Genome-wide structural variant analysis identifies risk loci for non-Alzheimer's dementias

Author list

Karri Kaivola, Ruth Chia, Jinhui Ding, Memoona Rasheed, Masashi Fujita, Vilas Menon, Ronald L. Walton, Ryan L. Collins, Kimberley Billingsley, Harrison Brand, Michael Talkowski, Xuefang Zhao, Ramita Dewan, Ali Stark, Anindita Ray, Sultana Solaiman, Pilar Alvarez Jerez, Laksh Malik, Ted M. Dawson, Liana S. Rosenthal, Marilyn S. Albert, Olga Pletnikova, Juan C. Troncoso, Mario Masellis, Julia Keith, Sandra E. Black, Luigi Ferrucci, Susan M. Resnick, Toshiko Tanaka; The American Genome Center; International LBD Genomics Consortium; International ALS/FTD Consortium; PROSPECT Consortium; Eric Topol, Ali Torkaman, Pentti Tienari, Tatiana M. Foroud, Bernardino Ghetti, John E. Landers, Mina Ryten, Huw R. Morris, John A. Hardy, Letizia Mazzini, Sandra D'Alfonso, Cristina Moglia, Andrea Calvo, Geidy E. Serrano, Thomas G. Beach, Tanis Ferman, Neill R. Graff-Radford , Bradley F. Boeve, Zbigniew K. Wszolek, Dennis W. Dickson , Adriano Chiò, David A. Bennett, Philip L. De Jager , Owen A. Ross, Clifton L. Dalgard, J. Raphael Gibbs, Bryan J. Traynor, Sonja W. Scholz

Summary

Initial submission:	Received : January 23rd 2023
	Scientific editor: Judith Nicholson
First round of review:	Number of reviewers: 3 Revision invited : March 21st 2023 Revision received : March 29th 2023
Second round of review:	Number of reviewers: N/A Accepted : April 6th 2023
Data freely available:	Yes
Code freely available:	Yes

This transparent peer review record is not systematically proofread, type-set, or edited. Special characters, formatting, and equations may fail to render properly. Standard procedural text within the editor's letters has been deleted for the sake of brevity, but all official correspondence specific to the manuscript has been preserved.

Referees’ reports, first round of review

Reviewer 1

This article is well analysed and presented. The authors have analysed structural variants in >4000 cases and controls of non Alzheimer's dementia. They have used the GATK-SV pipeline which combines five SV detection algorithms to increase the reliability of the findings. The GATK-SV pipeline was applied to short read WGS data. More importantly, the variants have been made available for public use which is important for future research and clinical application.

Please see my comments below:
1. How was the clinical diagnosis of LBD as well as FTD/ALS made in these cohorts? Please detail any criteria used for diagnosis (e.g. McKeith criteria for LBD). This is important as it is difficult to clinically distinguish between LBD and Alzheimer and the signals seen in LBD may represent Alzheimer. I will be hesitant to call the TPCN1 deletion as being specific to "non-Alzheimer's dementia"
2. The mean validation rate of 84.3% is not very high even for the deletions. This limitation of short read sequencing data needs to be acknowledged again in the discussion section

Overall, this study is an important step in understanding the role of structural variants in dementia.

Reviewer 2

This is a very nicely presented paper that uses up to date wet lab and analytical methods to investigate structural variation in a number of dementias. The data presented are interesting, with some novel findings and are made available to other researchers for future work.

The cohorts, methods and analyses are clearly and adequately described. The figures and tables are clear and readable with sufficient legends.

Expansion of the work to include integration of other omics to determine the potential function of the variants and perhaps suggest molecular follow up would be preferential and expand the readership of the manuscript but is not essential.

Two minor points below

Clarify this sentence are 4699 of the 4889 common?
After quality control filtering, 4,889 and 4,699 common (i.e., MAF ≥ 1%), high-quality structural variants were available for association testing in the LBD and the FTD/ALS cohorts

Typo
Figyre 5 (page 15)

Reviewer 3

Kaivola, Chia, Ding, and their colleagues have conducted a study to identify and characterize structural variations (SVs) in two non-Alzheimer's dementia disorders: Lewy body dementia (LBD) and frontotemporal dementia/amyotrophic lateral sclerosis (FTD/ALS). To discover SVs, they used GATK-SV, a sophisticated SV calling pipeline that is also used in gnomAD-SV, a reference map of SVs (Collins et al. Nature 2020). By analyzing whole genome sequencing data of 5,213 cases and 4,132 controls of European ancestry, they identified over 300,000 SVs, which corresponds to an average of about 800 per genome. With the SV catalog, they performed genome-wide association studies (GWAS) and identified common SVs that act as disease risk loci. Specifically, they detected a deletion in TPCN1 as a novel risk locus for LBD and known SVs at the C9orf72 and MAPT loci as risk loci for FTD/ALS. They also discovered and cataloged rare pathogenic SVs. This study is the first large cohort SV study that uses whole-genome sequencing data in non-Alzheimer's dementia to my knowledge. However, I have some concerns and questions regarding their analysis.

Concerns and questions:

1.My main concern is the relatively low number of SVs discovered in this study compared to recent population-scale SV studies (Collins et al. Nature 2020, Abel et al. Nature 2020) that utilized high-coverage WGS data. Abel et al. reported an average of 4,442 SVs per genome, while Collins et al. reported a median of 7,439 SVs per genome. It is worth noting that Collis et. al developed and used GATK-SV. A similar SV study in Parkinson's disease (Billingsley et al. Ann. Neurol. 2023), which also used GATK-SV, reported an average of 5,626 SVs per genome. However, in this study, the authors only found an average of 895 SVs per genome in the LBD case-control cohort and an average of 865 SVs per genome in the FTD/ALS case-control cohort. These counts are at least five times lower than those reported by other studies that used the same GATK-SV pipeline.

2.Assuming both cohorts share the same control samples for unaffected individuals, I do not understand why the numbers of controls are different in the SV mapping step and the following steps in Figure 1.

3.In relation to concern 1, I would suggest putting the violin plot of the SV count per genome for the final SV call set in Figure 2.

4.The stacked bar plots in Figure 2 make it difficult to compare allele frequency (AF) and size distributions by SV type. Instead, I would recommend using normalized distributions for the final SV call set, as demonstrated in Figure 3 of Billingsley et al.

Authors' response to the first round of review

REVIEWER 1:

1. This article is well analysed and presented. The authors have analysed structural variants in >4000 cases and controls of non Alzheimer's dementia. They have used the GATK-SV pipeline which combines five SV

detection algorithms to increase the reliability of the findings. The GATK-SV pipeline was applied to short read WGS data. More importantly, the variants have been made available for public use which is important for future research and clinical application.

Response: We thank the reviewer for their kind review of our study.

2. How was the clinical diagnosis of LBD as well as FTD/ALS made in these cohorts? Please detail any criteria used for diagnosis (e.g. McKeith criteria for LBD). This is important as it is difficult to clinically distinguish between LBD and Alzheimer and the signals seen in LBD may represent Alzheimer.

   Response: Thank you for this comment, and we appreciate the opportunity to elaborate on the case definitions. Patients with FTD (behavioral variant, primary progressive aphasia) were diagnosed according to the Neary criteria (Faber, 1999) or the Movement Disorders Society criteria for progressive supranuclear palsy (Höglinger et al., 2017), whereas ALS/FTD cases were diagnosed according to the El Escorial criteria (Brooks, 1994). LBD cases were diagnosed with either pathologically definite (69% of cases) or clinically probable disease (31% of cases), according to consensus criteria (Emre et al., 2007; McKeith et al., 2005). We updated the Methods to elaborate on the case definitions as follows (page 25, paragraph 1): "Briefly, LBD patients were diagnosed with pathologically definite (69.05% of the cohort) or clinically probable disease (30.95%) according to the McKeith and Emre consensus criteria 46,47 . These consensus criteria guide optimal methods to establish the clinical and pathological diagnosis of LBD, including diagnostic biomarkers. The FTD/ALS cohort included 1,377 patients diagnosed with FTD spectrum disorders, including the known subtypes of behavioral variant FTD, primary progressive aphasia, and progressive supranuclear palsy, and 1,065 patients diagnosed with ALS. Patients with FTD were diagnosed according to the Neary criteria 48 or the Movement Disorders Society criteria for progressive supranuclear palsy 49 . These criteria define core measures and several supportive and exclusion criteria for establishing a diagnosis of FTD. Patients with ALS were diagnosed according to the revised El Escorial criteria 50 . These criteria classify patients according to the level of diagnostic certainty and have been shown to be specific to the diagnosis of ALS."

3. I will be hesitant to call the TPCN1 deletion as being specific to "non-Alzheimer's dementia".

Response: We entirely agree with the reviewer on this point, and we did not mean to suggest the TPCN1 deletion is specific to non-Alzheimer's dementias. Although there is suggestive evidence for Alzheimer's disease, the role of the TPCN1 deletion has only been reproducibly associated with LBD. We reviewed our manuscript and confirmed that there is no language indicating that the TPCN1 deletion is specific to non-Alzheimer's dementia. We added the following line to the Discussion (page 15, paragraph 3): "The potential role of the TPCN1 locus in dementia is supported by a recent GWAS in Alzheimer's disease that reported a suggestive association between the intronic TPCN1 variant rs6489896 and Alzheimer's disease 22. This rs6489896 variant tagged the 309 base-pair TPCN1 deletion in our data, and there was a near-perfect correlation between the haplotypes in Alzheimer's disease GWAS and our study (Figure 5A). These data indicate that the TPCN1 deletion is not specific to LBD. The suggestive association between TPCN1 and Alzheimer's disease could be due to the well-established shared etiology between Alzheimer's disease and LBD, or a subpopulation of LBD-variant Alzheimer's disease patients."

4. The mean validation rate of 84.3% is not very high even for the deletions. This limitation of short read sequencing data needs to be acknowledged again in the discussion section.

Response: We agree with the reviewer that the validation rates of structural variants based on short-read sequencing data is still needs improvement. The issue affects the field using shortread sequencing, not just our paper. We were trying to convey in the manuscript that our data are comparable to other studies. Please see our response to Reviewer 3, point 2, and note the addition of Supplementary Table 3 comparing our study with previous publications. We have modified the Discussion to highlight this problem as follows (page 16, paragraph 3): "The main limitations of our study stem from the inherent difficulty of calling structural variants from short-read whole-genome sequencing data. As such, the validation rate of structural variants detected in short-read sequencing data is not ideal. To mitigate this problem, we used a multi-algorithm pipeline, GATK-SV 11, to create consensus structural variant calls and focused on a subset of high-quality structural variants in our analyses. Overall, the mean number of structural variant sites, the distribution of structural variant types, and the proportion of structural variant calls in Hardy-Weinberg equilibrium are within the expected limits. Moreover, we found good structural variant mapping precision and genotype concordance between short-read sequencing and long-read sequencing data (Supplementary Table 2). These findings indicate that our structural variant calls are robust. There are no accepted filtering standards following the GATK-SV pipeline, and direct comparison with other studies employing different filters can be complex (Supplementary Table 3). However, these findings indicate that our structural variant calls are robust.”

5. Overall, this study is an important step in understanding the role of structural variants in dementia.

Response: Thank you again for reviewing our study and for your positive comments.

REVIEWER 2

1. This is a very nicely presented paper that uses up to date wet lab and analytical methods to investigate structural variation in a number of dementias. The data presented are interesting, with some novel findings and are made available to other researchers for future work. The cohorts, methods and analyses are clearly and adequately described. The figures and tables are clear and readable with sufficient legends.

   Response: We thank the reviewer for their kind review of our study.

2. Expansion of the work to include integration of other omics to determine the potential function of the variants and perhaps suggest molecular follow up would be preferential and expand the readership of the manuscript but is not essential.”

   Response: We agree with the reviewer that our study will stimulate further follow-up research studies to assess the functional impact of the observed structural variants in non-Alzheimer dementia syndromes. While it is exciting to think of various multi-omic data integration efforts, these analyses would go beyond the scope of the current article. We thank the reviewer for their understanding. We have modified the Discussion to reflect the need for additional molecular and multi-omic work as follows (page 17, paragraph 1): “…Another challenge when studying structural variants is the assessment of pathogenicity and unraveling the mechanism by which they disrupt neuronal function. As previous data are scarce and the consequences of structural variants are not well understood, cell biology studies, especially those integrating other multi-omic data, are needed to fully understand the consequences of this mutation class.”

3. Clarify this sentence are 4699 of the 4889 common? “After quality control filtering, 4,889 and 4,699 common (i.e., MAF ≥ 1%), high-quality structural variants were available for association testing in the LBD and FTD/ALS

cohorts."

Response: Thank you for your comment. We defined common structural variants as variants with a minor allele frequency of ≥ 1%. To improve the readability of the statement above and avoid confusion, we restructured the Results as follows (page 10, paragraph 2): "After quality control filtering, we performed GWAS studies on common (i.e., MAF ≥ 1%), high-quality variants separately for the LBD and FTD-ALS case-control cohorts. In total, 4,899 structural variants were tested for association with LBD, and 4,699 variants were tested for FTD/ALS

4. - Typo Figyre 5 (page 15)

Response: Thank you for pointing out this typographical error. We corrected it in the revised version of our manuscript.

REVIEWER 3

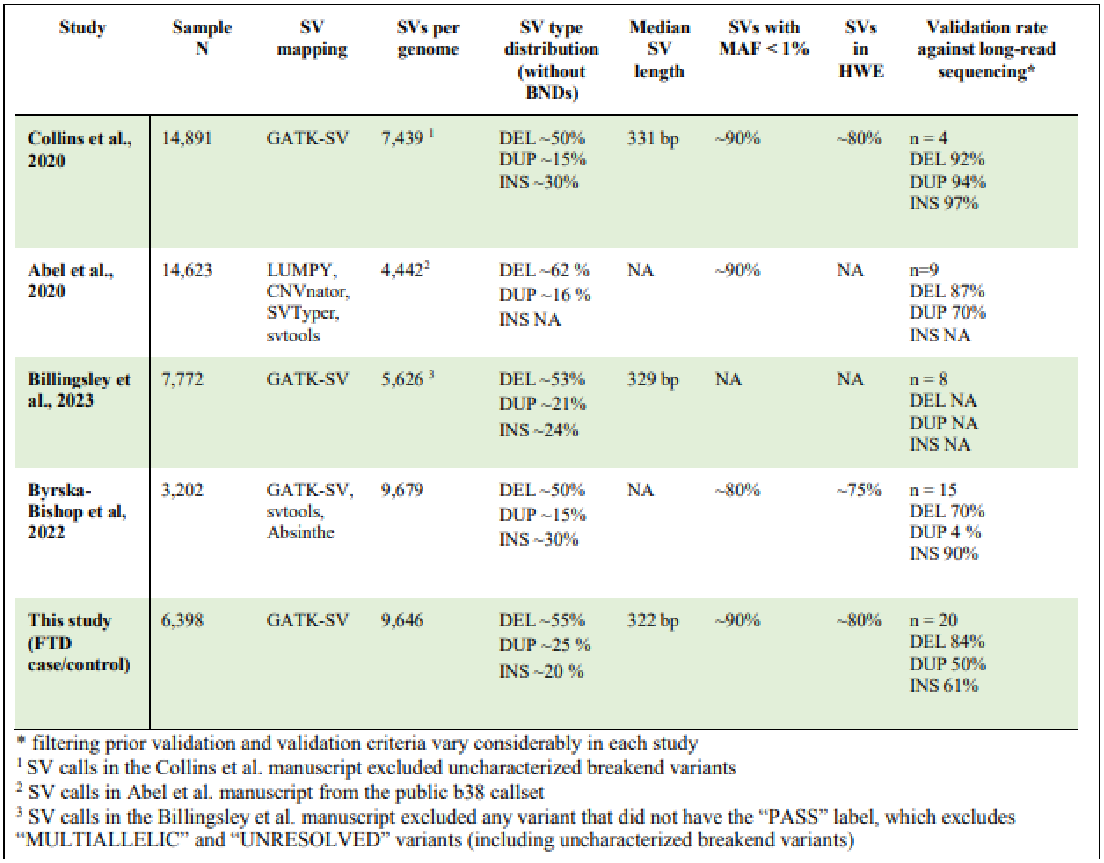
1. Kaivola, Chia, Ding, and their colleagues have conducted a study to identify and characterize structural variations (SVs) in two non-Alzheimer's dementia disorders: Lewy body dementia (LBD) and frontotemporal dementia/amyotrophic lateral sclerosis (FTD/ALS). To discover SVs, they used GATK-SV, a sophisticated SV calling pipeline that is also used in gnomADSV, a reference map of SVs (Collins et al. Nature 2020). By analyzing whole genome sequencing data of 5,213 cases and 4,132 controls of European ancestry, they identified over 300,000 SVs, which corresponds to an average of about 800 per genome. With the SV catalog, they performed genome-wide association studies (GWAS) and identified common SVs that act as disease risk loci. Specifically, they detected a deletion in TPCN1 as a novel risk locus for LBD and known SVs at the C9orf72 and MAPT loci as risk loci for FTD/ALS. They also discovered and cataloged rare pathogenic SVs. This study is the first large cohort SV study that uses whole-genome sequencing data in non-Alzheimer's dementia to my knowledge.

Response: Thank you for your kind summary of our study.

2. My main concern is the relatively low number of SVs discovered in this study compared to recent population-scale SV studies (Collins et al. Nature 2020, Abel et al. Nature 2020) that utilized high-coverage WGS data. Abel et al. reported an average of 4,442 SVs per genome, while Collins et al. reported a median of 7,439 SVs per genome. It is worth noting that Collis et. al developed and used GATK-SV. A similar SV study

in Parkinson's disease (Billingsley et al. Ann. Neurol. 2023), which also used GATK-SV, reported an average of 5,626 SVs per genome. However, in this study, the authors only found an average of 895 SVs per genome in the LBD case-control cohort and an average of 865 SVs per genome in the FTD/ALS casecontrol cohort. These counts are at least five times lower than those reported by other studies that used the same GATK-SV pipeline.

Response: We thank the reviewer for raising this critical point and apologize for not making this aspect of our work more transparent in the original manuscript. We identified a comparable number of structural variants in our study. The difference in the structural variant counts compared to other studies is mainly due to differences in the filtering that was applied following the GATK-SV pipeline to ensure we analyzed high-quality variants for association. In the supplemental table below, we now show the structural variant calls in several recent publications:



Study	Sample N	SV mapping	SVs per genome	SV type distribution (without BNDs)	Median SV length	SVs with MAF < 1%	SVs in HWE	Validation rate against long-read sequencing*
Collins et al., 2020	14,891	GATK-SV	7,439 [1]	DEL ~50% DUP ~15% INS ~30%	331 bp	~90%	~80%	n = 4 DEL 92% DUP 94% INS 97%
Abel et al., 2020	14,623	LUMPY, CNVnator, SVTyper, svtools	4,442[2]	DEL ~62 % DUP ~16 % INS NA	NA	~90%	NA	n=9 DEL 87% DUP 70% INS NA
Billingsley et al., 2023	7,772	GATK-SV	5,626 [3]	DEL ~53% DUP ~21% INS ~24%	329 bp	NA	NA	n = 8 DEL NA DUP NA INS NA
Byrska-Bishop et al, 2022	3,202	GATK-SV, svtools, Absinthe	9,679	DEL ~50% DUP ~15% INS ~30%	NA	~80%	~75%	n = 15 DEL 70% DUP 4 % INS 90%
This study (FTD case/control)	6,398	GATK-SV	9,646	DEL ~55% DUP ~25 % INS ~20 %	322 bp	~90%	~80%	n = 20 DEL 84% DUP 50% INS 61%

* filtering prior validation and validation criteria vary considerably in each study

[1] SV calls in the Collins et al. manuscript excluded uncharacterized breakend variants

[2] SV calls in Abel et al. manuscript from the public b38 callset

[3] SV calls in the Billingsley et al. manuscript excluded any variant that did not have the "PASS" label, which excludes "MULTIALLELIC" and "UNRESOLVED" variants (including uncharacterized breakend variants)

Several additional points are relevant to this issue:

- In the Collins et al. article, the median number of 7,439 structural variants per genome does not contain uncharacterized breakend (BND) variants. Breakend variants account for approximately one third of all structural variant calls. When we account for these breakend variants, the structural variant calls per genome in the Collins dataset (7,439 x 1.33 = 9,894) are almost identical to the number we observed in our study.
- In the Abel et al. article, the 5,626 structural variants per genome refers to filtered "high-confidence" structural variants.
- In the Billingsley et al. article, they identified 366,555 structural variants in 7,772 genomes with GATK-SV. They then applied additional filtering, leaving 227,357 structural variants across the cohort and averaging 5,626 structural variants per genome. There is no standard pipeline for structural variant filtering, so comparing structural variant counts with other publications is challenging. We have added this important point to the Discussion as follows (page 16, paragraph 3): "The main limitations of our study stem from the inherent difficulty of calling structural variants from short-read whole-genome sequencing data. As such, the validation rate of structural variants detected in short-read sequencing data is not ideal. To mitigate this problem, we used a multi-algorithm pipeline, GATK-SV 11, to create consensus structural variant calls and focused on a subset of high-quality structural variants in our analyses. Overall, the mean number of structural variant sites, the distribution of structural variant types, and the proportion of structural variant calls in Hardy-Weinberg equilibrium are within the expected limits. Moreover, we found good structural variant mapping precision and genotype concordance between short-read sequencing and long-read sequencing data (Supplementary Table 2). These findings indicate that our structural variant calls are robust. There are no accepted filtering standards following the GATK-SV pipeline, and direct comparison with other studies employing different filters can be complex (Supplementary Table 3). However, these findings indicate that our structural variant calls are robust." We have added the table above as Supplementary Table 3 to page 13 of the Supplementary Materials. The goal of a given study impacts structural variant filtering. For example, the Collins et al. study aimed to create a population-level reference of variants, while case-control association studies focus on a smaller subset of high-quality structural variants to minimize false-positive association signals. Finally, we now provide the unfiltered structural variant dataset in the

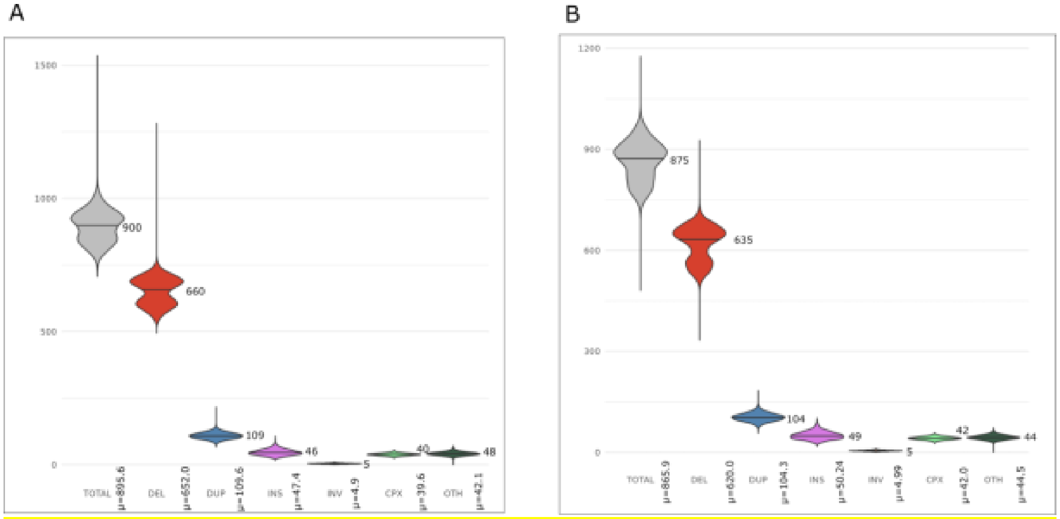
supplements (Supplementary Table 8) in addition to the filtered structural variant data (Supplementary Table 7). We already show the corresponding descriptive statistics (Supplementary Figures 4- 6) so other researchers can filter the data to match their study pipelines.

3. Assuming both cohorts share the same control samples for unaffected individuals, I do not understand why the numbers of controls are different in the SV mapping step and the following steps in Figure 1.

Response: It is correct that the LBD and FTD cohorts use the same initial set of control samples. However, after the GATK-SV pipeline was applied, there was a difference of ~ 20 samples due to the quality control steps in this pipeline. The thresholds for sample exclusions are not necessarily fixed. Instead, they are based on values derived from all the samples that are analyzed together, which includes metrics for ancestry and coverage. This filtering can (and did) lead to minor differences in which control samples are excluded when called together with cases. We added the following paragraph to the Methods to clarify this point (page 27, paragraph 2): "Of note, the LBD and FTD cohorts used the same initial set of control samples. However, there was a difference of 23 control samples after the GATK-SV pipeline was applied due to the initial quality control steps in the pipeline. The thresholds for sample exclusions are not fixed. Instead, they are based on values derived from the analyzed samples, including metrics for ancestry and coverage. This filtering approach led to minor difference in which control samples were excluded when they were called separately with the LBD cases and FTD cases."

4. In relation to concern 1, I would suggest putting the violin plot of the SV count per genome for the final SV call set in Figure 2.

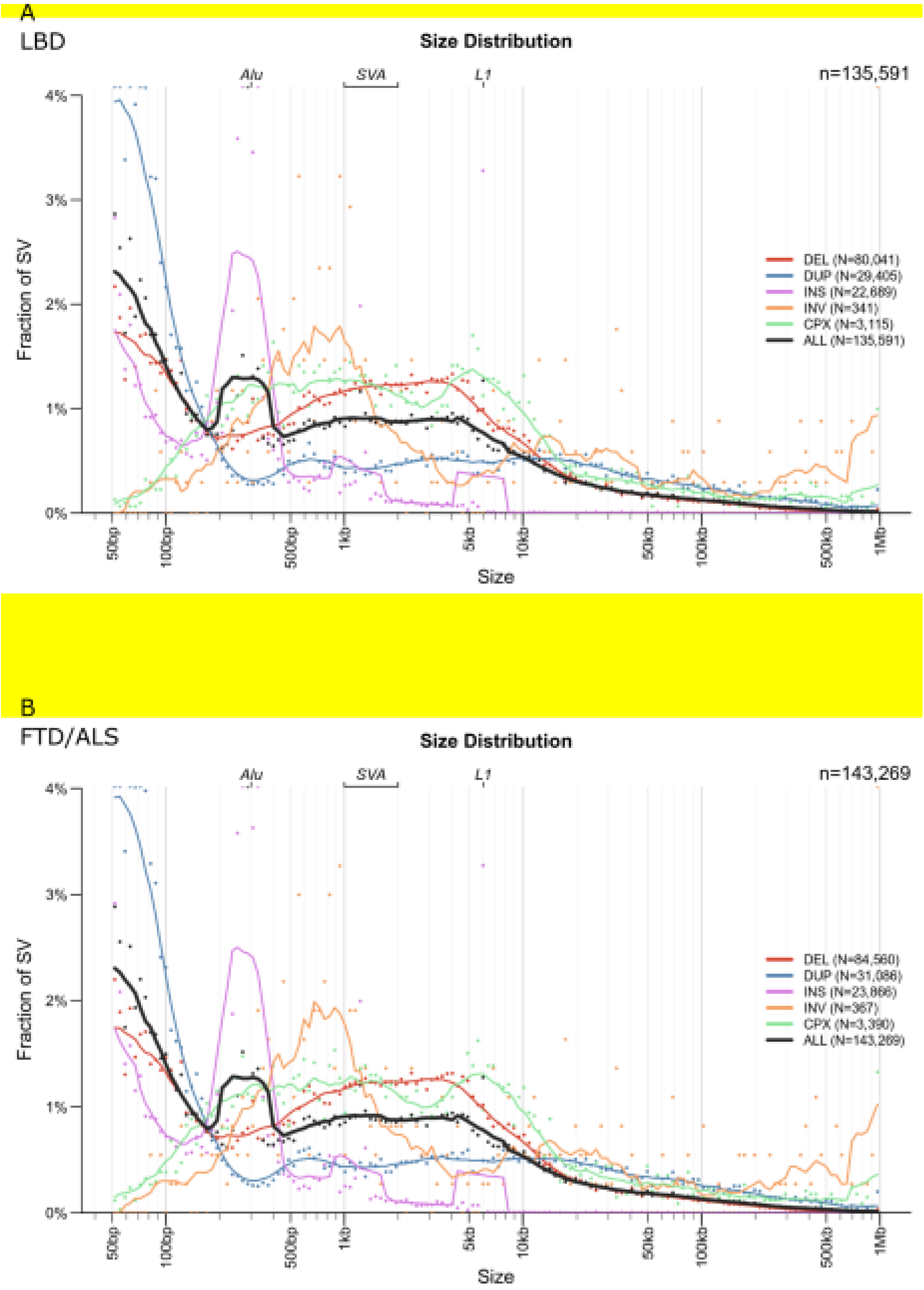
Response: Thank you for this suggestion. We now provide a violin plot summarizing the SV counts per genome as shown below (page 8, paragraph 3): "…In the FTD/ALS case-control cohort, there were 865 structural variants on average per participant (Figure 2 and Supplementary Figure 1-3 for descriptive statistics of filtered variants; Supplementary Figure 4-6 for summaries of unfiltered variants)." Supplementary Figure 1 | Structural variant counts per structural variant type



This figure shows the structural variant counts per variant type in the final filtered data used in the analyses. Panel **A** shows the results for the LBD case-control cohort, and **B** illustrates the FTD/ALS case-control cohort counts. The horizontal line in each violin plot represents the median value and μ refers to the mean. Abbreviations: DEL, deletions; DUP, duplications; INS, insertions; INV, inversions; CPX, complex structural variants; OTH, other structural variants.

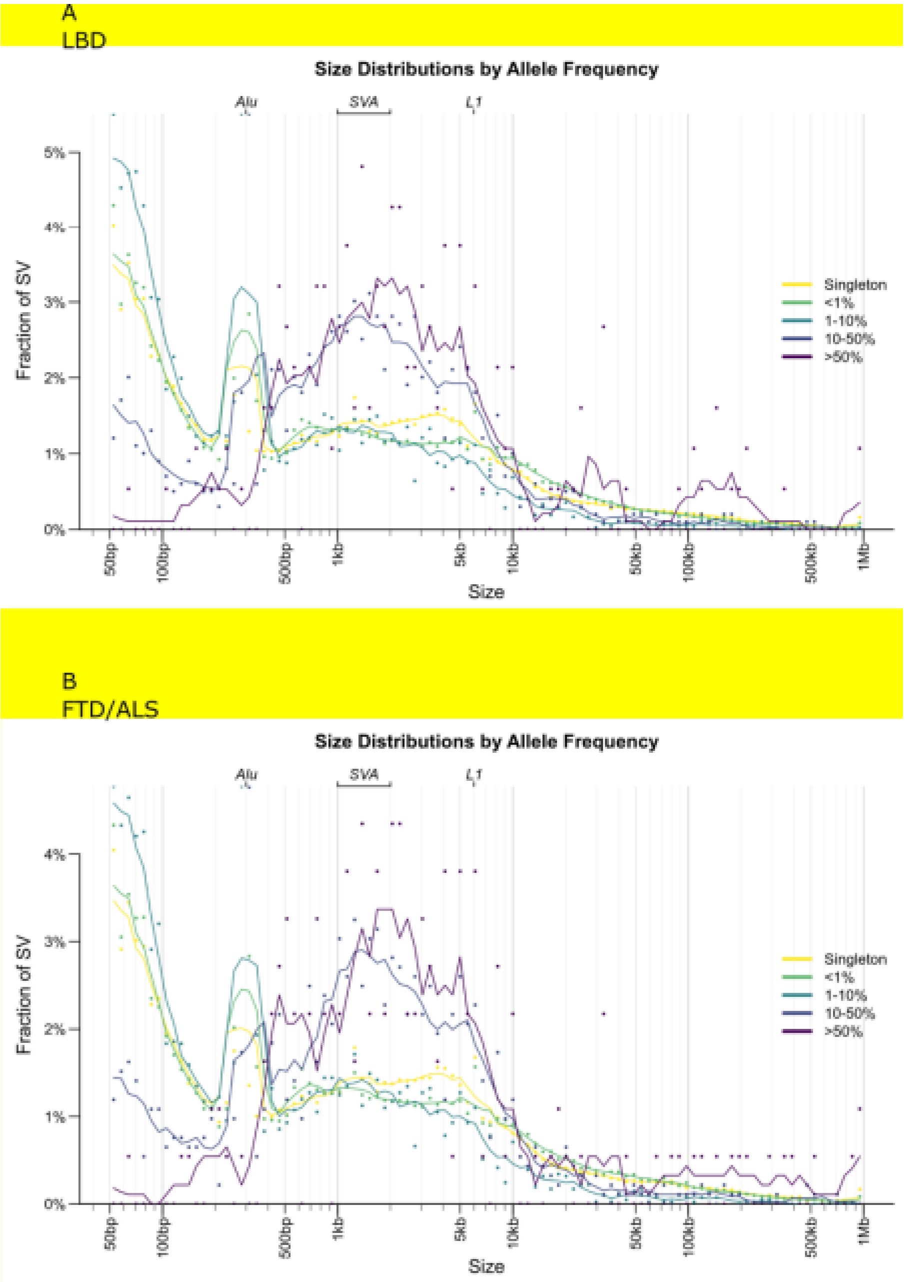
5. The stacked bar plots in Figure 2 make it difficult to compare allele frequency (AF) and size distributions by SV type. Instead, I would recommend using normalized distributions for the final SV call set, as demonstrated in Figure 3 of Billingsley et al.

Response: Thank you for this suggestion. We are now showing our results using a normalized distribution in Supplementary Figures 2 and 3 and updated our manuscript accordingly (page 8, paragraph 3): “In the FTD/ALS case-control cohort, there were 865 structural variants on average per participant (Figure 2 and Supplementary Figure 1-3, Supplementary Figure 4-6 for summaries of unfiltered variants).” Supplementary Figure 2 | Structural variant size per structural variant type



This figure shows the structural variant size for each variant type in the final filtered data used in the analyses. Panel **A** shows the results for the LBD case-control cohort, and **B** illustrates the FTD/ALS case-control results. Abbreviations: DEL, deletions; DUP, duplications; INS, insertions; INV, inversions; CPX, complex structural variants; OTH, other structural variants

Supplementary Figure 3 | Structural variant size and structural variant allele frequency



This figure shows the structural variant size per the structural variant allele frequency in the final filtered data used in the analyses. Panel **A** shows the results for the LBD case-control cohort, and **B** illustrates the FTD/ALS case-control results. Increased frequencies are observed at ~300 bp, ~1.2 kb, and ~6 kb, corresponding to abundant mobile elements in the human genome (Alu, SVA, L1).

Referees' report, second round of review

N/A

Authors’ response to the second round of review

N/A