Title: Reply: Evidence that APP gene copy number changes reflect recombinant vector contamination

Authors: Ming-Hsiang Lee[1*], Christine S. Liu[1,2*], Yunjiao Zhu[1], Gwendolyn E. Kaeser[1], Richard Rivera[1], William J. Romanow[1], Yasuyuki Kihara[1], Jerold Chun[1]§

Affiliations

[1]Sanford Burnham Prebys Medical Discovery Institute, La Jolla, CA 92037

[2]Biomedical Sciences Program, School of Medicine, University of California San Diego, La Jolla, CA, USA 92037

[*]Co-first authors

§Corresponding Author: Jerold Chun, jchun@sbpdiscovery.org

Kim *et al.*[1] conclude that somatic gene recombination (**SGR**) and amyloid precursor protein (***APP***) genomic complementary DNAs (**gencDNAs**) in brain are plasmid PCR artifacts and do not naturally exist. We disagree. Lee *et al.*[2] presented a total of nine distinct approaches, in addition to three from a prior publication[3], which support the existence of *APP* gencDNAs, and seven of these are independent of *APP* PCR (**Table 1**). Contamination in our pull-down dataset was identified after publication of Lee *et al.*[2]; however subsequent analyses showed that the contamination does not change any of our conclusions including those in our other publications (see below)[3,4]. Notably, alterations of *APP* gencDNA number and form by Alzheimer's disease (**AD**) and cell-type cannot be explained by plasmid contamination and PCR artifact. Here we provide data and discussion, which address the three analyses used by *Kim et al.*[1] to reach their conclusions: plasmid contaminant identification, plasmid PCR, and single-cell sequencing.

Ten uncontaminated whole-exome pull-down datasets identify *APP* gencDNAs

An *APP* plasmid contaminant (pGEMT *APP*) in our single pull-down dataset was found, however we could not definitively determine which *APP* exon::exon reads were due to gencDNAs vs. contamination, especially in view of the 11 other uncontaminated approaches that had independently identified *APP* gencDNAs. We therefore pursued confirmatory data through three additional pull-down experiments on seven human brain samples, using new reagents (*e.g.,* Agilent Human All Exon Kit (V6)). The resulting three datasets are free of *APP* plasmid contamination by both VecScreen[5] and subsequent use of *Kim et al.'s* Vecuum script[6]: critically, all of our uncontaminated datasets contain *APP* gencDNA sequences (**Figure 1a,b**). We also analyzed an independent whole-exome pull-down study from an unrelated laboratory (Park *et al.*)[7] and identified *APP* gencDNA sequences in five different sporadic Alzheimer's disease (**SAD**) brains

and two blood samples: again all are plasmid-free by Vecuum[6] (**Figure 1a-e**). Notably, Park *et a*l.'s pull-down contained probes recognizing 5' and 3' UTRs and we identified *APP* UTR sequences paired with reads containing *APP* gencDNA exon::exon junctions (**Figure 1d,e**). Thus, uncontaminated pull-down datasets from two different laboratories have now identified *APP* gencDNA sequences, including some with an UTR, all of which are consistent with our 11 unchallenged approaches that support the biological existence of *APP* gencDNAs (**Table 1**).

Detection of gencDNAs without use of *APP* PCR

Kim *et al*.[1] inaccurately states that "IEJs were exclusively reported in the PCR-based methods". Lee *et al*.[2], presented four lines of evidence for *APP* gencDNAs containing intraexonic junctions **(IEJs)** that are independent of *APP* PCR: two different commercially produced cDNA single molecule real-time (**SMRT**) sequencing libraries, DNA *in situ* hybridization and RNA *in situ* hybridization (**DISH** and **RISH**, respectively). SMRT sequencing libraries revealed IEJs within *APP* (**Lee *et al.* Extended Data Figure 1E**)[2] as well as other genes (**Extended Data Figure 1)**, which cannot be attributed to plasmid contamination or PCR amplification. DISH and RISH results each separately support the existence of *APP* gencDNAs (see **Extended Data Discussion and Lee *et al*., Figure 2, Extended Data Figures 1 and 2**)[2] using custom-designed and validated commercial probe technology (Advanced Cell Diagnostics), independently shown to detect exon::exon junctions[8] and single nucleotide mutations[9]. Thus, gencDNAs and IEJs are detectable in the absence of targeted PCR.

Importantly, contamination as proposed by Kim *et al.* cannot account for the dramatic change in number and form of gencDNAs occurring with disease state. Change is also apparent when

comparing cell types, where signals are more prevalent in SAD neurons compared to non-neurons from the same brain and processed at the same time by SMRT sequencing or DISH (**Lee *et al.* Figures 3, 5**)[2]. Independent PNA-FISH and dual-point-paint experiments from our previous work further support *APP* gencDNAs[3] (**Table 1**). Critically, SMRT sequencing identified 11 single nucleotide variations that are considered pathogenic in familial AD, which were only present in our SAD samples, while none exist as plasmids in our laboratory: contamination cannot explain these results.

Non-biological data are generated by PCR of *APP* plasmids

Kim *et al*. then used PCR of *APP* splice variant plasmids, which generated sequences containing IEJs. However, multiple discrepancies in approach and results differ from our biological IEJs and gencDNAs.: **1)** experimental conditions beyond our primer sequences were different: Kim *et al*. employed twice the concentration of primers and >1 million times more template (250 picograms of *APP* plasmid is 4.6 x 10^7 copies vs. ~40 gencDNA copies in our PCR of 20 nuclei (**Lee *et al.* Figure 5**[2]: DISH 16/17 averaged ~1.8 copies/SAD nucleus); **2**) both gencDNA and IEJ sequences can be detected with as few as 30 cycles of PCR as we used in SMRT sequencing (**Lee *et al.* Figure 3**)[2]; **3**) their agarose gels are uniformly and unambiguously dominated by a vastly over-amplified ~2 kb band (**Kim *et al.* Figure 1c and Extended Data Figure 3a**) that is never seen using human neurons despite our routine identification of myriad smaller bands (*c.f.,* **Lee *et al.* Figure 2b**)[2]. We did observe an over-amplified ~2 kb band in our purposeful plasmid transfection experiments that also utilized PCR; however, gencDNA and IEJ formation was comparatively limited and critically, required both reverse transcriptase activity and DNA strand breakage (**Lee**

***et al.*, Figure 4[2]**); and **4**) only 45 unique IEJs from AD and 20 from non-diseased brains were identified (**Lee *et al.* Figure 3**; with some overlap, fewer than 65 together)[2] compared to the 12,426 identified by Kim *et al.* (~200-fold increase over biological IEJs; Kim *et al.* Supplementary Table 1). We wish to note that micro-homology regions within *APP* exons are intrinsic to *APP's* DNA sequence, and that micro-homology mediated repair mechanisms involve DNA polymerases[10,11]. Kim *et al.'s* PCR results differ from our biological data yet may inadvertently support endogenous formation of at least some IEJs within DNA rather than requiring RNA.

Single-cell whole genome sequencing limitations may prevent *APP* gencDNA detection

Kim *et al.*[1]*'s* third line 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 per SAD brain were examined, raising immediate sampling issues required to detect mosaic *APP* gencDNAs. Self-identified "uneven genome amplification" (Kim *et al.* and [12-14]) produces ~20% of the genome having less than 10X depth of coverage[14] with potential amplification failure at one (~9% allelic dropout rate) or both alleles (~2.3% locus dropout rate)[12,14]. The limitations are compounded by potential amplification biases reflected by whole-genome amplification failure rates that may miss neuronal subtypes and/or disease states, especially relevant to single copies of *APP* gencDNAs that are as small as 0.15 kb (but still detectable by DISH). All of these issues are further compounded by genomic mosaicism[15] demonstrated by Kim *et al.'s* data on *ZNF100*, where somatic insertion was identified by exonic read depth gain in only 3 out of 10 cells from just 1 out of 7 individuals (Kim *et al.*[1] Figure 2b; 3/64 cells). Still further, Kim *et al.'s* informatic analysis is predicated on the unproven assumption that gencDNA structural features are shared with processed pseudogenes and LINE1 elements

(Kim *et al.*[1] Figure 2a and Extended Data Figure 1a), and possible differences could prevent straightforward detection under even ideal conditions. These issues could explain Kim *et al.'s* negative results.

Considering these many points, we believe that our data and conclusions supporting SGR and *APP* gencDNAs remain intact, warranting their further study in the normal and diseased brain.

Author Contributions

MHL, YK, and WR completed laboratory experiments and CSL and YZ completed all bioinformatics; All authors wrote and edited the manuscript.

Competing interests

None declared.

Data availability

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

Code availability

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

Acknowledgements

We thank Dr. Laura Wolszon and Ms. Danielle Jones for copy editing help on this manuscript.

This work was supported by The Shaffer Family Foundation, The Bruce Ford & Anne Smith Bundy Foundation, NIH P50-AG00513 (UCSD pilot grant), and SBP institutional funds (J.C.) and R01NS103940 (Y.K.).

Figures

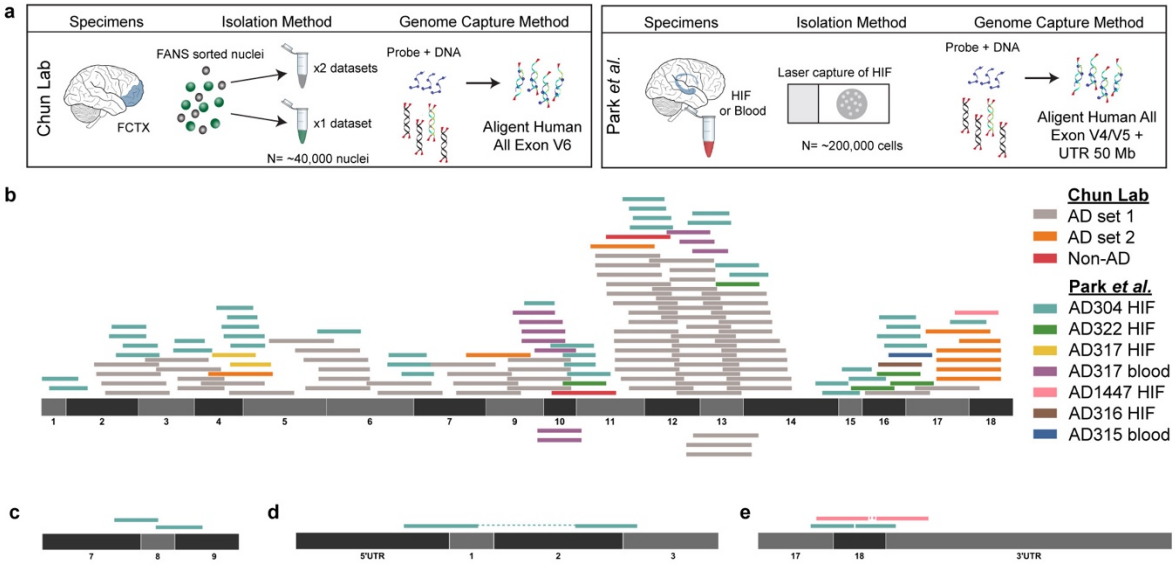

Figure 1: Identification of APP gencDNA sequences in whole-exome pull-down from 10 datasets and two independent laboratories. a) Methods schematic depicting the standard protocol followed for all-exome pull-downs and methodological differences between the independent laboratories. b) *APP-751* sequence with non-duplicate gencDNA reads from uncontaminated datasets; color key indicates source of unique reads for all panels. c) Reads mapping 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.

Table 1: Summary of targeted and non-targeted *APP* PCR methods and results that support APP gencDNAs and IEJs.

	Method	Targeted APP PCR	Support for the existence of IEJs and gencDNAs	Reference
Approaches without targeted *APP* PCR				
1	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 sequencing	None	An independent commercial source identified IEJs in *APP* and other genes	Public data set[1]; Lee *et al.*; this Reply.
3	Targeted RNA SMRT sequencing	None	RNA pull-down that identified *APP* IEJs	Public data set[1]; Lee *et al.*
4	Agilent SureSelect Targeted Pull-down	None	Identified *APP* gencDNAs in SAD brains; contains plasmid sequence contamination	Lee *et al.*; Kim *et al.*; This Reply.
5	DISH of gencDNAs	None	IEJ 3/16 and exon::exon junction 16/17 show increases in SAD neurons compared to non-neurons from the same brain and non-diseased neurons; J20 mice containing the *APP* transgene under a PDGF-β-promoter show increased number and size of signal compared to non-neurons and WT mice	Lee *et al.*
6	Dual point-paint FISH	None	Identified *APP* CNVs of variable puncta size that were not always associated with Chr21	Bushman *et al.*
7	PNA-FISH	None	*APP* exon copy number increases show variable signal size and shape with semi-quantitative exonic probes	Bushman *et al.*
New	Agilent All-Exon pull-down	None	All-Exon pull-downs with no plasmid contamination by Vecuum contain *APP* gencDNA sequences and evidence of gencDNA UTRs	Park *et al.*; This Reply.
Approaches with targeted *APP* PCR				
8	RT-PCR and Sanger sequencing	OligoDT primed and targeted *APP* primers	Novel *APP* RNA variants with IEJs; predominantly in neurons from SAD brains	Lee *et al.*
9	Genomic DNA PCR and Sanger Sequencing	Yes	Identified *APP* gencDNAs with IEJs; predominantly in neurons from SAD brains	Lee *et al.*
10	Genomic DNA PCR and SMRT sequencing	Yes	IEJ/gencDNAs were more prevalent in number and form in SAD neurons compared to non-diseased neurons; Identified 11 pathogenic SNVs that were only present in SAD samples	Lee *et al.*
11	APP-751 over-expression in CHO cells	Yes	IEJ and gencDNA formation required DNA strand breakage and reverse transcriptase	Lee *et al.*
12	Single-cell qPCR	Yes; individual exon	Intragenic Exon 14 single-cell qPCR showed copy number increases in SAD prefrontal cortex neurons over cerebellar neurons from the same brain	Bushman *et al.*

CNV, copy number variation; DISH, DNA in situ hybridization; FISH, fluorescence in situ hybridization; IEJ, intra-exonic junction; PNA, peptide nucleic acid; RISH, RNA in situ hybridization; SAD, sporadic Alzheimer's disease; SMRT, single molecule real-time.

[1]The Alzheimer brain Iso-Seq dataset was generated by Pacific Biosciences, Menlo Park, California, and additional information about the sequencing and analysis is provided at https://downloads.pacbcloud.com/public/dataset/Alzheimer_IsoSeq_2016/.

References

1 Junho Kim, B. Z., August Yue Huang, Michael B. Miller, Michael A. Lodato, Christopher A. Walsh, Eunjung Alice Lee. in *bioRxiv* (2019).

2 Lee, M. H. *et al.* 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. *et al.* 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 Rohrback, S. *et al.* 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, doi:10.1073/pnas.1812702115 (2018).

5 <https://www.ncbi.nlm.nih.gov/tools/vecscreen/about/>

6 Kim, J. *et al.* Vecuum: identification and filtration of false somatic variants caused by recombinant vector contamination. *Bioinformatics* **32**, 3072-3080, doi:10.1093/bioinformatics/btw383 (2016).

7 Park, J. S. *et al.* 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).

8 Visualize EGFRvIIIsplice variant in glioblastoma sample: Enabled by BaseScope Assay https://acdbio.com/science/applications/research-areas/egfrviii. (2019).

9 Baker, A. M. *et al.* Robust RNA-based in situ mutation detection delineates colorectal cancer subclonal evolution. *Nat Commun* **8**, 1998, doi:10.1038/s41467-017-02295-5 (2017).

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

11 Sfeir, A. & Symington, L. S. Microhomology-Mediated End Joining: A Back-up Survival Mechanism or Dedicated Pathway? *Trends in Biochemical Sciences* **40**, 701-714, doi:10.1016/j.tibs.2015.08.006 (2015).

12 Evrony, G. D. *et al.* Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell* **151**, doi:10.1016/j.cell.2012.09.035 (2012).

13 Cai, X. *et al.* Single-cell, genome-wide sequencing identifies clonal somatic copy-number variation in the human brain. *Cell reports* **8**, 1280-1289, doi:10.1016/j.celrep.2014.07.043 (2014).

14 Evrony, G. D. *et al.* 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. & Chun, J. Genomic mosaicism in the developing and adult brain. *Dev Neurobiol* **78**, 1026-1048, doi:10.1002/dneu.22626 (2018).