues, J. F., Tzourio, C., Gut, I., Van Broeckhoven, C., Alpérovitch, A., Lathrop, M. and Amouyel, P., Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. *Nat Genet* **41**, 1094-1099 (2009).

65. Chen, G. F., Xu, T. H., Yan, Y., Zhou, Y. R., Jiang, Y., Melcher, K. and Xu, H. E., Amyloid beta: structure, biology and structure-based therapeutic development. *Acta Pharmacol Sin* **38**, 1205-1235 (2017).

66. Rohan de Silva, H. A., Jen, A., Wickenden, C., Jen, L. S., Wilkinson, S. L. and Patel, A. J., Cell-specific expression of beta-amyloid precursor protein isoform mRNAs and proteins in neurons and astrocytes. *Brain Res Mol Brain Res* **47**, 147-156 (1997).

67. Keren-Shaul, H., Spinrad, A., Weiner, A., Matcovitch-Natan, O., Dvir-Szternfeld, R., Ulland, T. K., David, E., Baruch, K., Lara-Astaiso, D., Toth, B., Itzkovitz, S., Colonna, M., Schwartz, M. and Amit, I., A Unique Microglia Type Associated with Restricting Development of Alzheimer's Disease. *Cell* **169**, 1276-1290 e1217 (2017).

68. Mathys, H., Adaikkan, C., Gao, F., Young, J. Z., Manet, E., Hemberg, M., De Jager, P. L., Ransohoff, R. M., Regev, A. and Tsai, L. H., Temporal Tracking of Microglia Activation in Neurodegeneration at Single-Cell Resolution. *Cell Rep* **21**, 366-380 (2017).

69. Hickman, S., Izzy, S., Sen, P., Morsett, L. and El Khoury, J., Microglia in neurodegeneration. *Nat Neurosci* **21**, 1359-1369 (2018).

70. Miyoshi, G., Hjerling-Leffler, J., Karayannis, T., Sousa, V. H., Butt, S. J., Battiste, J., Johnson, J. E., Machold, R. P. and Fishell, G., Genetic fate mapping reveals that the caudal ganglionic eminence produces a large and diverse population of superficial cortical interneurons. *J Neurosci* **30**, 1582-1594 (2010).

71. Hansen, D. V., Lui, J. H., Flandin, P., Yoshikawa, K., Rubenstein, J. L., Alvarez-Buylla, A. and Kriegstein, A. R., Non-epithelial stem cells and cortical interneuron production in the human ganglionic eminences. *Nat Neurosci* **16**, 1576-1587 (2013).

72. Kleschevnikov, A. M., Belichenko, P. V., Villar, A. J., Epstein, C. J., Malenka, R. C. and Mobley, W. C., Hippocampal long-term potentiation suppressed by increased inhibition in the Ts65Dn mouse, a genetic model of Down syndrome. *J Neurosci* **24**, 8153-8160 (2004).

73. Siarey, R. J., Carlson, E. J., Epstein, C. J., Balbo, A., Rapoport, S. I. and Galdzicki, Z., Increased synaptic depression in the Ts65Dn mouse, a model for mental retardation in Down syndrome. *Neuropharmacology* **38**, 1917-1920 (1999).

74. Goldberg-Stern, H., Strawsburg, R. H., Patterson, B., Hickey, F., Bare, M., Gadoth, N. and Degrauw, T. J., Seizure frequency and characteristics in children with Down syndrome. *Brain & development* **23**, 375-378 (2001).

75. Scotti, M. M. and Swanson, M. S., RNA mis-splicing in disease. *Nat Rev Genet* **17**, 19-32 (2016).

76. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res* **47**, D330-d338 (2019).

Acknowledgments

General: We thank Richard R. Rivera, Danielle Jones, and Gwendolyn E. Kaeser for their discussions and key input; Laura Wolszon for her efforts to source and obtain human specimens; and Brian James and Kang Liu at the Sanford Burnham Prebys Medical Discovery Institute Genomics Core for RIN analysis of brain samples. Brain specimens were obtained from the University of Maryland Brain and Tissue Bank, the Goizueta Alzheimer's Disease Research Center (ADRC) at Emory University, the London Neurodegenerative Diseases Brain Bank, the Newcastle Brain Tissue Resource, and the Southwest Dementia Brain Bank for human brain specimens. We also thank the donors and families who shared these precious brain materials.

Funding: Research reported in this publication was supported by the National Institute on Aging of the National Institutes of Health under award numbers R56AG073965, R01AG065541, R01AG071465 (J.C.), as well as the National Institute of General Medical Sciences through the UCSD Graduate Training Program in Cellular and Molecular Pharmacology institutional training grant under award number T32 GM007752 (awarded to C.S.L.). This work was also supported by non-Federal funds from The Shaffer Family Foundation and The Bruce Ford & Anne Smith Bundy Foundation (J.C.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions: **C.R.P.**: Conceptualization, Investigation, Validation, Formal analysis, Visualization, Writing – Original Draft Preparation, Review and Editing. **C.S.L.**: Conceptualization, Investigation, Software, Formal Analysis, Visualization, Data Curation, Writing – Original Draft Preparation, Review and Editing. **W.J.R.**: Investigation, Writing – Review and Editing. **M.-H.L.**: Investigation. **J.C.**: Conceptualization, Writing – Review and Editing, Supervision, Funding acquisition.

Competing interests: None.

Data and materials availability: Fastq files for Illumina reads and bam files for PacBio reads

will be available upon publication through the European Genome-Phenome Archive (EGA). Isoforms identified by long-read sequencing are available through the UCSC Genome browser:

Untargeted dataset: genome.ucsc.edu/s/csl022/DSND_snIsoSeq_sample

SPP1 targeted dataset: genome.ucsc.edu/s/csl022/SPP1_scIsoSeq

APP targeted dataset: genome.ucsc.edu/s/csl022/APP_scIsoSeq

BIN1 targeted dataset: genome.ucsc.edu/s/csl022/BIN1_snIsoSeq

Chapter 4, in full, is a reprint of the material as it appears in *PNAS* 2021. Palmer, C.R.*, Liu, C.S.*, Romanow, W.J., Lee, M-H., Chun, J. Altered cell and RNA isoform diversity in aging Down syndrome brains revealed by snRNA-seq. The dissertation author was a co-primary researcher and author of this paper.

CHAPTER 5

IDENTIFYING NOVEL ISOFORM FEATURES THROUGH MODIFICATION OF SQANTI3

Advances in long-read sequencing have made it possible to examine the transcriptome at isoform-level resolution, and bioinformatics efforts to support analysis of these datasets have also kept pace. Numerous software packages have been written to process the data from raw data to isoforms and their corresponding structural characteristics, abundances, etc. These tools are used perform quality control (QC) on the sequencing reads to remove potential artifacts from the library preparation and sequencing processes, align reads to the genome/transcriptome, and compare the sequences to the reference annotation. While many of the QC tools are technology-specific to better account for artifacts/errors that each technology is more prone to, downstream analysis after read alignment is often technology-agnostic.

A highly-cited software package from the Conesa lab, called SQANTI3, is a technology-agnostic tool that uses a gff/gtf file of mapped, non-redundant transcripts as an input and outputs numerous files detailing the characteristics of transcripts (gene, structural category, percent of A's in the downstream sequence, junction canonical status, etc) (*1*). These metrics can then be used to filter out potential artifacts, resulting in a confident set of isoforms that were detected in the sample. Read counts can be provided as an input as well, making it possible to calculate the relative abundances of each of the detected isoforms. Its greatest utility is its ability to compare the isoforms to the reference annotation, making it possible to identify potentially undiscovered, novel isoforms.

SQANTI3 categorizes isoforms relative to the reference annotation, labeling isoforms as belonging to one of four major groups: full-splice match (FSM), incomplete-splice match (ISM), novel in catalog (NIC) and novel not in catalog (NNC) (Fig 1A), in addition to a few smaller categories: antisense, fusion, genic, and intergenic. An FSM isoform contains identical junctions and splice sites as a known isoform. An ISM isoform is a truncated form of a known isoform, and the junctions that are present match those of that isoform. NIC and NNC isoforms are novel forms of annotated genes that can contain novel combinations of known splice sites (NIC) or include non-annotated splice sites (NNC). These larger categories can be further broken down into small subcategories with additional detail about how they compare to reference

isoforms. FSM isoforms can be truncated on the 5' or 3' ends without affecting the splice junctions. ISM isoforms likewise may be truncated on the 5' or 3' end or may even represent an internal fragment. NIC isoforms can result from a combination of known splice sites/junctions or the absence of a junction through intron retention. NNC isoforms in particular have various features that differentiate them from the isoforms in the reference annotation. Identification and annotation of these features are not part of the current version of SQANTI3. We made some fully-integrated modifications to SQANTI3 in order to characterize these features as part of its normal implementation.

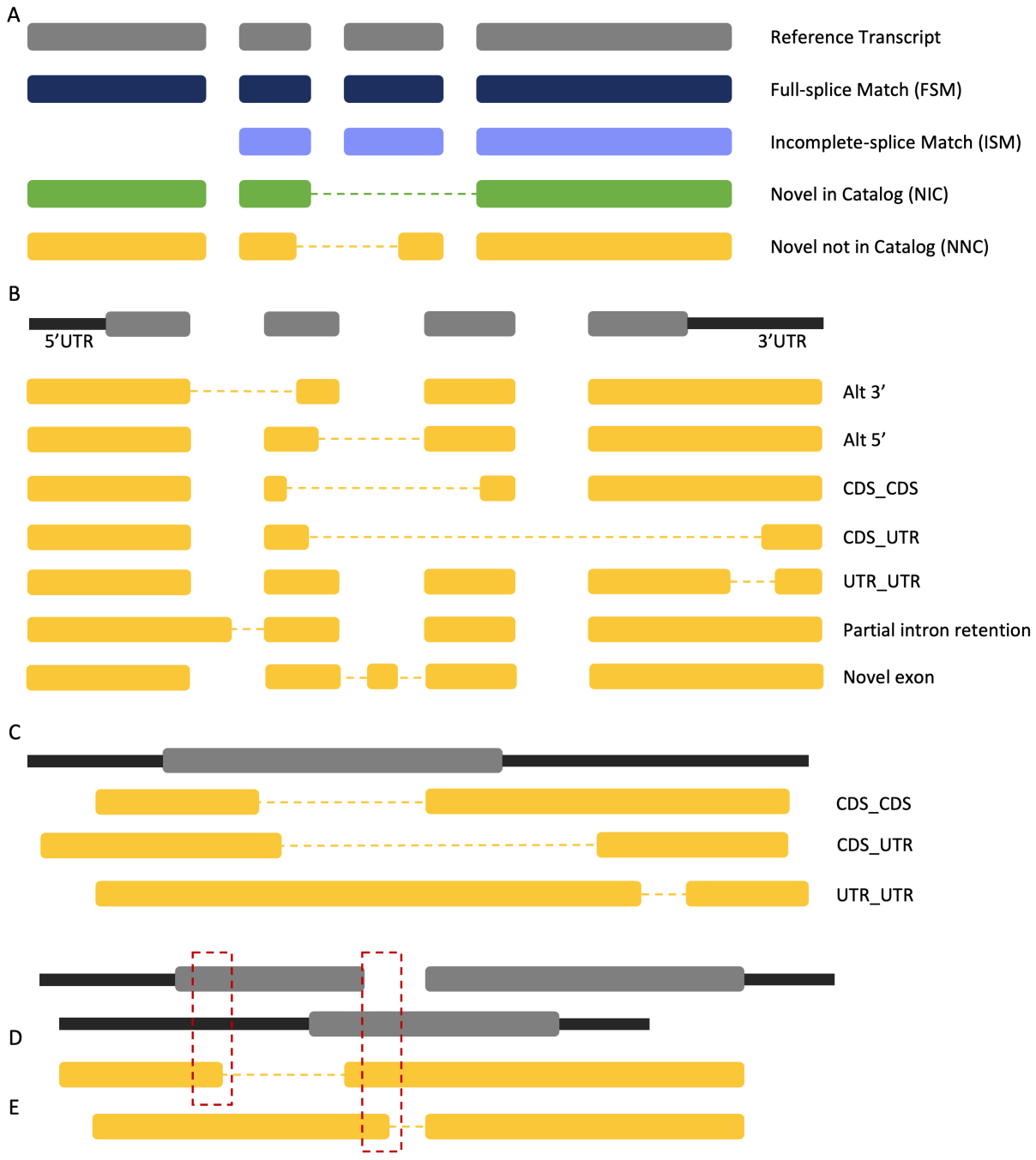

Figure 5.1. SQANTI3 categories and NNC features.
(**A**) The four main SQANTI3 isoform categories. Dashed lines indicate novel junctions that differ from the reference annotation transcript. (**B**) Seven features of NNC isoforms that create a novel junction (or two). Multiple features can be observed in a single transcript, but only one is necessary to assign the isoform to the NNC category. (**C**) New categorization of spliced transcripts that map to a gene with a single exon. (**D**) Example gene with multiple isoforms and a novel transcript with a donor site that overlaps a coordinate overlapping both a coding exon and UTR in different reference isoforms.

The modifications to SQANTI3 only involve the part of the tool that determines which category an isoform belongs to. Once an isoform is determined to diverge from the known set in the reference, its junctions and splice sites are examined more closely to determine whether it belongs in the NIC or NNC category. In our edited script, additional features of the NNC isoforms are identified and noted in the "subcategory" section of the classification output file. Each of these features on its own can make an isoform NNC, or several can be present in a single isoform. These seven features are: Alt 3' junction, Alt 5' junction, CDS_CDS junction, CDS_UTR junction, UTR_UTR junction, partial intron retention, and novel exon (Fig 1B). Alt 3', Alt 5', CDS_CDS, CDS_UTR, and UTR_UTR junctions all increase the size of junctions, and new splice sites are generated within known exons, while partial intron retentions and novel exons decrease and split junctions respectively with new splice sites in annotated introns. Alt 3' junctions result from a new acceptor splice site on the 3' end of the junction that truncates the "acceptor" exon; the 5' donor splice site is unchanged and matches a known splice site. Alt 5' junctions follow the same pattern as Alt 3' junctions, but the novel splice site is the 5' donor. CDS_CDS junctions are created from two new splice sites within the coding sequence. These exons do not need to be adjacent, and both novel splice sites can occur in the same exon. CDS_UTR junctions have two new splice sites within exons where one splice site is in the coding sequence and one splice site is in a UTR (on either end). Similarly to the CDS_CDS junction, the exons do not need to be adjacent, and the junction can occur within a single exon. UTR_UTR junctions have two novel splice sites that are both within a UTR; they can occur within the same UTR (5' or 3') or different ones (5' and 3'). The extension of an exon into previously annotated intronic space by creation of a new splice site in intronic sequence describes partial intron retention. Novel exons create two new splice sites in intronic space by including sequence that does not overlap any known exon (Fig 1B).

An additional small modification was made with regards to how SQANTI3 handles single-exon genes. Originally, SQANTI3 would label an isoform with two exons that mapped to a single-exon gene, as intergenic. This splice pattern matches a CDS_CDS, CDS_UTR, or

UTR_UTR junction, and modifications were made to the code to label these isoforms in this manner (Fig 1C).

Many genes have more than one transcript, which can complicate the assignment of whether a novel splice site occurs in the CDS or UTR. While many transcripts often share exons and differ by which exons are present in combination, there are genes with vastly differing transcripts because of alternate start codons or sequence that can be spliced in or out in different isoforms. In the case where a novel splice site in a transcript overlaps a coordinate annotated as CDS in one transcript and UTR in the other, priority is given to labeling that site as being in CDS (Fig 1D). Likewise, for a coordinate that is part of an exon in one isoform and part of an intron in another, priority is given to labeling that site as being in CDS (Fig 1E).

The ability to identify the features that make an isoform novel relative to what has been reported has many potential applications. Identifying features commonly present in artifacts generated through sample preparation and sequencing could improve filtering and quality control. It is known that "RT-switching" during the reverse transcription process for converting RNA to cDNA can create "new" splice sites (*2, 3*). Additional details about what these splice sites look like can better inform researchers of potential artifacts that may not actually have biological significance. One way to evaluate this would be to compare isoforms from cDNA sequencing and direct-RNA sequencing using Oxford Nanopore Technologies' (ONT) approach for directly sequencing RNA without the reverse transcription step. NNC features that are more prominent in cDNA libraries may be indicative of reverse transcription artifacts.

Many diseases are associated with altered splicing whether through observation of uncommon isoforms or through mutations in splicing factors/regulators. Full-length transcriptome analyses can highlight which types of novel isoforms are more prevalent in disease states. Amyotrophic lateral sclerosis (ALS) is on example of a disease that could be studied this way. Several studies have linked ALS to splicing disruption because many ALS-associated inherited mutations occur in splicing factors (*4-7*). Others have observed isoforms with altered splicing resulting from mutations that alter the splice sites that are recognized by the spliceosome (*8, 9*). These

novel features also have the potential to affect gene expression and protein translation. Altered splice sites could cause frame shifts that lead to increased nonsense-mediated decay or translation of novel proteins. This opens up the possibility for NNC isoform features to contribute to altered cellular function in disease.

This work was inspired by a previous publication from our lab (see Chapter 4). While examining the isoforms that we detected in our Iso-Seq samples, we identified the seven NNC features as being comprehensive of the different variations that could result in the isoforms being labeled as NNC. The modifications to SQANTI3 were made to automate this process and characterize novel isoforms on a greater scale.

References

1. Tardaguila, M., de la Fuente, L., Marti, C., Pereira, C., Pardo-Palacios, F. J., Del Risco, H., Ferrell, M., Mellado, M., Macchietto, M., Verheggen, K., Edelmann, M., Ezkurdia, I., Vazquez, J., Tress, M., Mortazavi, A., Martens, L., Rodriguez-Navarro, S., Moreno-Manzano, V. and Conesa, A., SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res* **28**, 396-411 (2018).

2. Cocquet, J., Chong, A., Zhang, G. and Veitia, R. A., Reverse transcriptase template switching and false alternative transcripts. *Genomics* **88**, 127-131 (2006).

3. Houseley, J. and Tollervey, D., Apparent non-canonical trans-splicing is generated by reverse transcriptase in vitro. *PLoS One* **5**, e12271 (2010).

4. Arnold, E. S., Ling, S. C., Huelga, S. C., Lagier-Tourenne, C., Polymenidou, M., Ditsworth, D., Kordasiewicz, H. B., McAlonis-Downes, M., Platoshyn, O., Parone, P. A., Da Cruz, S., Clutario, K. M., Swing, D., Tessarollo, L., Marsala, M., Shaw, C. E., Yeo, G. W. and Cleveland, D. W., ALS-linked TDP-43 mutations produce aberrant RNA splicing and adult-onset motor neuron disease without aggregation or loss of nuclear TDP-43. *Proc Natl Acad Sci U S A* **110**, E736-745 (2013).

5. Butti, Z. and Patten, S. A., RNA Dysregulation in Amyotrophic Lateral Sclerosis. *Front Genet* **9**, 712 (2018).

6. La Cognata, V., Gentile, G., Aronica, E. and Cavallaro, S., Splicing Players Are Differently Expressed in Sporadic Amyotrophic Lateral Sclerosis Molecular Clusters and Brain Regions. *Cells* **9**, (2020).

7. Perrone, B., La Cognata, V., Sprovieri, T., Ungaro, C., Conforti, F. L., Ando, S. and Cavallaro, S., Alternative Splicing of ALS Genes: Misregulation and Potential Therapies. *Cell Mol Neurobiol* **40**, 1-14 (2020).

8. Brown, A. L., Wilkins, O. G., Keuss, M. J., Hill, S. E., Zanovello, M., Lee, W. C., Bampton, A., Lee, F. C. Y., Masino, L., Qi, Y. A., Bryce-Smith, S., Gatt, A., Hallegger, M., Fagegaltier, D., Phatnani, H., Consortium, Nygc Als, Newcombe, J., Gustavsson, E. K., Seddighi, S., Reyes, J. F., Coon, S. L., Ramos, D., Schiavo, G., Fisher, E. M. C., Raj, T., Secrier, M., Lashley, T., Ule, J., Buratti, E., Humphrey, J., Ward, M. E. and Fratta, P., TDP-43 loss and ALS-risk SNPs drive mis-splicing and depletion of UNC13A. *Nature* **603**, 131-137 (2022).

9. Ma, X. R., Prudencio, M., Koike, Y., Vatsavayai, S. C., Kim, G., Harbinski, F., Briner, A., Rodriguez, C. M., Guo, C., Akiyama, T., Schmidt, H. B., Cummings, B. B., Wyatt, D. W., Kurylo, K., Miller, G., Mekhoubad, S., Sallee, N., Mekonnen, G., Ganser, L., Rubien, J. D., Jansen-West, K., Cook, C. N., Pickles, S., Oskarsson, B., Graff-Radford, N. R., Boeve, B. F., Knopman, D. S., Petersen, R. C., Dickson, D. W., Shorter, J., Myong, S., Green, E. M., Seeley, W. W., Petrucelli, L. and Gitler, A. D., TDP-43 represses cryptic exon inclusion in the FTD-ALS gene UNC13A. *Nature* **603**, 124-130 (2022).

CHAPTER 6

ISOSEQL: COMPARING LONG-READ ISOFORMS ACROSS MULTIPLE DATASETS

Background

Advances in long-read sequencing technology have made it possible to examine the transcriptome at isoform-resolution. Short-read RNA sequencing technologies are not able to capture entire isoforms and can only be used to infer exon combinations from reads spanning exon-exon junctions. The long reads obtained through either Oxford Nanopore Technologies (ONT) or PacBio sequencing technologies readily exceed the length of the average human mRNA of approximately 3kb, making it possible to examine the combination of exons that are expressed together in a single transcript (*1, 2*). Bioinformatics pipelines and tools have been developed to characterize these isoforms and examine abundance, structure, and protein-coding potential among other properties. Isoform diversity has been explored in several organisms, including humans, mice, and various other animals and plants (*3-11*). Novel isoforms are identified in many studies, indicating that there is a whole set of isoforms that have not been captured and annotated up to this point. This highlights the potential for long-read isoform profiling to help identify new transcripts that could code for unknown proteins with different functions.

One tool in particular, SQANTI3, provides extensive quality control and structural feature characterization of sequenced transcripts (*12*). SQANTI3 is widely used and compatible with both ONT and PacBio sequencing datasets. To expand upon the utility of SQANTI3, we present isoSeQL, a new tool for comparing isoform profiles across multiple datasets that is intended for use with SQANTI3 output files. We demonstrate its usage by comparing Iso-Seq datasets from twelve samples from the Human Genome Structural Variation Consortium phase 3.

Results and discussion

A significant challenge of working with long-read isoform data is comparing across several different samples. Currently analysis pipelines like SQANTI3 are better suited to handling individual samples, and during the classification process, novel isoform IDs (eg. PB.1111.22) are assigned to isoforms in the sample. These isoform IDs are not standardized across analyses, and isoform PBXX.Y in one sample is not the same isoform as PBXX.Y in another sample that was analyzed separately. A few suggested solutions exist, but each has limitations. One solution,

"chaining samples" with cDNA_Cupcake, is only recommended for two to four samples and has known redundancy issues (*13*). A second approach is to utilize TAMA Merge, a script from another isoform classification tool, but this method does not keep track of isoform abundances (*14*). Another solution, using TALON, re-categorizes the isoforms and removes all classification details associated with each isoform from SQANTI3 (*15*). A final solution, used in a previous publication from our group, is to combine all the sample reads into a single "mega-sample" after labeling reads to indicate their sample of origin but prior to alignment, collapsing, and annotation (*8, 9*). This approach allows comparison across samples by using the same isoform IDs, however, it is severely limited in the number of samples that can be studied and by the computing power/memory required.

In order to examine isoform diversity across many samples in a manner that is compatible with the classification and filtering steps from SQANTI3, we developed a software package for comparing isoform characterizations, isoSeQL. isoSeQL uses two output files from SQANTI3 and two user-supplied files with sample information to create a SQLite database (Fig. 1a). The SQLite database is made up of several tables to keep track of isoform characteristics (chromosome, strand, gene, number of exons, SQANTI3 category, etc), read counts, experiment parameters (software versions, date of sequencing run, etc), and sample information (sample ID, tissue, age, etc). All of this information can be queried and used to compare samples of interest.

One unique feature of the database is the ability to group isoforms that only differ very slightly by the start/end coordinates. A challenge of third generation sequencing is generating long fragments of cDNA/RNA to sequence. The length of these fragments is limited by polymerase processivity and stability of the cDNA or RNA (*3, 16*). It is not a trivial task to determine whether reads originate from different isoforms if they only differ by a few hundred nucleotides on the 5' or 3' end. cDNA_Cupcake's collapse_isoforms processing step tries to account for this by collapsing reads that otherwise have the same junctions but differ by fewer than 1000bp on the 5' end and 100bp on the 3' end (*13*). When comparing across several samples, it may not always be clear if certain isoforms would have been collapsed together or not had they been processed

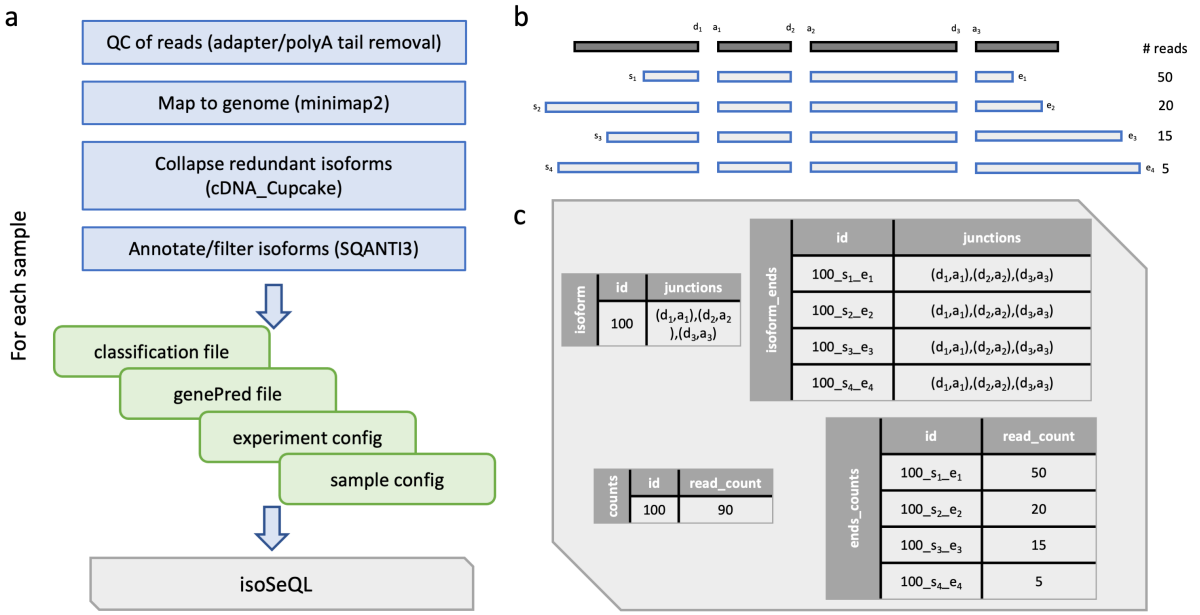

Figure 6.1. isoSeQL workflow and handling of isoforms with common junctions.
(**a**) Generalized workflow for processing long-read isoform sequencing data from raw data to addition to database. Every sample is processed separately and added to the database using two SQANTI3 output files and two user-supplied config files with sample and experiment information. (**b**) Schematic of a reference isoform and many transcripts that map to it. Each transcript differs from the annotation on the 5' or 3' ends. (**c**) Simplified tables in the database demonstrating how read counts for common junction isoforms and isoforms with variable ends are recorded.

all together. In order to emphasize which sequences the samples have in common, the user can decide to analyze isoforms with common junctions, ignoring the end variability completely. This effectively combines all isoforms with the exact same exons and splice junctions (donor/acceptor sites) together and defines that as a single "common junction" isoform that can be identified across multiple samples of interest.

Isoform end variability can, however, be indicative of different transcript start sites (TSS) and transcript termination sites (TTS). Without CAGE peak or additional orthogonal data indicating evidence for start/end sites, it is difficult to interpret whether the ends of the observed isoforms are real or a result of fragmentation (*17*). Graphs showing the spread of the end coordinates can be generated to visualize read counts for each TSS/TTS.

Several built-in functions create plots and tables that illustrate some of the commonly

explored metrics in RNA isoform analysis. The outputs of these functions were demonstrated on Iso-Seq samples from the Human Genome Structural Variation Consortium phase 3. Plots showing the proportion of isoforms or reads belonging to each structural category can be generated with a single command (Fig. 2a). Isoforms in each sample can be visualized in the UCSC Genome Browser or Integrative Genomics Viewer (IGV) by creating bed files color-coded by isoform category (Fig. 2b). Input files to tappAS (read count matrix and transcriptome gff) can be generated for downstream analysis of differential isoform usage/expression (*18*). Gene-level comparisons can be visualized to investigate the distribution of reads corresponding to different isoforms of a specific gene (Fig. 2c). UpSet plots can be made to show how many isoforms were found to be in common or unique amongst several samples (Fig. 2d). For more complicated or study-specific questions, the database can be loaded in Python, and custom queries can be used to incorporate additional filtering or sample grouping.

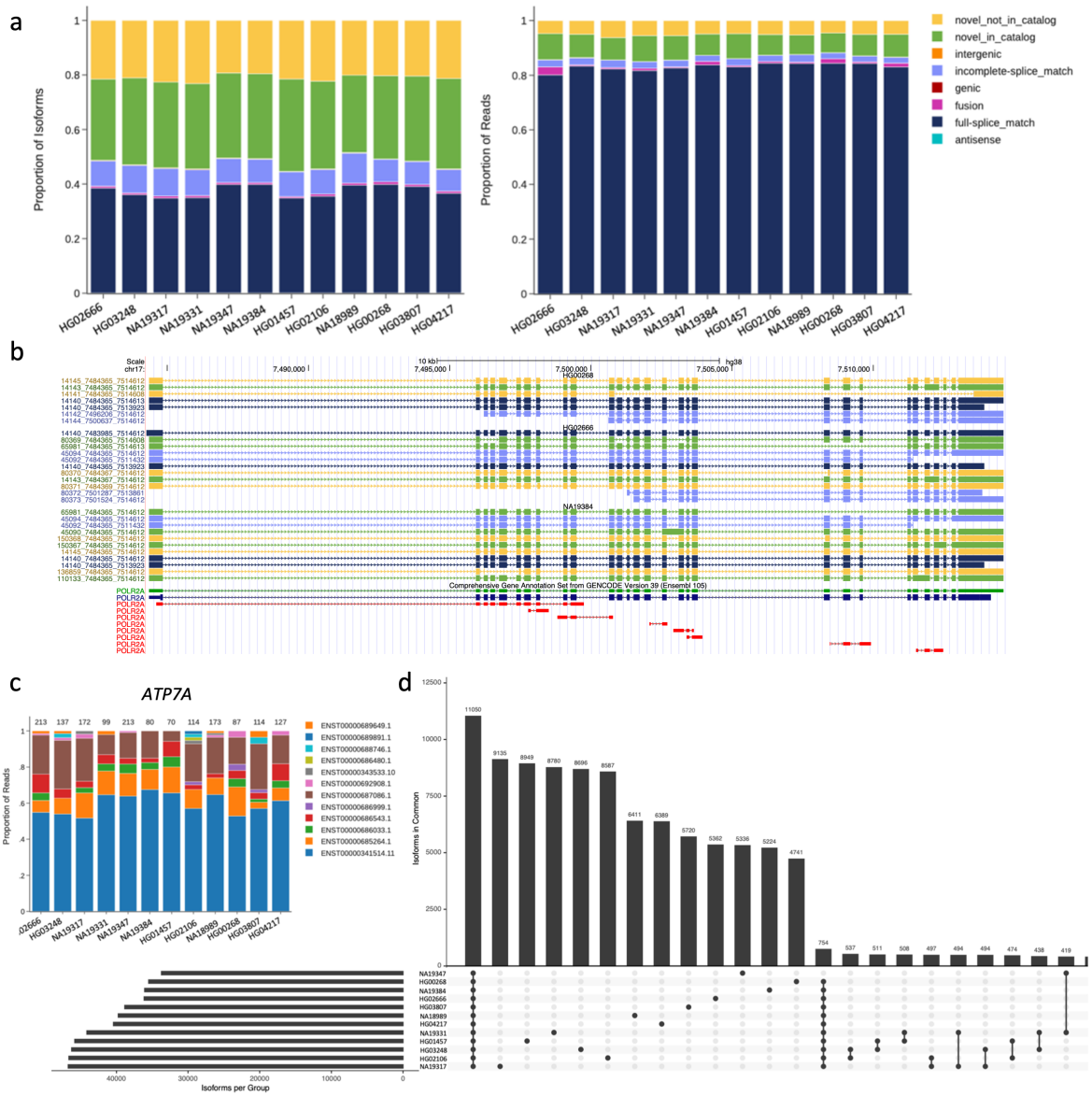

Figure 6.2. Plots generated through isoSeQL's built-in functions.
(**a**) Plots of structural category proportion of isoforms (left) and reads (right). (**b**) UCSC genome browser track visualization. (**c**) Plot showing the proportion of reads from each annotated isoform of *ATP7A*. (**d**) UpSet plot showing how many isoforms with common junctions are shared between samples.

Conclusion

This report describes isoSeQL, the first program for comparing full-length isoform profiles resulting from SQANTI3 analysis. Several studies prior to this have relied on methods with limitations in order to compare across samples or have previously only reported isoforms that come from a very limited sample size. Future iterations of isoSeQL will include additional built-in features and improvements to the current implementation. Currently isoSeQL only tracks long-read abundances (ie read counts), but many people supply matching short-read sequencing data that can be mapped to the long-read-generated transcriptome for junction confirmation and higher-depth estimates of relative abundance. Incorporating short-read counts will not only add further validation of novel junctions but also allow correlation between short- and long-read expression estimates. Another addition will be the ability to group together samples for comparisons with multiple replicates. This would also be helpful to combine multiple runs of the same sample that were intended to increase the sequencing depth and capture rare isoforms. Another important enhancement of the current implementation will be to make it compatible with single-cell long-read isoform data. Several studies have already shown that cell types express different isoforms, and new advances in long-read methods (MAS-seq) have increased the throughput of single-cell long-read sequencing, making this application particularly relevant (*5, 6, 9, 19*). In summary, this package improves upon the utility of SQANTI3 and opens the door for comparative isoform analysis across different conditions, diseases, tissues, etc.

References

1. Piovesan, A., Antonaros, F., Vitale, L., Strippoli, P., Pelleri, M. C. and Caracausi, M., Human protein-coding genes and gene feature statistics in 2019. *BMC Res Notes* **12**, 315 (2019).

2. Piovesan, A., Caracausi, M., Antonaros, F., Pelleri, M. C. and Vitale, L., GeneBase 1.1: a tool to summarize data from NCBI gene datasets and its application to an update of human gene statistics. *Database (Oxford)* **2016**, (2016).

3. Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E. and Gouil, Q., Opportunities and challenges in long-read sequencing data analysis. *Genome Biol* **21**, 30 (2020).

4. De Paoli-Iseppi, R., Gleeson, J. and Clark, M. B., Isoform Age - Splice Isoform Profiling Using Long-Read Technologies. *Front Mol Biosci* **8**, 711733 (2021).

5. Gupta, I., Collier, P. G., Haase, B., Mahfouz, A., Joglekar, A., Floyd, T., Koopmans, F., Barres, B., Smit, A. B., Sloan, S. A., Luo, W., Fedrigo, O., Ross, M. E. and Tilgner, H. U., Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat Biotechnol* **36**, 1197-1202 (2018).

6. Hardwick, S. A., Hu, W., Joglekar, A., Fan, L., Collier, P. G., Foord, C., Balacco, J., Lanjewar, S., Sampson, M. M., Koopmans, F., Prjibelski, A. D., Mikheenko, A., Belchikov, N., Jarroux, J., Lucas, A. B., Palkovits, M., Luo, W., Milner, T. A., Ndhlovu, L. C., Smit, A. B., Trojanowski, J. Q., Lee, V. M. Y., Fedrigo, O., Sloan, S. A., Tombacz, D., Ross, M. E., Jarvis, E., Boldogkoi, Z., Gan, L. and Tilgner, H. U., Single-nuclei isoform RNA sequencing unlocks barcoded exon connectivity in frozen brain tissue. *Nat Biotechnol* **40**, 1082-1092 (2022).

7. Joglekar, A., Prjibelski, A., Mahfouz, A., Collier, P., Lin, S., Schlusche, A. K., Marrocco, J., Williams, S. R., Haase, B., Hayes, A., Chew, J. G., Weisenfeld, N. I., Wong, M. Y., Stein, A. N., Hardwick, S. A., Hunt, T., Wang, Q., Dieterich, C., Bent, Z., Fedrigo, O., Sloan, S. A., Risso, D., Jarvis, E. D., Flicek, P., Luo, W., Pitt, G. S., Frankish, A., Smit, A. B., Ross, M. E. and Tilgner, H. U., A spatially resolved brain region- and cell type-specific isoform atlas of the postnatal mouse brain. *Nat Commun* **12**, 463 (2021).

8. Leung, S. K., Jeffries, A. R., Castanho, I., Jordan, B. T., Moore, K., Davies, J. P., Dempster, E. L., Bray, N. J., O'Neill, P., Tseng, E., Ahmed, Z., Collier, D. A., Jeffery, E. D., Prabhakar, S., Schalkwyk, L., Jops, C., Gandal, M. J., Sheynkman, G. M., Hannon, E. and Mill, J., Full-length transcript sequencing of human and mouse cerebral cortex identifies widespread isoform diversity and alternative splicing. *Cell Rep* **37**, 110022 (2021).

9. Palmer, C. R., Liu, C. S., Romanow, W. J., Lee, M. H. and Chun, J., Altered cell and RNA isoform diversity in aging Down syndrome brains. *Proc Natl Acad Sci U S A* **118**, (2021).

10. Sharon, D., Tilgner, H., Grubert, F. and Snyder, M., A single-molecule long-read survey of the human transcriptome. *Nat Biotechnol* **31**, 1009-1014 (2013).

11. Veiga, D. F. T., Nesta, A., Zhao, Y., Deslattes Mays, A., Huynh, R., Rossi, R., Wu, T. C., Palucka, K., Anczukow, O., Beck, C. R. and Banchereau, J., A comprehensive long-read isoform analysis platform and sequencing resource for breast cancer. *Sci Adv* **8**, eabg6711 (2022).

12. Tardaguila, M., de la Fuente, L., Marti, C., Pereira, C., Pardo-Palacios, F. J., Del Risco, H., Ferrell, M., Mellado, M., Macchietto, M., Verheggen, K., Edelmann, M., Ezkurdia, I., Vazquez, J., Tress, M., Mortazavi, A., Martens, L., Rodriguez-Navarro, S., Moreno-Manzano, V. and Conesa, A., SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res* **28**, 396-411 (2018).

13. Tseng, E.. (github, 2017). github.com/Magdoll/cDNA_Cupcake

14. Kuo, R. I., Cheng, Y., Zhang, R., Brown, J. W. S., Smith, J., Archibald, A. L. and Burt, D. W., Illuminating the dark side of the human transcriptome with long read transcript sequencing. *BMC Genomics* **21**, 751 (2020).

15. Wyman, Dana, Balderrama-Gutierrez, Gabriela, Reese, Fairlie, Jiang, Shan, Rahmanian, Sorena, Forner, Stefania, Matheos, Dina, Zeng, Weihua, Williams, Brian, Trout, Diane, England, Whitney, Chu, Shu-Hui, Spitale, Robert C., Tenner, Andrea J., Wold, Barbara J. and Mortazavi, Ali, A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. *bioRxiv*, 672931 (2020).

16. Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., Tyson, J. R., Beggs, A. D., Dilthey, A. T., Fiddes, I. T., Malla, S., Marriott, H., Nieto, T., O'Grady, J., Olsen, H. E., Pedersen, B. S., Rhie, A., Richardson, H., Quinlan, A. R., Snutch, T. P., Tee, L., Paten, B., Phillippy, A. M., Simpson, J. T., Loman, N. J. and Loose, M., Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol* **36**, 338-345 (2018).

17. Bertin, N., Mendez, M., Hasegawa, A., Lizio, M., Abugessaisa, I., Severin, J., Sakai-Ohno, M., Lassmann, T., Kasukawa, T., Kawaji, H., Hayashizaki, Y., Forrest, A. R. R., Carninci, P. and Plessy, C., Linking FANTOM5 CAGE peaks to annotations with CAGEscan. *Sci Data* **4**, 170147 (2017).

18. de la Fuente, L., Arzalluz-Luque, A., Tardaguila, M., Del Risco, H., Marti, C., Tarazona, S., Salguero, P., Scott, R., Lerma, A., Alastrue-Agudo, A., Bonilla, P., Newman, J. R. B., Kosugi, S., McIntyre, L. M., Moreno-Manzano, V. and Conesa, A., tappAS: a comprehensive computational framework for the analysis of the functional impact of differential splicing. *Genome Biol* **21**, 119 (2020).

19. Al'Khafaji, Aziz M., Smith, Jonathan T., Garimella, Kiran V, Babadi, Mehrtash, Sade-Feldman, Moshe, Gatzen, Michael, Sarkizova, Siranush, Schwartz, Marc A., Popic, Victoria, Blaum, Emily M., Day, Allyson, Costello, Maura, Bowers, Tera, Gabriel, Stacey, Banks, Eric, Philippakis, Anthony A., Boland, Genevieve M., Blainey, Paul C. and

Hacohen, Nir, High-throughput RNA isoform sequencing using programmable cDNA concatenation. *bioRxiv*, 2021.2010.2001.462818 (2021).

isoSeQL will be made publicly available upon publication: github.com/christine-liu/isoSeQL

Chapter 6, in part, is currently being prepared for submission for publication of the material. Liu, C.S., Chun, J. The dissertation author was the primary researcher and author of this material.

CHAPTER 7

TRANSCRIPTOMIC HALLMARKS AND RNA ISOFORM DIVERSITY IN HUMAN NEURODEGENERATIVE DISEASE

This chapter describes novel bioinformatic methods extending the utility of isoSeQL (Chapter 6) for use with single-cell long-read RNA-sequencing. An ongoing project at the time of my graduation examined six different neurodegenerative diseases (Alzheimer's disease (AD), cortical basal degeneration (CBD), dementia with Lewy bodies (DLB), Parkinson's disease (PD), Pick's disease (PiD), and progressive supranuclear palsy (PSP)) using single-cell sequencing technologies. This study utilized targeted PacBio Iso-Seq to examine a panel of 50 genes at single-cell isoform resolution.

43 samples in total were sequenced using the 10X Genomics Single Cell 3' v3.1 kit. Prior to fragmentation, the libraries were split for short-read sequencing (Illumina) and long-read sequencing (PacBio). A portion was fragmented to construct the final library for Illumina sequencing. The unfragmented library was used for target gene enrichment with a custom probe panel (Twist Biosciences). The 50 genes used for probe design were chosen for having a predicted correlation with neurological disorders from literature or for being differentially expressed in a disease group compared to controls.

We utilized a targeted approach to address the limited throughput of single-cell long-read sequencing. Previous work in the lab (see Chapter 4) and other studies have shown that limited sequencing depth is a challenge of identifying novel isoforms and their prevalence (*1-3*). While short-read sequencing approaches can output hundreds of millions of reads from a single sequencing lane, a single SMRTcell will output approximately five million reads on average. The resulting read count per cell is much lower from long-read sequencing than from short-read sequencing. By targeting a specific number of genes, we hoped to obtain sufficient coverage of those genes' isoforms across different cell types.

Several modifications had to be made to isoSeQL to accommodate the additional information provided by single-cell sequencing. All the structural and experimental/sample information tables stayed the same, but the read counts were handled very differently. The analysis process of individual sequencing runs had a few additional steps to demultiplex and deduplicate the cellular barcodes and unique molecular identifiers (UMIs). This information was then annotated with

cell types determined from corresponding short-read single-cell sequencing on the same samples. When running isoSeQL, an additional file containing the barcode and UMIs associated with various transcripts is provided.

This file is parsed into a Python dictionary to keep track of the number of unique UMIs observed in support of particular transcripts. Through the process of adding samples to the database, these counts are added to three new tables that were initialized for use with single-cell sequencing. The first table, scInfo, keeps track of the experiment ID, cellular barcode sequence, and assigned cell type. Each entry is given a unique scID number for reference in the count tables. The second and third tables keep track of counts for common junction isoforms as well as isoforms with variable ends, consistent with isoSeQL's design to group isoforms with the same common junctions as well as treat them as unique isoforms. The original counts tables (counts and ends_counts) are then populated by summing the UMIs over all the cells of a sample, reporting "pseudobulk" abundances. These numbers are not exactly the same as those that would result from analyzing the data as a bulk sample instead of single-cell because UMI duplicates are detected and removed. Without UMIs, PCR duplicate reads could be counted individually.

Once the data were consolidated into the isoSeQL database, custom queries were used to access read counts and generate plots showing the distribution of isoform structural categories, the proportion of reads from a variety of isoforms of a particular gene, the number of isoforms in common between the different disease groups, etc. The data could be queried as pseudobulk (ignoring cell type assignments) or at the cell type level, providing information about the isoforms detected in specific cell types.

References

1. Joglekar, A., Prjibelski, A., Mahfouz, A., Collier, P., Lin, S., Schlusche, A. K., Marrocco, J., Williams, S. R., Haase, B., Hayes, A., Chew, J. G., Weisenfeld, N. I., Wong, M. Y., Stein, A. N., Hardwick, S. A., Hunt, T., Wang, Q., Dieterich, C., Bent, Z., Fedrigo, O., Sloan, S. A., Risso, D., Jarvis, E. D., Flicek, P., Luo, W., Pitt, G. S., Frankish, A., Smit, A. B., Ross, M. E. and Tilgner, H. U., A spatially resolved brain region- and cell type-specific isoform atlas of the postnatal mouse brain. *Nat Commun* **12**, 463 (2021).

2. Leung, S. K., Jeffries, A. R., Castanho, I., Jordan, B. T., Moore, K., Davies, J. P., Dempster, E. L., Bray, N. J., O'Neill, P., Tseng, E., Ahmed, Z., Collier, D. A., Jeffery, E. D., Prabhakar, S., Schalkwyk, L., Jops, C., Gandal, M. J., Sheynkman, G. M., Hannon, E. and Mill, J., Full-length transcript sequencing of human and mouse cerebral cortex identifies widespread isoform diversity and alternative splicing. *Cell Rep* **37**, 110022 (2021).

3. Palmer, C. R., Liu, C. S., Romanow, W. J., Lee, M. H. and Chun, J., Altered cell and RNA isoform diversity in aging Down syndrome brains. *Proc Natl Acad Sci U S A* **118**, (2021).

Chapter 7, in part, is currently being prepared for submission for publication of the material. Park, C., Liu, C.S., Ngo, T., Saikumar, J., Palmer, C.R., Costantino, I., Romanow, W.J., Chun, J. The dissertation author was a co-primary researcher and author of this material.

CHAPTER 8

CONCLUSION/FUTURE DIRECTIONS

This chapter summarizes earlier chapters and provides insight into what I believe are directions that my work can take in the future.

Chapter 2

Chapter 2 is the response to a Matters Arising issue written by the Walsh lab, challenging the conclusions of our November 2018 *Nature* paper describing the first gencDNA. This publication resulted from a few rounds of revision during which they were able to edit their manuscript in response to ours and vice versa. Their main issues with the original *Nature* paper were the presence of plasmid contamination (which we confirmed), lack of insertion site evidence (which we provided), and the inability to replicate our findings with their own data.

In our response, we acknowledged the presence of plasmid contamination but argued that the plasmid sequences could not account for every read that was interpreted as evidence of gencDNAs (exon-exon junction spanning reads). While we can't prove this directly from those sequencing datasets, we cited 12 lines of evidence that provide orthogonal support for our findings. We additionally analyzed another group's published, whole-exome sequencing dataset from Alzheimer's disease (AD) brain samples. Their data did not have any plasmid contamination, as determined through analysis using Vecuum and VecScreen, and we found exon-exon spanning reads from *APP* (*1, 2*). These data also contained insertion site information about these *APP* gencDNAs in the form of read pairs that contained either a "clipped read" mapping across the UTR-insertion site boundary with its mate also mapping to the new insertion site or an exon-exon spanning read with its mate mapping to an entirely different locus. The additional exon-exon spanning reads were dismissed as potential mRNA contamination in view of many additional genes that were identified as having similar intron-less structures.

The Walsh group used their own single-neuron whole genome sequencing (WGS) data from AD patients to search for *APP* gencDNAs and their insertion sites. Their samples were sequenced to an average depth of 45X, but they were unable to identify *APP* retroinsertions in any of their samples. Limitations in their genome amplification method and small sample size

(average of nine neurons per individual, seven individuals total) did not convince us that these data could disprove the existence of *APP* gencDNAs.

The process of writing this response really emphasized the need for us to show evidence for each step of the gencDNA mechanism of creation. In defining a novel genomic structure, I believe that the burden of proof falls on us as a lab to convince the scientific community that it truly exists and can have ramifications. Chapter 3 details work that I did in order to unbiasedly identify other gencDNAs that could exist and be correlated with disease. Additional projects in the lab have been started to look for gencDNAs and their insertion sites using long-read sequencing, to examine potential reverse transcriptases that could facilitate the formation of gencDNAs, and to prevent their formation using reverse transcriptase inhibitors. There are a lot of unanswered questions regarding gencDNAs and several other projects could more closely examine other aspects of their existence:

- mRNA expression
- protein translation
- pattern of gencDNA insertion across different brain regions or other tissues
- relation to other diseases besides AD
- prevalence/function in health
- differences and similarities with processed pseudogenes
- insertion site influence on expression
- chromatin configuration influence on insertion
- SNVs in gencDNA sequences

There are many challenges associated with detecting gencDNAs with confidence, and full characterization of their properties may potentially be stymied by the limitations of current

technology. For example, in sequencing-based experiments, read depth and abundance cutoffs are currently implemented to more confidently distinguish between variants and noise. Variants that only occur in a single neuron may not reach those cutoffs or be distinguishable from noisy data until sequencing technology improves. Replicability and sampling are also ever-present challenges that may make it more difficult to prove that a gencDNA exists with a really low prevalence.

Chapter 3

Chapter 3 describes a novel pipeline for identifying gencDNAs unbiasedly from short-read sequencing data. The novel bioinformatics pipeline parses the alignment from STAR (a short-read aligner typically used for RNA-sequencing datasets) and identifies sequencing reads that could be indicative of gencDNA formation (*3*). This pipeline was then applied to over 3,000 samples representing various brain regions, disease states, and sequencing technologies. No clear patterns were observed linking certain genes' gencDNAs with disease or brain region. A vast majority of samples had zero gencDNAs, and from this large study, exon-exon spanning reads were estimated to occur with a frequency of 1/500,000,000 reads. More surprisingly, *APP* gencDNAs were not detected in any of the 1,000+ samples obtained from AD patients.

These results strongly contradict the findings of our paper detailing the existence of the first gencDNA. That paper estimated that *APP* gencDNAs were present in 60% of neurons in sporadic AD, and from that estimate, we would have expected that a large number of the examined samples would have *APP* exon-exon spanning reads. Potential explanations for the low frequency of exon-exon reads and the lack of patterns linking gencDNAs to disease include limitations in sequencing approaches and sampling issues. Read depth of sequencing experiments is calculated with the assumption that a single genome is the subject of the experiment. Assuming complete mosaicism, individual cells' genomes can vary, and the coverage of each cell's genome is far from complete. Current sequencing technologies often require several thousands of cells as input to obtain $\sim$30X genome coverage. The probability that a sequencing read covers an

exon-exon junction of a gencDNA that occurs in one small portion of the genome in a particular cell is extremely slim. Additionally, the expected inter-cell variability makes it possible that cells that were sampled just happened to not be affected by the variation of interest. In our case, we could interpret the lack of *APP* gencDNAs in AD samples as a result of poor sampling that may have unintentionally selected cells that were less likely to contain a gencDNA insertion or lack of sequencing coverage across the *APP* gencDNAs in cells that were present in the sample.

The pipeline that was presented in Chapter 3 is fairly incomplete in its assessment of potential gencDNA structures. The pipeline I wrote primarily focuses on identifying the exon-exon junction spanning reads. Additionally, we wanted to look for insertion site evidence to support our claims that these sequences were reverse transcribed and inserted into the genome. I made use of STAR's ability to look for chimeric reads that mapped improbably far away from each other. These types of reads would theoretically be able to link a gencDNA sequence to its insertion site even if it was inserted on an entirely separate chromosome. Yunjiao Zhu, a post-doc in the lab wrote an additional part of the pipeline for examining reads containing insertion site information from sequences that flanked UTRs. Neither pipeline resulted in convincing conclusions that the exon-exon spanning reads originated from a reverse-transcribed sequence inserted into the genome. The number of insertion site reads tracked well with expectations for known processed pseudogenes, but did not correlate well with other genes that had a lot of exon-exon spanning reads. This could be the result of insertion sites occurring in regions that are challenging to map with only a portion of a short sequencing read or the ever-present sampling issue. UTRs are not confirmed to be included as part of the retrotranscribed sequence either, which may present an incorrect assumption that the discovery pipeline was based on.

Overall, much more work needs to be put into identifying insertion sites to better confirm that these gencDNAs exist according to our original definition. Several of the shortcomings of this chapter can be alleviated through long-read sequencing. Long-read sequencing can't address the sampling issues or lack of individual genomic coverage, but the reads themselves will have less ambiguity about whether they represent a gencDNA structure. Similar considerations will

need to be made in terms of accounting for artifacts of the sample preparation and sequencing processes and determining appropriate cutoffs.

Chapter 4

Chapter 4 describes our single-cell RNA-sequencing study of Down syndrome (DS). At the time, it represented the first single-nucleus transcriptomic analysis of the post-natal human DS brain. Three main takeaways from this study were: 1) there is an increased ratio of inhibitory to excitatory neurons in DS; 2) microglia in DS appeared to be activated and have signatures of AD-related aging prior to AD-related neuropathology; and 3) vast RNA isoform diversity exists amongst different cell types of the brain.

Focusing on the long-read bioinformatics analysis in this paper - the long-read isoform analysis - there are a couple of ideas that can be expanded upon:

- Long-read-only cell type identification: We showed a proof of concept that it was possible to identify cell types using only long reads. I presented this work to Jean Fan's lab at Johns Hopkins University, and she suggested that I use a projection algorithm that she developed, MUDAN, with the intent of reducing noise in the dataset and grouping the cells by their similarities instead of trying to cluster maximizing the differences between them (*4*). While I believe that the main limitation of trying to identify cell types solely from long read sequencing is the depth (long-read sequencing has much lower throughput than short-read sequencing), removing noise from the data may also be another potential option. In terms of increasing the read depth/improving throughput, technology has already improved since we published our paper. PacBio recently released MAS-Seq, a new kit for improving single-cell long read isoform throughput by concatenating several barcoded cDNAs together to take advantage of the long polymerase read lengths that exceed the length of a single cDNA. MAS-Seq improves the throughput of a single run by 16X (*5, 6*). Additional new developments in technology include the release of the PacBio Revio - the

number of ZMWs per SMRTcell is increased to 25M ZMWs, more than three times the ZMWs on the Sequel II/e.

- Increased sequencing depth for confidence & setting cutoffs: As stated in the previous point, depth is a real limitation of identifying meaningful novel isoforms. It's pretty difficult to estimate how many reads are required to reach saturation at an isoform level given that the number of isoforms is not fixed (almost every long-read RNA-seq study identifies novel isoforms) and relative abundances will determine how easily the isoforms are captured. Unlike the genome where the number of copies of each gene is (mostly) fixed at two and there are a known number of genes, the transcriptome is a lot more variable. We've tried to address this through targeted sequencing (Chapters 4 and 7), either through PCR amplification or probe-based pulldown. PCR amplification was only able to capture isoforms that contained the primer sequences, which is potentially very limiting in cases where the 5' or 3' ends vary a lot. Twist Biosciences' probe design was based off of known isoforms; as long as novel isoforms had some sort of sequence overlap with the known ones, they could be captured, but this approach could still fail to capture novel isoforms that differ significantly from the known transcripts. Again, a solution is to start using new technologies that have significantly increased throughput and read depth, but I think more efforts could be put into establishing a general consensus read depth sufficient for isoform studies. Increased throughput will make it challenging to utilize the same bioinformatic tools for analysis that were not necessarily optimized for so much data, however, the solutions in Chapter 6 and 7 hopefully provide a starting point for starting to integrate large isoform sequencing datasets.

Chapter 5

Chapter 5 describes the modifications I made to an already existing bioinformatics tool, SQANTI3. SQANTI2 (the previous version) was used for long-read Iso-Seq analysis in Chapter 4, and while looking through the data, we noticed that there could be a few "subcategories" that

the NNC isoforms could belong to that more specifically described how the structures were novel. After trying to parse through the outputs to automate identification of novel features like intra-exonic junctions, it made more sense to incorporate the novel feature annotation into the initial SQANTI2/3 run. I first identified the seven different features and then figured out which metrics could be used to differentiate them.

The initial plan was to show the application of these modifications using a variety of datasets which I still think are potentially good future applications:

- Quality control: one of SQANTI3's main functions is to annotate many different structural characteristics of the isoforms that could be potentially used to identify artifacts. For example, intra-priming is identified by a high percentage of A's downstream of the detected end of the transcript; a stretch of sequential A's could be erroneously picked up by the poly-A selection step, leading to detection of a more truncated isoform (*7, 8*). SQANTI3 is fairly conservative when it comes to identification of novel splice sites, and without additional short-read sequencing support, isoforms with novel, non-canonical splice sites (splice sites that don't use canonical donor/acceptor sequences) are removed as potential artifacts from RT-switching during the mRNA to cDNA conversion. While the creators of SQANTI used a spike-in control sequence to identify anomalous sequence that's generated through the library preparation process, another control could be direct RNA-sequencing data. Direct RNA-sequencing datasets potentially better represent the sequences that are present in the sample without any potential modification or selection bias from RT-PCR. The additional structural information provided by my SQANTI3 modifications may be helpful for identifying features that are predominantly created artifactually. If any of the seven features is particularly enriched in cDNA libraries compared to direct-RNA or if a feature is unique to cDNA libraries, then it can be used as an additional filtering criterion.

- Disease vs non-disease: alternative splicing is implicated in several neurological and neuropsychiatric diseases in addition to cancers. In some conditions, certain splicing

patterns are expected to be affected. For example, novel exons have been reported in ALS. Being able to globally assess the ways in which novel isoforms differentiate from the expected transcripts could identify which splicing mechanisms are altered in disease.

Chapter 6

Chapter 6 describes isoSeQL, a tool I created for the purpose of unifying isoform IDs across different datasets and therefore making the datasets comparable. Personal experience from trying to compare sixteen different samples in the DS single-cell project (Chapter 4) really highlighted the need for this type of analysis, especially if I wanted to continue to use SQANTI3 for analysis (which of course I did after making my own modifications to it). Although there are a number of suggestions for comparing isoforms across datasets, all of them have limitations that make them less than ideal. isoSeQL overcomes most of the issues associated with the current recommendations and further accounts for 5'/3' end diversity.

I tried to build in some basic plotting and visualization functions to cover some basic numbers that people tend to report in isoform analyses. These types of plots don't cover all the possible different ways that people want to examine their data, but the database is query-able through loading it into Python. Several additions I'd like to add to isoSeQL include (may contain ideas that are in progress but not finished at the time of writing this dissertation):

- Ability to process single-cell data - single-cell data includes another layer of information to store in the database and query. Individual cells' isoform profiles need to be accessible in addition to aggregating and averaging counts over a specific cell type. This work is already in progress and is being used to analyze a single-cell isoform sequencing dataset from examination of six different neurodegenerative diseases (Chapter 7). The SQLite database has been configured to included UMI counts, celltypes, barcodes, etc, but many more built-in queries need to be written to make it more user-friendly.

- Designation of groups of interest - there are many different comparisons that can be made between groups of replicates that have differences in age, sex, disease, etc. Built-in

comparisons examine individual samples, but most experiments are set up to compare different groups, each with multiple replicates. I would like to be able to designate the groups of interest and which samples belong to the groups and make plots and statistical comparisons with a single command.

- Merging of multiple sequencing runs from a single sample - to increase sequencing depth, one could sequence the same library multiple times. While it's possible to merge the files of a few runs together for analysis prior to addition to the database, it would be more convenient to be able to merge replicates within the database and treat it as a single sample.

- Incorporation of short-read sequencing data - short-read sequencing abundances are still the gold standard because of the high-throughput used to obtain reliable counts. Some tools estimate isoform abundance by aligning short-read sequencing from the same sample to its own long-read transcriptome. These types of numbers can be used for validation of novel structures and for potentially more accurate estimations of isoform expression. isoSeQL currently does not keep track of corresponding short-read counts for isoforms.

Chapter 7

Chapter 7 briefly describes the alterations made to the isoSeQL pipeline for use with single-cell sequencing datasets. This project is still in progress at the time of writing this document.

The addition of single-cell sequencing support suffices; however, I can already think of things to improve:

- Speed - in the current implementation, adding isoforms from single-cell datasets and querying the resulting database is significantly slower than with bulk sequencing datasets. I would like to determine if there are ways to speed this up (more redundancy, indexing, etc)

- Merging samples - increased depth is even more imperative for single-cell sequencing datasets, and low throughput can be partially remedied by more sequencing. At the moment, the way the UMIs and numbers are stored, there is no way to simply aggregate counts across the individual sequencing runs once they are added to the isoSeQL database. Samples need to be merged during the previous analysis steps to ensure that UMIs are not counted multiple times.

References

1. Kim, J., Maeng, J. H., Lim, J. S., Son, H., Lee, J., Lee, J. H. and Kim, S., Vecuum: identification and filtration of false somatic variants caused by recombinant vector contamination. *Bioinformatics* **32**, 3072-3080 (2016).

2. Schaffer, A. A., Nawrocki, E. P., Choi, Y., Kitts, P. A., Karsch-Mizrachi, I. and McVeigh, R., VecScreen_plus_taxonomy: imposing a tax(onomy) increase on vector contamination screening. *Bioinformatics* **34**, 755-759 (2018).

3. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T. R., STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).

4. J. Fan. (github, 2018). github.com/JEFworks/MUDAN

5. PacBio (PacBio, 2022), vol. 2022. https://www.pacb.com/products-and-services/applications/rna-sequencing/single-cell-rna-sequencing/

6. Al'Khafaji, Aziz M., Smith, Jonathan T., Garimella, Kiran V, Babadi, Mehrtash, Sade-Feldman, Moshe, Gatzen, Michael, Sarkizova, Siranush, Schwartz, Marc A., Popic, Victoria, Blaum, Emily M., Day, Allyson, Costello, Maura, Bowers, Tera, Gabriel, Stacey, Banks, Eric, Philippakis, Anthony A., Boland, Genevieve M., Blainey, Paul C. and Hacohen, Nir, High-throughput RNA isoform sequencing using programmable cDNA concatenation. *bioRxiv*, 2021.2010.2001.462818 (2021).

7. Nam, D. K., Lee, S., Zhou, G., Cao, X., Wang, C., Clark, T., Chen, J., Rowley, J. D. and Wang, S. M., Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription. *Proc Natl Acad Sci U S A* **99**, 6152-6156 (2002).

8. Spies, N., Burge, C. B. and Bartel, D. P., 3' UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. *Genome Res* **23**, 2078-2090 (2013).