

1 **A Specialized Reference Panel with Structural Variants Integration for Improving Genotype**
2 **Imputation in Alzheimer's Disease and Related Dementias (ADRD)**
3

4 Po-Liang Cheng^{1,2}, Hui Wang^{1,2}, Beth A Dombroski^{1,2}, John J Farrell³, Iris Horng², Tingting Chung²,
5 Giuseppe Tosto^{4,5}, Brian W Kunkle^{6,7}, William S Bush^{8,9}, Badri Vardarajan^{4,5}, Gerard D Schellenberg^{1,2},
6 Wan-Ping Lee^{1,2}
7

8 ¹Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of
9 Pennsylvania, Philadelphia, PA, USA, ²Penn Neurodegeneration Genomics Center, Perelman School of
10 Medicine, University of Pennsylvania, Philadelphia, PA, USA, ³Biomedical Genetics, Department of
11 Medicine, Boston University Medical School, Boston, MA, USA, ⁴Taub Institute for Research on
12 Alzheimer's Disease and the Aging Brain, College of Physicians and Surgeons, Columbia University, NY
13 10032, USA, ⁵Department of Neurology, College of Physicians and Surgeons, Columbia University and
14 the New York Presbyterian Hospital, NY 10032, USA, ⁶John P Hussman Institute for Human Genomics,
15 Miami, FL, USA, ⁷John T Macdonald Department of Human Genetics, Miami, FL, USA, ⁸Cleveland
16 Institute for Computational Biology, Cleveland, OH, USA, ⁹Department of Population and Quantitative
17 Health Sciences, Case Western Reserve University, Cleveland, OH, USA

18 **Summary**

19 We developed an imputation panel for Alzheimer's disease (AD) and related dementias (ADRD)
20 using whole-genome sequencing (WGS) data from the Alzheimer's Disease Sequencing Project (ADSP).
21 Recognizing the significant associations between structural variants (SVs) and AD, and their
22 underrepresentation in existing public reference panels, our panel uniquely integrates single
23 nucleotide variants (SNVs), short insertions and deletions (indels), and SVs. This panel enhances the
24 imputation of disease susceptibility, including rare AD-associated SNVs, indels, and SVs, onto genotype
25 array data, offering a cost-effective alternative to whole-genome sequencing while significantly
26 augmenting statistical power. Notably, we discovered 10 rare indels nominal significant related to AD
27 that are absent in the TOPMed-r2 panel and identified three suggestive significant (p -value $< 1E-05$)
28 AD-associated SVs in the genes *EXOC3L2* and *DMPK*, were identified. These findings provide new
29 insights into AD genetics and underscore the critical role of imputation panels in advancing our
30 understanding of complex diseases like ADRD.

31 Introduction

32 Genome-wide association studies (GWAS) aim to identify genomic variants linked to disease
33 risks or specific traits by analyzing the genomes of numerous individuals. GWAS seeks to identify
34 variants that occur more frequently in individuals with a particular disease compared to those without
35 it. GWAS primarily employs either whole-genome sequencing (WGS) or genotyping arrays to identify
36 genomic variants. Despite the rapid advancements and increasing affordability of WGS technology, it
37 still remains prohibitively expensive and computationally demanding for large-scale cohorts.
38 Consequently, genotype arrays provide a pragmatic and valuable tool due to their cost-effectiveness
39 and the availability of extensive disease data.

40 Genotype arrays assay variants relying on a pre-designed set of a small fraction of variants
41 chosen by the linkage disequilibrium (LD) structure of the human genome. Variants not directly
42 genotyped on arrays can be statistically inferred through a process called genotype imputation, which
43 compares variants in haplotypes to an external reference panel containing known haplotypes of a large
44 number of individuals, who have been genotyped using high-density genotype arrays or WGS. Usually,
45 imputation algorithms first estimate haplotypes between each individual in a study cohort utilizing
46 genotype arrays and a reference panel, and then use this information to infer missing alleles of the
47 individual. The accuracy of imputation depends on several crucial factors, including haplotype size, the
48 accuracy of genotypes in individuals, and the population diversity of the reference panel.

49 Currently, several public reference panels exist, such as the International HapMap Project¹, the
50 1000 Genomes Project (1000GP)², the UK10K Project³, the Haplotype Reference Consortium (HRC)⁴,
51 and the Trans-Omics for Precision Medicine (TOPMed) program^{5,6}. Among these, the TOPMed-r2 panel
52 stands out with its reference panel including 97,256 WGS samples, making it the largest reference

53 panel for genotype imputation to date⁶. The most recent version, TOPMed-r3, was released in
54 December 2023. However, during the experiments conducted for this study, only TOPMed-r2 was
55 available.

56 While public reference panels demonstrate high imputation accuracy in European populations,
57 their effectiveness is limited when applied to other ethnicity groups. Population-specific reference
58 panels, such as those tailored to Asian⁷ and African⁸ populations, show improved performance by
59 capturing recently evolved population-specific variants. Similarly, public reference panels, composed of
60 common populations, may potentially neglect rare variants in particular diseases. Therefore, we
61 hypothesize that utilization of disease-specific imputation panels may improve imputation accuracy for
62 disease studies.

63 Another rationale for the necessity of disease-specific imputation panels is that current public
64 reference panels either lack or have a limited number of structural variants (SVs) that have been
65 implicated in the association with human diseases^{9,10}. For example, the association of the inverted H2
66 haplotype with reduced risk of a range of neurodegenerative diseases¹¹, an 18Kb copy number
67 variation in *CR1* was found to associate with AD¹² and an 8 kb deletion upstream of *CREB1* is also
68 associated with AD¹³. Recent advancements underscore the importance of SV imputation. One study
69 developed a multi-ancestry SV imputation panel using long-read sequencing data of 888 samples from
70 1000GP¹⁴, and another study on the *CYP2A6* gene emphasized genotyping and imputing known and
71 novel SVs to understand genetic influences on traits like nicotine metabolism¹⁵. These findings
72 illustrate that incorporating SVs into imputation panels enhances the resolution and accuracy of
73 genetic association studies, providing deeper insights into the genetic underpinnings of complex
74 diseases such as AD. By capturing a broader spectrum of genetic variation, including SVs, disease-

75 specific imputation panels offer a more comprehensive tool for genomic research, facilitating better
76 disease risk prediction and understanding of disease mechanisms.

77 The Alzheimer's Disease Sequencing Project (ADSP) is a collaborative research effort,
78 sequencing diverse individuals across populations. The ADSP Release 3 (R3) 17K contains 16,905
79 samples with WGS data. Leveraging data from ADSP, we built ADSP-Short-Var (single nucleotide
80 variants [SNVs] and short insertion/deletions [indels]) and ADSP-All-Var (SNVs, indels, and SVs)
81 reference imputation panels, tailored to capture AD-enriched variants, particularly for SVs. We
82 demonstrated the strengths of these specialized panels by applying them to genotype data of 38,271
83 subjects of multiple ethnicities from the Alzheimer's Disease Genetics Consortium (ADGC).

84 **Results**

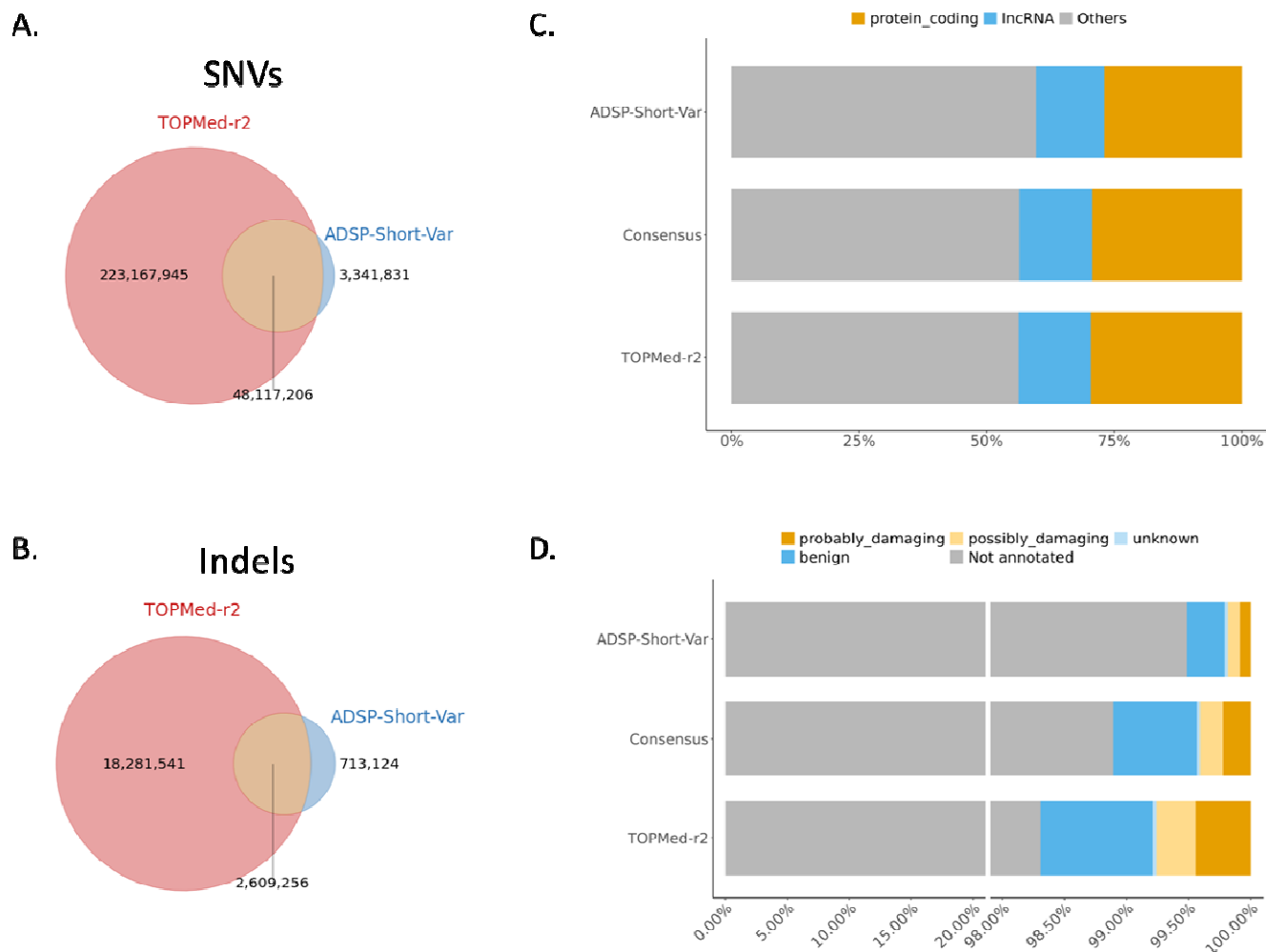
85 Overview of ADSP-Short-Var, ADSP-All-Var, and TOPMed-r2 Panels

86 The ADSP-Short-Var panel contained 54 million variants (51,459,037 [93.94%] SNVs and
87 3,322,380 [6.06%] indels in chromosomes 1-22) derived from 16,564 sequenced genomes,
88 representing diverse ethnic backgrounds including 62.76% non-Hispanic white, 18.68% Hispanic,
89 18.11% African American, 0.3% Asian, and 0.14% other ethnicities. In comparison, the TOPMed-r2
90 panel included 295 million variants (274,388,520 [92.85%] SNVs and 20,899,436 [7.15%] indels in
91 chromosomes 1-22) derived from 97,256 sequenced genomes, 48.49% European, 24.95% African,
92 17.57% admixed American, 1.22% East Asian, 0.66% South Asian, and 7.11% unassigned-ethnic
93 individuals (**Table S1**). We categorized variants into three categories for comparison: consensus
94 imputed variants shared between the imputations generated against the ADSP-Short-Var and TOPMed-
95 r2 panels, ADSP-Short-Var-specific imputed variants, and TOPMed-r2-specific imputed variants. Among
96 these variants, 50 million consensus variants were shared between the two panels (94.86% SNVs and

97 5.14% indels; **Figures 1A-B**). Panel-specific variants included 4 million variants (82.41% SNVs and
98 17.59% indels) unique to the ADSP-Short-Var panel and 241 million variants (92.42% SNVs and 7.57%
99 indels) unique to the TOPMed-r2 panel.

100 Using Variant Effect Predictor¹⁶ (VEP v105.0) for annotation, 29.32% of consensus variants were
101 situated in protein-coding regions. For ADSP-Short-Var-specific and TOPMed-r2-specific variants, the
102 percentages were 26.98% and 29.64%, respectively (**Figure 1C**). The second most common biotype
103 observed was Long Non-Coding RNA (LncRNA), accounting for 14.26%, 13.38%, and 14.04% of
104 consensus, ADSP-Short-Var-specific, and TOPMed-r2-specific variants, respectively (**Figure 1C**).
105 Regarding variants with PolyPhen, the prediction labeled by possibly damaging and probably damaging,
106 we observed 0.41%, 0.18%, and 0.76% in consensus, ADSP-Short-Var-specific, and TOPMed-r2-specific
107 variants, respectively (**Figure 1D**).

108 The ADSP-All-Var panel was constructed by integrating SNVs and indels in the ADSP-Short-Var
109 panel along with the SVs (231,385 deletions, 119,648 insertions, 45,839 duplications, and 3,362
110 inversions) identified on the same sample set in our previous study¹⁷. This integration enriched a
111 diverse genomic landscape in the ADSP-All-Var panel, with proportions of 93.25% SNVs, 6.02% indels,
112 0.42% deletions, 0.22% insertions, 0.08% duplications, and 0.01% inversions, respectively. The
113 incorporation of SVs into the ADSP-All-Var panel enables a more comprehensive discovery of variants
114 associated with AD on large sample sets with genotype data.



115

116 **Figure 1.** Comparison of variants between TOPMed-r2 and ADSP-Short-Var panel. (A) Venn diagrams
 117 showing the number of Single Nucleotide Variants (SNVs). (B) Venn diagrams showing the number of
 118 insertions/deletions (Indels). (C) Distribution of annotated biotypes. (D) PolyPhen predictions for
 119 TOPMed-r2 specific, ADSP-Short-Var specific, and consensus variants.

120

121 Discovery of Novel Suggestive Significant and Disease Susceptibility SVs Through Imputation

122

123 In an effort to enhance the statistical power of SV analysis, we performed imputation on the
 ADGC genotype dataset (Ncase=16,779 and Ncontrol=21,492) against the ADSP-All-Var panel. By

124 increasing the sample size, we aimed to uncover novel significant SVs. Our subsequent single variant
125 association test on the imputed SVs revealed three suggestively significant (p-value < 1E-05) SVs
126 (**Figure S1**).

127 The most notable discovery was an Alu insertion in the intron of *EXOC3L2*, exhibiting a p-value
128 of 1.78E-07, with allele frequencies (AF) of 0.01632 in AD cases and 0.01061 in controls. The further
129 experimental validation by PCR also confirmed this insertion (**Figure S2**). This insertion is also present
130 in the gnomAD database with an AF of 0.01275, similar to the AF of controls in our dataset. Another
131 significant SV identified was a deletion at chr19:45775716 (p-value = 9.94E-06) located in intron 8 of
132 *DMPK*. However, this region is complex, containing multiple Alu elements, such as AluSx1, AluJo, AluSz,
133 AluSx3, AluY, and AluJb, which may affect the quality of the deletion call. It is crucial to note that the
134 significance of the insertion at chr19:45216933 and the deletion at chr19:45775716 diminished under
135 the condition of *APOE e4*, suggesting the confounding impact of the SVs and *APOE e4*. **Table 1** provides
136 detailed information on these findings.

137 In the previous study on ADSP R3 17K WGS¹⁷, 107 SVs (72 deletions, 20 duplications, and 15
138 insertions) were reported. We found that 97.20% (104 out of 107) SVs were successfully imputed by
139 ADSP-All-Var panel. With this larger sample set, 65.38% (68 out of 104) SVs exhibited an increase in
140 allele count, and 39.71% (27 out of 68) showed enhanced statistical significance (**Table S2**). All the well-
141 imputed SVs had highly similar AFs ($r = 0.9931$) to SVs discovered from ADSP (**Figure S3**).

142 Among these SVs, some were specifically located in important AD genes and were in the same
143 LD block with known SNVs, which facilitated quality of the SV imputation of SVs. For instance, a
144 5,505bp deletion at chr2:105731359 in the upstream of *NCK2* ($R^2 = 0.7$), which is a gene highly
145 expressed in amyloid-responsive microglial cells, was in the same LD block with the known SNV

146 rs143080277, associated with late onset AD¹⁸. Similarly, a 238bp deletion at chr17:46009357 in *MAPT*
147 ($R^2 = 0.98$), which encodes tau protein implicated in AD pathology, was in the same LD block with the
148 SNV rs8070723, which is associated with reduced risk of LOAD¹¹. This 238-bp deletion, located
149 between exons 9 and 10 on the H2 background, is commonly used to differentiate between H1 or H2
150 haplotypes.

151 **Table 1.** Three suggestive significant structure variants imputed by the ADSP-All-Var panel.

SV	Size	AF Case	AF Control	OR	P	Gene
chr19:45216933:INS	327	0.0163	0.0106	1.5393	1.78E-07	EXOC3L2
chr11:47775210:INS	113+	0.2934	0.2966	0.9892	1.86E-06	
chr19:45775716:DEL	641	0.4813	0.4654	1.0343	9.94E-06	DMPK

152

153 Discovery of Disease Susceptibility SNVs and indels Through Imputation

154 In the analysis of WGS data from the ADSP R3 NHW cohorts (Ncontrol = 2,601 and Ncase =
155 4,053), 69 exonic rare indels (MAF < 1%) suggestively associated with AD were identified. These indels
156 met the criteria with CADD > 20, p-value < 0.05 (Fisher's test), and odds ratio (OR) exceeding 1.5 or
157 under 0.5. Out of these 69 indels, 55 were confirmed through experimental validation. Given their
158 presence in our ADSP-Short-Var panel, we imputed these indels on genotype data from the ADGC NHW
159 cohorts (Ncontrol = 15,216 and Ncase = 13,182) to enhance statistical power by increasing sample size.
160 As a result, the nominal significance of a deletion, chr20-663684-CCGGCGGGGGT-C in the exon 2 of
161 *SCRT2*, increased with p-value from 0.0096 to 0.0034. *SCRT2* is a neuron-specific gene involved in
162 neuronal survival, neuronal migration, and neurogenesis during brain development¹⁹⁻²¹. A study
163 indicated that the *SCRT2* expression was altered after surgery in aged mice with impaired cognition²².
164 Of note, 10 out of 55 indels were absent in TOPMed-r2 imputation (**Table S3**), which are in genes,
165 *C12orf81*, *TOMM20L*, *FAM174B*, *NTN3*, *RGL3*, *PNKP*, *PPDPF*, and *PCDHB13*.

166 To evaluate the accuracy of imputed genotypes for those disease susceptibility indels, PCR
167 validation was conducted on six indels, which are located in *DNAH14*, *ANO7*, *ZNF655*, *PTGER1*, *SCRT2*,
168 and *PPDPF*, on 17 available DNA samples from the ADGC cohorts (**Table S4**). The results revealed that
169 86.67% (13/15) indels were accurately genotyped by the ADSP-Short-Var panel, while only 66.67%
170 (10/15) were accurately genotyped by the TOPMed-r2 panel.

171 We also investigated the 263 SNVs and 10 indels in *ABCA7*, previously discovered through the
172 association test of aggregate of rare coding variants²³ on the WGS data of ADSP R3 NHW cohorts.
173 Among these variants, when imputed them onto ADGC NHW genotype data, 23.81% (65 out of 273)
174 were well imputed by the ADSP-Short-Var panel, while 76.19% (208 out of 273) could not be imputed
175 due to their rarity ($AC < 5$). For those well-imputed SNVs and indels, 75.38% (49 out of 65) showed an
176 increase in allele counts as sample size expanded, and 67.35% (33 out of 49) increased statistical
177 significance (**Table S5**). Importantly, all well imputed SNVs and indels had similar AFs ($r = 0.9375$)
178 compared to AFs discovered from the WGS data of ADSP R3 NHW cohorts (**Figure S4**).

179

180 Assessment of Imputation Accuracy of SNVs and Indels Compared to WGS

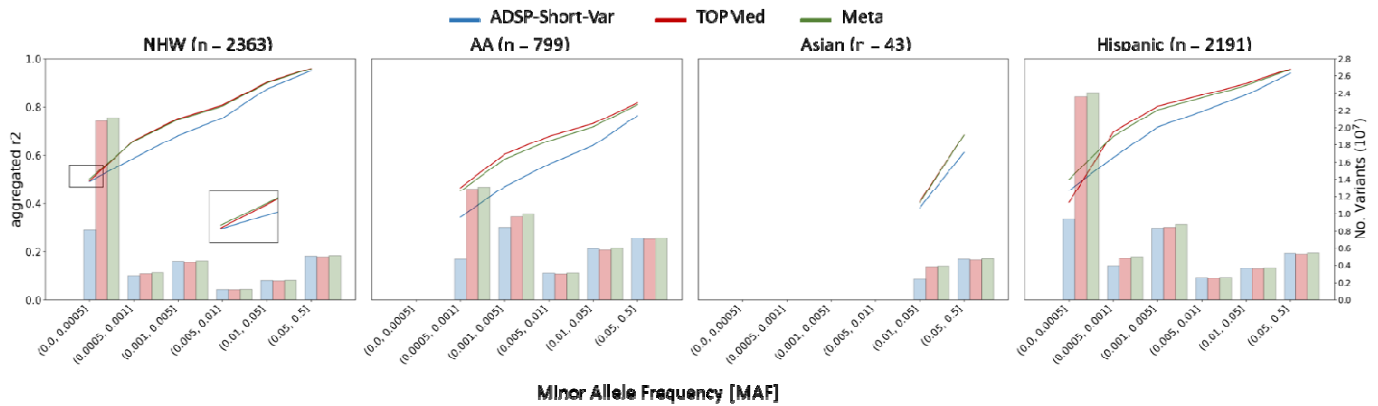
181 To evaluate imputation accuracy, we compared genotypes of SNVs and indels derived from
182 imputations with those obtained from WGS. This analysis was conducted on a dataset with 2,363 Non-
183 Hispanic White (NHW), 2,191 Hispanic, 799 African American (AA), and 43 Asian individuals with both
184 genotype and WGS data available. Notice that these samples were independent from the ADSP-Short-
185 Var and ADSP-All-Var panels.

186 Using aggRsquare^{24} , we reported an aggregated R^2 , the squared Pearson correlation between
187 genotypes obtained from imputations and WGS, across all SNVs and indels, stratified by minor allele

188 frequency (MAF) bins: <0.0005, 0.0005-0.001, 0.001-0.005, 0.005-0.01, 0.01-0.05, and >0.05. The
189 ADSP-Short-Var panel demonstrated improved performance over the TOPMed-r2 panel for variants
190 with MAF < 0.0005 in Hispanic cohorts and performed comparably well for variants with MAF < 0.0005
191 in NHW cohorts (**Figure 2, Table S6**). Due to limited sample sizes, we were unable to examine the
192 variants with MAF < 0.005 for AA and Asian cohorts. Conversely, the TOPMed-r2 panel outperformed
193 for variants with MAF > 0.005 (**Figure 2**).

194 Using the merge feature of Meta-Minimac2²⁴, a sophisticated algorithm designed to merge
195 imputations from multiple reference panels into a unified imputation, we obtained a meta-imputation
196 by merging imputations from the ADSP-Short-Var and TOPMed-r2 panels. The meta-imputation
197 improved the imputation quality for variants with MAF < 0.005, elevating aggregated R^2 from 0.4542
198 and 0.4045 (ADSP-Short-Var and TOPMed-r2; difference 0.0497) to 0.4995 (meta-imputation) for
199 Hispanic cohorts and from 0.4911 and 0.4935 (difference 0.0024) to 0.5014 for NHW cohorts.

200 Our analyses, however, revealed a nuanced performance landscape. While Meta-Minimac2
201 generally improved imputation quality when initial accuracies were closely matched, such as in the
202 MAF < 0.0005 bin for NHW cohorts, divergent outcomes were observed where substantial disparities
203 existed between the initial imputations. Specifically, the differences in the average aggregated R^2 for
204 MAFs > 0.005 were 0.0471 ± 0.0269 for NHW, 0.1031 ± 0.0324 for AA, 0.0492 ± 0.0340 for Asian, and
205 0.0652 ± 0.0340 for Hispanic cohorts, indicating decreased performance in these scenarios. These
206 findings underscore the potential of tailored reference panels in improving the imputation of rare SNVs
207 and indels and demonstrate that combining the strengths of different panels through Meta-Minimac2
208 can optimize imputations.



209
 210 **Figure 2.** Comparison of aggregated R^2 , the squared Pearson correlation between genotypes obtained
 211 from imputations and WGS, for four ethnicities. The bars indicate the total number of variants
 212 analyzed for each ethnicity.

213 Assessment of Imputation Accuracy of SVs Compared to WGS

214 Since aggRsquare is not well-suited for accurately assessing SVs, we calculated the precision
 215 and recall of SVs obtained from imputation compared to those identified by Manta²⁵ on WGS for each
 216 sample, using the same dataset utilized for SNV and indel imputation accuracy. We filtered SVs
 217 identified from WGS data with the label “PASS”, and for imputed SVs, we selected the high-quality SVs
 218 listed in the previous study¹⁷. Given that imputation quality (R^2) is indicative of imputation accuracy,
 219 we applied different R^2 thresholds (i.e., 0.2, 0.5, and 0.8) to filter the imputed SVs and compared them
 220 to SVs identified from WGS data. On average, there were $8,523.65 \pm 803.92$ SVs per sample on WGS,
 221 whereas high-quality imputed SVs on genotype array data is on average 11211.25 ± 303.14 ,
 222 10461.67 ± 298.75 , 9696.10 ± 290.25 and $5,418.22 \pm 242.01$ for R^2 filter set at 0, 0.2, 0.5 and 0.8 (**Figure**
 223 **S5**).

224 Both false positive (FP) and true positive (TP) SVs decreased as the R^2 threshold increased
 225 (**Table S7**). For FPs, the numbers of deletions, duplications, and insertions dropped from

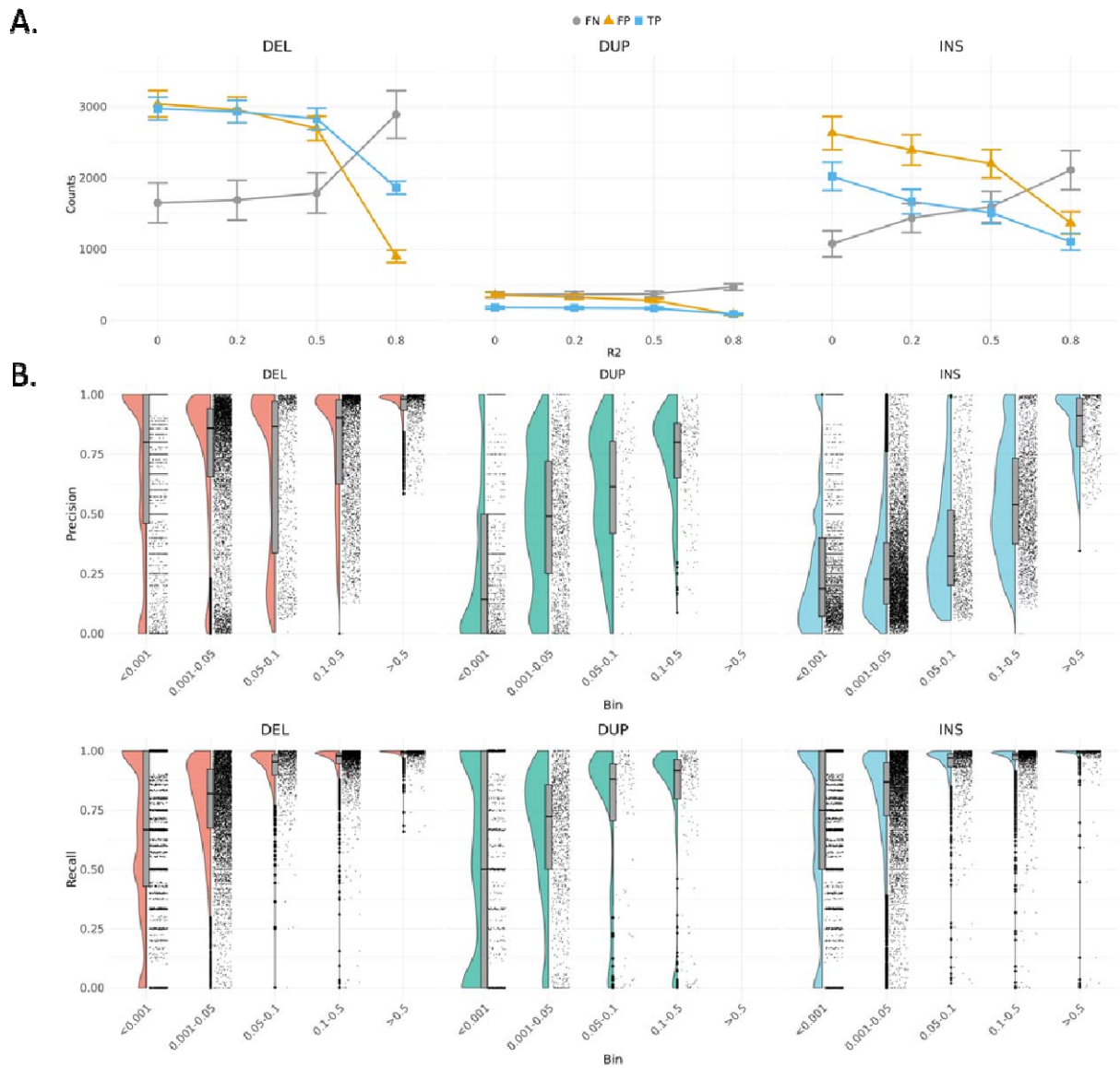
226 2,693.80±171.58, 282.69±23.71, and 2,199.62±197.65 with $R^2 > 0.5$ to 899.42±87.26, 84.00±12.01, and
227 1,368.21±156.96 with $R^2 > 0.8$, respectively. For TPs, the number of deletions, duplications, and
228 insertions dropped from 2,829.95±150.55, 171.90±16.47, and 1,510.85±157.17 with $R^2 > 0.5$ to
229 1,864.14±91.49, 90.50±10.31, and 1,104.55±118.25 with $R^2 > 0.8$, respectively (**Figure 3A**). We noted
230 that FPs dropped faster than TPs as R^2 increased, thereby improving precision (TP / (TP + FP)).
231 Deletions overall showed the higher precision among SV types. Specifically, average precisions were
232 0.6747±0.0267 for deletions, 0.5194±0.0441 for duplications, and 0.4473±0.0455 for insertions.

233 Conversely, the numbers of false negative (FN) increased substantially as R^2 increases. FNs rose
234 from 1,786.66±282.19, 372.36±40.11, and 1,596.07±217.25 with $R^2 > 0.5$ to 2,890.00±332.61,
235 469.52±44.31, and 2,107.35±273.75 with $R^2 > 0.8$ for deletions, duplications, and insertions,
236 respectively (**Figure 3A**). Deletions stood out again with higher recall (TP / (TP + FN)) compared to
237 other SV types, with average recalls of 0.3936±0.0204 for deletions, 0.1619±0.0149 for duplications,
238 and 0.3446±0.0166 for insertions (**Figure S6**). The higher precision and recall of deletions reflect the
239 better calling quality of deletions on short-read WGS compared to other types of SVs. To exhibit high
240 accurate SVs for downstream association analysis and functional validation, we set the threshold at R^2
241 > 0.8 to filter imputed SVs to obtain more promising outcomes in the analysis.

242 Regarding AF, we observed outstanding performance in imputing SVs with higher AF (**Figure**
243 **3B**). The average precisions were 0.9430±0.0833 for deletions and 0.8677±0.1318 for insertions with
244 MAF > 0.5 . Notice that we did not find any duplications with MAF > 0.5 . Deletions maintained higher
245 average precision in all AF bins, even at MAF < 0.001 (0.6742±0.3618). In contrast, precisions of
246 insertions decreased rapidly from MAF > 0.5 (0.8677±0.1318) to $0.1 < \text{MAF} < 0.5$ (0.5563±0.2330).
247 Similarly, precisions of duplications dropped gradually as the AF decreased. Both deletions and

248 insertions remained with high average recall at $0.05 < \text{MAF} < 0.1$, with average recall values of
249 0.9233 ± 0.0963 for deletions, 0.9337 ± 0.1164 for insertions, and 0.7275 ± 0.3301 for duplications.
250 Further investigation revealed that the performance of imputing SVs is poorly correlated with SV
251 lengths (**Figure S7**).

252



253

254 **Figure 3. A.** The changes of true positive, false positive and false negative among different R^2
255 threshold. **B.** The precision and recall of deletions, duplications, and insertions across different allele
256 frequency.

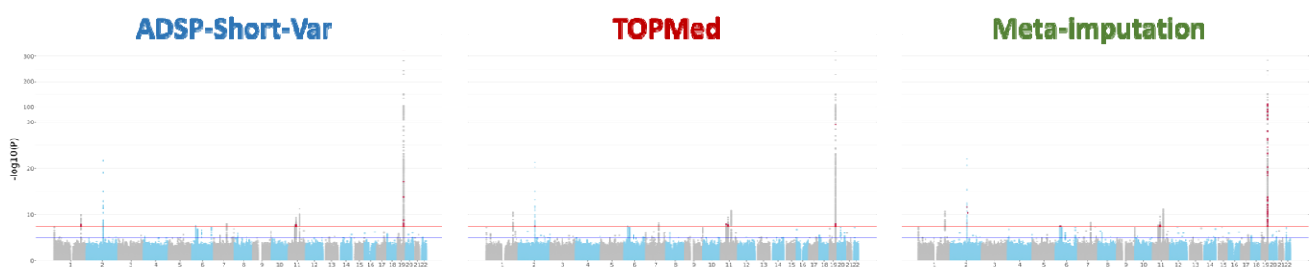
257 Association Analyses on Different Imputations

258 In the ADGC genotype dataset (Ncase=16,779 and Ncontrol=21,492), we conducted single
259 variant association tests on three distinct imputations, derived from the ADSP-Short-Var, TOPMed-r2,
260 and meta-imputation generated by Meta-Minimac2, to evaluate the impact of various reference
261 panels. Our analysis on pooled samples across ethnicities revealed eight genome-wide significant loci,
262 including 495 genome-wide significant variants across 36 genes, that were concordant in all three
263 association tests (**Table S8**). Regarding discordant signals, we found 32, 16, and 71 genome-wide
264 significant variants uniquely identified in the ADSP-Short-Var, TOPMed-r2, and meta-imputation tests,
265 respectively (**Figure 4, Figure S9A**). Notice that no novel significant locus was identified by those
266 discordant variants. Furthermore, 18 suggestive significant variants in the tests on imputations from
267 the ADSP-Short-Var (average p-value $9.76E-08 \pm 4.37E-08$) and TOPMed-r2 (average p-value $6.85E-$
268 $08 \pm 2.03E-08$) panels showed increased significance in the test on meta-imputation (average p-value
269 $3.86E-08 \pm 9.90E-09$).

270 Ethnicity-specific analysis in NHW cohorts (Ncase=13,182 and Ncontrol=15,216) revealed eight
271 genome-wide significant loci, including 765 genome-wide significant variants across 44 genes, in all
272 three association tests (**Table S8**). Six of these loci, located in chromosomes 1, 2, 11, and 19, were also
273 identified in the pooled-sample analysis. Two loci on chromosomes 1 and 3 emerged, suggesting that
274 the approach for pooled samples might obscure specific genetic signals due to differences in
275 population genetic structures (**Table S9**). For the panel-specific genome-wide significant variants, there
276 were unique 36, 29, and 79 genome-wide significant variants from the ADSP-Short-Var, TOPMed-r2
277 and meta-imputation tests, respectively (**Figures S8A,S9B**). Most of the panel-specific suggestive
278 significant variants were at the borderline of genome-wide significance.

279 In AA cohorts (Ncase=1,795 and Ncontrol=3,784), three consensus loci, including 45 variants in
280 nine genes at chromosomes 19, 21, and 22, were identified in all three association tests. An additional
281 genome-wide significant locus with 15 significant variants in AC019063.4 on chromosome 7 was
282 detected in the TOPMed-r2 imputation, but the locus vanished after meta-imputation. We also found
283 that ADSP-Short-Var panel had better imputation quality ($R^2 = 0.882 \pm 0.0288$) than TOPMed ($R^2 =$
284 0.865 ± 0.0415) for the 15 genome-wide significant variants on chromosome 7. This finding showed that
285 the ADSP-Short-Var panel could help refine the imputation results through meta-imputation. The
286 number of variants unique to each test was 7 for ADSP-Short-Var, 21 for TOPMed-r2, and 14 for meta-
287 imputation (**Figures S8B and S9C**).

288 No genome-wide significant loci were observed in the Asian (Ncase=1,576 and Ncontrol=1,951)
289 and Hispanic (Ncase=226 and Ncontrol=541) cohorts, except for *APOE-ε4* SNV rs429358, which was
290 observed genome-wide significantly associated with AD in the TOPMed-r2 test (**Figures S8C-D and S9D-**
291 **E**). Overall, the results of single variant association tests indicated that the ADSP-Short-Var and
292 TOPMed-r2 imputations were largely similar. Enhancing suggestive significant signals demonstrated
293 the potential for optimizing imputation results through meta-imputation.



294
295 **Figure 4.** Single variant association tests performed on different imputations against the ADSP-Short-
296 Var and TOPMed-r2 panels and meta-imputation. Red dots represent the genome-wide significant
297 signals uniquely present in each imputation.

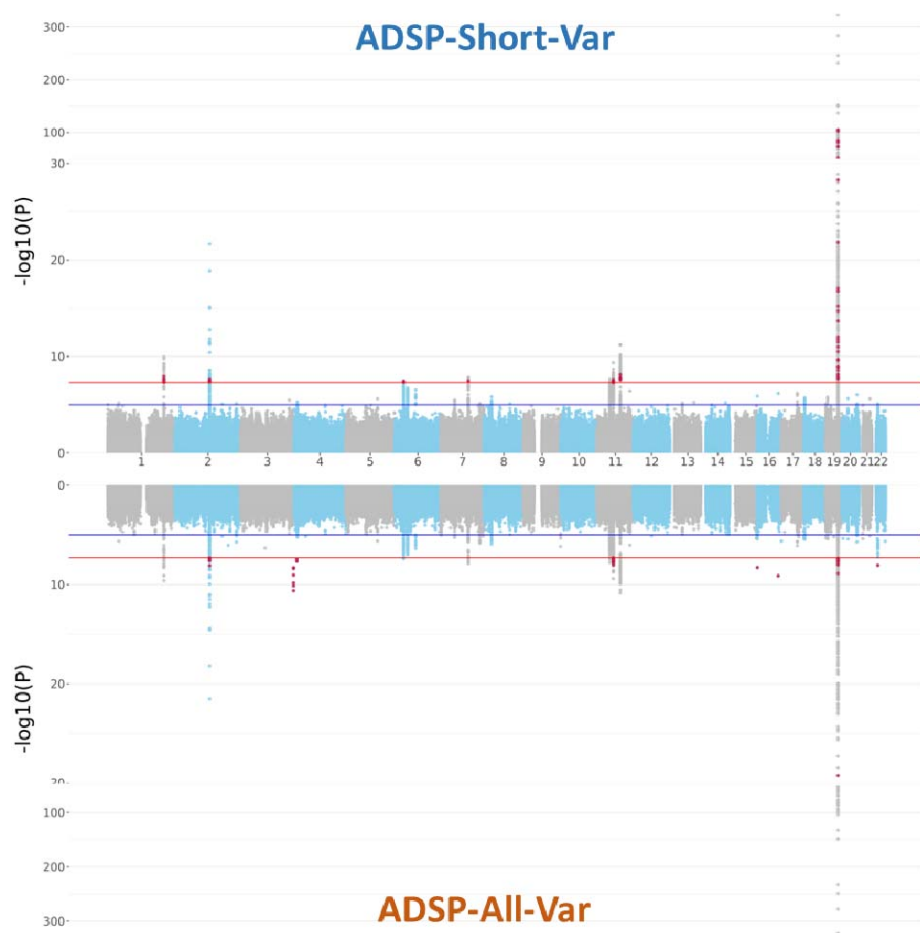
298 Investigation of the Impact of SV Integration in Reference Panel

299 To evaluate the impact of integrating SVs into the ADSP-Short-Var panel for the ADSP-All-Var
300 panel, we assessed the imputation accuracy of these two panels. We utilized a dataset containing
301 samples with both genotype and WGS data available, performing imputations against the ADSP-Short-
302 Var and ADSP-All-Var panels separately. The imputation accuracy was assessed by calculating
303 aggregated R^2 values between SNVs from imputations and WGS, revealing nearly identical imputation
304 accuracies (**Figure S10**).

305 Upon comparing the results of single variant association tests on pooled samples
306 (Ncase=16,779 and Ncontrol=21,492) imputations from the ADSP-Short-Var and ADSP-All-Var panels,
307 we identified eight consensus genome-wide significant loci across 36 genes (**Table S8**) in both tests
308 (**Figure 5**). In details, there were 75 and 139 genome-wide significant variants uniquely from ADSP-
309 Short-Var and ADSP-All-Var, respectively. For the ADSP-All-Var specific variants, 93.53% (130 out of
310 139) were located in the eight consensus genome-wide significant loci. Two additional genome-wide
311 significant loci were discovered in chromosomes 3 and 22 from the imputation against the ADSP-All-
312 Var panel.

313 Compared to the ADSP-Short-Var panel, we observed that the ADSP-All-Var panel altered 8.30%
314 of haplotypes defined by the six genome-wide significant variants on chromosome 3 and 3.06% of
315 haplotypes defined by one genome-wide significant variant along with 48 suggestive significant
316 variants on chromosome 22. These changes may result in differing outcomes when the same
317 covariates are adjusted in association tests between imputations from the two panels, despite the AFs
318 being quite similar (**Table S10**). Additionally, all the signals on chromosome 3 were within the same LD
319 block, while all the signals on chromosome 22 were within the same LD block (**Figure S11**). Including

320 SVs into SNV and indel panel did not dramatically alter LD structure but might have caused part of the
321 haplotypes to change while inferring haplotypes from genotype data.



322
323 **Figure 5.** Single variant association test of imputations performed by ADSP-Short-Var and
324 ADSP-All-Var reference panels. Red dots represent the genome-wide significant signals uniquely
325 present in each imputation.

326 Discussion

327 In this study, we constructed a reference panel of 16,564 whole-genome sequenced genomes
328 from ADSP R3 with diverse populations including NHW, AA, Asian, and Hispanic to provide high-quality
329 reference panels for ADRD research. To assess the performance of our reference panels, we performed

330 imputation on ADGC datasets and compared to it from the public reference panel, TOPMed-r2. Our
331 panel captured several rare but potential causal indels that were missed by TOPMed-r2. Furthermore,
332 imputation from our panel provided high-quality SVs that were absent in TOPMed-r2.

333 We identified 3 suggestive AD associated SVs that located in two genes, *EXOC3L2* and *DMPK*.
334 *EXOC3L2* is a component of the exocyst complex, involved in the regulation of the readily releasable
335 pool of synaptic vesicles via the binding of NSF and SNARE proteins²⁶. A variant rs597668 near *EXOC3L2*
336 is identified as a risk factor for AD in European population²⁷, but played a protective role in AD in East
337 Asian population²⁸ in previously studies. The suggestive AD associated SV in *EXOC3L2* that we identified
338 was also recognized as a risk factor in the NHW population and as a protective factor in the Asian
339 population. *DMPK* is a serine/threonine kinase that could prevent ROS-induced cell death²⁹, and its
340 gene mutations cause myotonic dystrophy type 1 (DM1)³⁰. A study indicated that the expression of
341 *DMPK* in the brain follows an age-related pattern³¹, but its role in aging or in AD is still unknown.

342 There were 10 out of 55 rare indels absent in TOPMed-r2 imputation, which are located in
343 genes, *C12orf81*, *TOMM20L*, *FAM174B*, *NTN3*, *RGL3*, *PNKP*, *PPDPF*, and *PCDHB13*. *NTN3* (netrin-3) is a
344 member of the netrin family, a kind of extracellular protein that directs cell and axon migration during
345 embryogenesis³² and is highly expressed in sensory ganglia³³. This protein family includes the other
346 famous protein netrin-1 which is highly correlated with A β levels in the brain tissue of AD patients³⁴⁻³⁶.
347 *PNKP* (Polynucleotide Kinase-Phosphatase) is involved in DNA repair processing³⁷, that might associate
348 with AD pathogenesis and its dysfunction of this gene can result in microcephaly or
349 neurodegeneration³⁸. The *PPDPF* was predicted to be involved in cell differentiation and mainly
350 expressed in oligodendrocytes based on the data from the human protein atlas. It was downregulated

351 in the dorsolateral prefrontal cortex of AD patients comparing to people with NCI (no cognitive
352 impairment) or MCI (mild cognitive impairment)³⁹.

353 Each reference panel offers unique strengths, leading us to employ Meta-Minimac2 for meta-
354 imputation. This tool integrates imputations from multiple reference panels, leveraging their collective
355 strengths to enhance imputation accuracy. Our application of Meta-Minimac2 improved imputation
356 results for ultra-rare variants in the NHW and Hispanic ethnic groups. However, in AA and Asian
357 groups, the accuracy of meta-imputation was not improved. This suggests that meta-imputation does
358 not universally enhance accuracy, particularly when substantial discrepancies exist between initial
359 imputation results. Despite these challenges, meta-imputation still enabled the identification of AD-
360 specific genotypes absent in TOPMed-r2. From single variant association tests, we found that most
361 genome-wide significant signals were consistent across imputations from the ADSP-Short-Var and
362 TOPMed-r2 panels, and through meta-imputation. However, meta-imputation introduced some noise
363 signals, which lacks LD support.

364 In this study, we expanded our methodologies by constructing a reference panel integrated
365 with SVs. Previous research has applied similar approaches to impute SVs for general populations and
366 cardiometabolic traits^{40,41}; however, these studies did not specifically address AD. Efforts have been
367 made to use high-quality SV datasets to improve SV imputation on genotype data. Our inclusion of
368 high-quality SVs into the imputation panel resulted in commendable imputation quality. Nevertheless,
369 the existing pipeline fails to address potential conflicts among SNVs, indels, and SVs. For example, SNVs
370 should not be present within homozygous deletions. This oversight indicates a clear necessity for novel
371 phasing or imputation methods specifically designed for SVs. Ultimately, our tailored reference panel

372 promises to significantly advance genetics research in Alzheimer's Disease and Related Dementias
373 (ADRD), especially concerning rare variants and SVs.

374 **Materials and Methods**

375 Whole Genome Sequence Samples From ADSP

376 Alzheimer's Disease Sequencing Project (ADSP)⁴² is a collaborative project aiming at identifying
377 new variants, genes, and therapeutic targets in AD. Data from the ADSP are available to qualified
378 investigators via the National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site
379 (NIAGADS) (<https://dss.niagads.org/>). This work focused on participants with WGS in the NIAGADS data
380 named "R3 17K WGS Project Level VCF", which contains 16,905 subjects (6,646 AD cases, 6,938
381 controls and 3,321 subjects with unknown AD status) collected across 24 cohorts and whole genome
382 sequencing was performed by Illumina HiSeqX, HiSeq2000, HiSeq2500, and NovaSeq platforms. The
383 ADSP dataset included 10,517 non-Hispanic white, 3,018 African American, 3,296 Hispanic white, 50
384 Asian and 24 unknown or other ethnicities.

385 From the initial pool of 16,905 individuals of the ADSP R3 17K, we first removed 341 related
386 samples through the identity by descent (IBD) analysis ($PI_HAT > 0.4$). Subsequently, we conducted a
387 rigorous variant quality control process, starting with the assessment of Hardy-Weinberg Equilibrium
388 (HWE) using RUTH⁴³ with the top 10 principal components (PCs) that calculate by plink in control
389 subjects and filtering variants violating the principle ($SLE_P_I < -4$). Variants with an allele count (AC)
390 less than 5 and a missing genotype rate exceeding 90% were also removed. This stringent filtering
391 resulted in a final dataset of 16,564 sequenced genomes with 51,459,037 SNVs and 3,322,380 indels
392 for a reference panel construction (**Figure S12A**).

393 Genotype Array Samples From ADGC

394 The Alzheimer's Disease Genetics Consortium (ADGC) is a collection of GWAS data funded by
395 the NIH, aiming to collaboratively use the collective resources of the AD research community to resolve
396 Alzheimer's disease (AD) genetics. In total, 51 available cohorts with 21,492 control and 16,779 AD
397 cases were used in this study. There were 15,176 male and 23,095 female in this dataset. This dataset
398 consists of 4 ethnicities that include 28,398 non-Hispanic white, 5,579 African American, 3,527 Asian
399 and 767 Hispanic samples.

400 Whole Genome Sequence Data Process

401 The Genome Center for Alzheimer's Disease (GCAD) mapped short reads against the reference
402 genome hg38 using BWA MEM⁴⁴, called SNVs and indels using the GATK HaplotypeCaller V2.6⁴⁵ for
403 each sample, and then performed joint genotyping across all samples using GATK. The GCAD quality
404 control (QC) working group performed quality checks of variants and genotypes and assigned a quality
405 annotation⁴⁶.

406 The SV callset is available on NIAGADS as well¹⁷ For each sample, Manta²⁵ v1.6.0 and Smoove
407 v0.2.5 (<https://github.com/brentp/smoove>) with default parameters were used. Calls from Manta and
408 Smoove were merged by Svimmer⁴⁷ to generate a union of two call sets for a sample. Unresolved non-
409 reference 'breakends' (BNDs) and SVs > 10 Mb were filtered. Then, all individual sample VCF files were
410 merged together by Svimmer as input to GraphTyper2⁴⁸ v2.7.3 for joint genotyping. This study utilizes
411 the SVs from the callset¹⁷.

412 Workflow of Reference Panel Building

413 We first phased 51,459,037 SNVs and 3,322,380 indels derived from the 16,564 whole genome
414 sequenced genomes by SHAPEIT4⁴⁹, followed by converting the vcf format into m3vcf using
415 minimac3⁵⁰. At last, we could get the ADSP-Short-Var panel. In order to extend our research into SVs,
416 we augmented our reference panel to include not only SNVs and indels but also SVs. Leveraging SV
417 callset obtained from our previous study¹⁷ on ADSP R3 WGS data, we selected and incorporated
418 231,385 deletions, 119,648 insertions, 45,839 duplications and 3,362 inversions were selected and
419 incorporated into ADSP-Short-Var panel construction. The high-quality SVs were merged with a
420 stringent filtered ADSP 17K dataset. Then phased the dataset, which contained SNVs, indels, and SVs,
421 by SHAPEIT4 and then turned the vcf format into m3vcf by minimac3⁵⁰. After this process, we obtained
422 the ADSP-All-Var panel.

423 Workflow of Genotype Imputation

424 The imputation strategy was shown in **Figure S12B**. Total 51 ADGC cohort was first phased by
425 SHAPEIT4-4.2.2⁴⁹ and imputed on the TOPMed imputation server⁶ by TOPMed-r2 panel. The phased
426 datasets also imputed to ADSP-Short-Var panel and ADSP-All-Var panel by minimac4-1.0.2⁵⁰. Then we
427 utilized metaminimac2-1.0.0²⁴ to combine imputation results generated using TOPMed and ADSP-
428 Short-Var panel into a consensus imputed dataset. To merge all imputed cohorts of each imputation,
429 the imputation quality scores (R^2) were calculated and combined using Fisher z-transformation and
430 generated lists of excluded and retained variants from information files (.info.gz) by IMMerge⁵¹. We
431 removed SNVs which R^2 labeled as NA in information files that were generated by TOPMed imputation
432 server in order to avoid the failed calculation of Fisher z-transformation before the merging process

433 start. At last, 38,271 samples with known AD status were selected from the merged cohort to form the
434 final dataset.

435 Single variant association analysis

436 For the single variant association test, variants with R^2 over 0.8, MAF over 0.5%, and the
437 HWE_SLP_I value range from -4 to 4 were used in the task. We used a R package GENESIS⁵² v. 2.28.0 to
438 perform single variant and structure variant association test with an additive genotype model adjusting
439 for age, sex and population substructure using top 10 principal components.

440 Imputation accuracy and quality measurement

441 Imputation accuracy was determined by comparing genotypes from imputation to genotypes
442 from WGS. The WGS data of 36,361 individuals of the ADSP R4 36K were utilized to evaluate the
443 imputation accuracy of imputed genotypes. The variants were called by GATK⁴⁵ v.4.1.1 and SVs were
444 generated by manta. We utilized 5396 samples (2363 non-Hispanic white, 2191 Hispanic, 799 African
445 American, 43 Asian) which both had genotype array data and WGS data, independent to samples used
446 in building panel, for evaluating the imputation accuracy. The validation samples were selected from
447 each imputed cohort and merge together by ethnicity using bcftools⁵³ for the three imputations. For
448 each ethnicity, all three imputations were compared to WGS data through calculating aggregated r^2 by
449 aggRsquare^{24} , which is calculated as the squared Pearson correlation between the imputed genotypes
450 and the WGS genotypes. Imputation quality was determined by R^2 score, that generated from
451 Minimac4. The threshold of well-imputed variants was setting at the R^2 over 0.8.

452 Imputed SVs with R^2 over 0.8 were kept for validation. SVs callset were from Manta²⁵. The same
453 SVs were defined by the covered region of each structure variant in imputed SVs reciprocal overlap

454 more than 50% with the SVs in validation call set. BEDTools⁵⁴ was used to intersect the SVs. For each
455 sample, the SVs discovered in both imputation and validation dataset were deemed as true positive.
456 The SVs only discovered in imputation dataset were defined as false positive, in contrast, the SVs only
457 discovered in validation dataset were defined as false negative. The precision of each SVs was
458 calculated by the number of all true positive in the SV divide the sum of the number of the true
459 positive and the number of false positive in the same SV. On the other hand, the recall of each SVs was
460 calculated by the number of all true positive in the SV divide the sum of the number of the true
461 positive and the number of false negative in the same SV. The allele frequency of validation dataset
462 was used to assign SVs to specific allele frequency bin.

463 PCR validation

464 For variant's genotyping, primers were designed at 200bp upstream and downstream of the
465 target position. 50ng Genomic DNA was amplified by SimpliAmp Thermal Cycler (Applied Biosystems)
466 in a 20ul reaction volume with HotStarTaq Master Mix (Qiagen) in the presence of 2uM primers (IDT).
467 PCR was performed at: 95°C for 15min; 30 cycles at 95°C for 20sec, 55°C for 30sec, 72°C for 2min; with
468 a final extension of 72°C for 7min. The amplified target sequences were cleaned up with ExoSAP-IT
469 (USB) by incubating at 37°C for 45min followed by 80°C for 15min. The target sequences after being
470 cleaned up, were then used to perform Sanger sequencing by using the BigDye® Terminator v3.1 Cycle
471 Sequencing kit (Part No. 4336917 Applied Biosystems) at: 96°C for 1min; 25 cycles at 96°C for 10sec,
472 50°C for 5sec, 60°C for 1min15sec. The products were then cleaned up by using XTerminator and SAM
473 Solution (Applied Biosystems) with 30min of shaking at 1800rpm. The sequencing products were
474 analyzed on a SeqStudio Genetic Analyzer (Applied Biosystems) and the sequencing traces were
475 analyzed using Sequencher 5.4 (Gene Code).

476 Statistical analysis

477 For comparing the rare variants and SVs of ADSP dataset to imputations of ADSP, associations
478 of case and control were calculated by Fischer's exact test. Pearson correlation was used to estimate
479 the correlation of AFs between ADSP-Short-Var imputation and AFs discovered from ADSP R3 WGS
480 data. A P value that less than 0.05 was determined as nominal significant. All statistical analyses were
481 performed in R.

482 **Acknowledgements**

483 See supplementary text. PLC, HW, IH, and TC report grant support from RF1-AG074328. WPL
484 reports grant support from RF1-AG074328, P30-AG072979, U54-AG052427, and U24-AG041689.

485 **Author contributions**

486 PLC, HW, and IH performed statistical analyses. PLC and HW performed phenotype acquisition
487 and/or harmonization. PLC, HW, and WPL performed Genotype acquisition and/or QC. BAD, PLC, and
488 GDS performed experimental validation. PLC, HW, JJF, IH, TC, GT, BWK, WSB, BV, GDS, and WPL
489 interpreted results. PLC and WPL wrote the first draft of the manuscript. All authors read, critically
490 revised, and approved the manuscript.

491 **Data availability**

492 <https://github.com/whtop/SV-ADSP-Pipeline> <https://dss.niagads.org/>

493 **Code availability**

494 ADSP-Short-Var and ADSP-All-Var panel building codes are publicly accessible at

495 <https://github.com/plCas/SNP-SV-imputation-panel-building-pipeline>

496 **Competing interests**

497 None

498 **Reference**

- 499 1 International HapMap, C. *et al.* Integrating common and rare genetic variation in diverse human
500 populations. *Nature* **467**, 52-58 (2010). <https://doi.org/10.1038/nature09298>
- 501 2 Genomes Project, C. *et al.* A map of human genome variation from population-scale
502 sequencing. *Nature* **467**, 1061-1073 (2010). <https://doi.org/10.1038/nature09534>
- 503 3 Huang, J. *et al.* Improved imputation of low-frequency and rare variants using the UK10K
504 haplotype reference panel. *Nat Commun* **6**, 8111 (2015). <https://doi.org/10.1038/ncomms9111>
- 505 4 McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet*
506 **48**, 1279-1283 (2016). <https://doi.org/10.1038/ng.3643>
- 507 5 Bick, A. G. *et al.* Inherited causes of clonal haematopoiesis in 97,691 whole genomes. *Nature*
508 **586**, 763-768 (2020). <https://doi.org/10.1038/s41586-020-2819-2>
- 509 6 Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program.
510 *Nature* **590**, 290-299 (2021). <https://doi.org/10.1038/s41586-021-03205-y>
- 511 7 Choi, J. *et al.* A whole-genome reference panel of 14,393 individuals for East Asian populations
512 accelerates discovery of rare functional variants. *Sci Adv* **9**, eadg6319 (2023).
513 <https://doi.org/10.1126/sciadv.adg6319>
- 514 8 O'Connell, J. *et al.* A population-specific reference panel for improved genotype imputation in
515 African Americans. *Commun Biol* **4**, 1269 (2021). <https://doi.org/10.1038/s42003-021-02777-9>
- 516 9 Girirajan, S. *et al.* Relative burden of large CNVs on a range of neurodevelopmental phenotypes.
517 *PLoS Genet* **7**, e1002334 (2011). <https://doi.org/10.1371/journal.pgen.1002334>
- 518 10 de Cid, R. *et al.* Deletion of the late cornified envelope LCE3B and LCE3C genes as a
519 susceptibility factor for psoriasis. *Nat Genet* **41**, 211-215 (2009).
520 <https://doi.org/10.1038/ng.313>
- 521 11 Allen, M. *et al.* Association of MAPT haplotypes with Alzheimer's disease risk and MAPT brain
522 gene expression levels. *Alzheimers Res Ther* **6**, 39 (2014). <https://doi.org/10.1186/alzrt268>
- 523 12 Brouwers, N. *et al.* Alzheimer risk associated with a copy number variation in the complement
524 receptor 1 increasing C3b/C4b binding sites. *Mol Psychiatry* **17**, 223-233 (2012).
525 <https://doi.org/10.1038/mp.2011.24>
- 526 13 Li, Y. *et al.* Integrated copy number and gene expression analysis detects a CREB1 association
527 with Alzheimer's disease. *Transl Psychiatry* **2**, e192 (2012). <https://doi.org/10.1038/tp.2012.119>
- 528 14 Noyvert, B. *et al.* Imputation of structural variants using a multi-ancestry long-read sequencing
529 panel enables identification of disease associations. *medRxiv*, 2023.2012.2020.23300308
530 (2023). <https://doi.org/10.1101/2023.12.20.23300308>
- 531 15 Langlois, A. W. R. *et al.* Genotyping, characterization, and imputation of known and novel
532 CYP2A6 structural variants using SNP array data. *J Hum Genet* **68**, 533-541 (2023).
533 <https://doi.org/10.1038/s10038-023-01148-y>
- 534 16 McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016).
535 <https://doi.org/10.1186/s13059-016-0974-4>
- 536 17 Wang, H. *et al.* Structural Variation Detection and Association Analysis of Whole-Genome-
537 Sequence Data from 16,905 Alzheimer's Diseases Sequencing Project Subjects. *medRxiv* (2023).
538 <https://doi.org/10.1101/2023.09.13.23295505>

- 539 18 Schwartzenuber, J. *et al.* Genome-wide meta-analysis, fine-mapping and integrative
540 prioritization implicate new Alzheimer's disease risk genes. *Nat Genet* **53**, 392-402 (2021).
541 <https://doi.org/10.1038/s41588-020-00776-w>
- 542 19 Paul, V. *et al.* Scratch2 modulates neurogenesis and cell migration through antagonism of bHLH
543 proteins in the developing neocortex. *Cereb Cortex* **24**, 754-772 (2014).
544 <https://doi.org/10.1093/cercor/bhs356>
- 545 20 Itoh, Y. *et al.* Scratch regulates neuronal migration onset via an epithelial-mesenchymal
546 transition-like mechanism. *Nat Neurosci* **16**, 416-425 (2013). <https://doi.org/10.1038/nn.3336>
- 547 21 Rodriguez-Aznar, E. & Nieto, M. A. Repression of Puma by scratch2 is required for neuronal
548 survival during embryonic development. *Cell Death Differ* **18**, 1196-1207 (2011).
549 <https://doi.org/10.1038/cdd.2010.190>
- 550 22 Schenning, K. J. *et al.* Gene-Specific DNA Methylation Linked to Postoperative Cognitive
551 Dysfunction in Apolipoprotein E3 and E4 Mice. *J Alzheimers Dis* **83**, 1251-1268 (2021).
552 <https://doi.org/10.3233/JAD-210499>
- 553 23 Lee, W. P. *et al.* Association of Common and Rare Variants with Alzheimer's Disease in over
554 13,000 Diverse Individuals with Whole-Genome Sequencing from the Alzheimer's Disease
555 Sequencing Project. *medRxiv* (2023). <https://doi.org/10.1101/2023.09.01.23294953>
- 556 24 Yu, K. *et al.* Meta-imputation: An efficient method to combine genotype data after imputation
557 with multiple reference panels. *Am J Hum Genet* **109**, 1007-1015 (2022).
558 <https://doi.org/10.1016/j.ajhg.2022.04.002>
- 559 25 Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer
560 sequencing applications. *Bioinformatics* **32**, 1220-1222 (2016).
561 <https://doi.org/10.1093/bioinformatics/btv710>
- 562 26 Robbins, M., Clayton, E. & Kaminski Schierle, G. S. Synaptic tau: A pathological or physiological
563 phenomenon? *Acta Neuropathol Commun* **9**, 149 (2021). [https://doi.org/10.1186/s40478-021-](https://doi.org/10.1186/s40478-021-01246-y)
564 [01246-y](https://doi.org/10.1186/s40478-021-01246-y)
- 565 27 Seshadri, S. *et al.* Genome-wide analysis of genetic loci associated with Alzheimer disease.
566 *JAMA* **303**, 1832-1840 (2010). <https://doi.org/10.1001/jama.2010.574>
- 567 28 Wu, Q. J. *et al.* EXOC3L2 rs597668 variant contributes to Alzheimer's disease susceptibility in
568 Asian population. *Oncotarget* **8**, 20086-20091 (2017).
569 <https://doi.org/10.18632/oncotarget.15380>
- 570 29 Pantic, B. *et al.* Myotonic dystrophy protein kinase (DMPK) prevents ROS-induced cell death by
571 assembling a hexokinase II-Src complex on the mitochondrial surface. *Cell Death Dis* **4**, e858
572 (2013). <https://doi.org/10.1038/cddis.2013.385>
- 573 30 Kaliman, P. & Llagostera, E. Myotonic dystrophy protein kinase (DMPK) and its role in the
574 pathogenesis of myotonic dystrophy 1. *Cell Signal* **20**, 1935-1941 (2008).
575 <https://doi.org/10.1016/j.cellsig.2008.05.005>
- 576 31 Langbehn, K. E. *et al.* DMPK mRNA Expression in Human Brain Tissue Throughout the Lifespan.
577 *Neurol Genet* **7**, e537 (2021). <https://doi.org/10.1212/NXG.0000000000000537>
- 578 32 Rajasekharan, S. & Kennedy, T. E. The netrin protein family. *Genome Biol* **10**, 239 (2009).
579 <https://doi.org/10.1186/gb-2009-10-9-239>
- 580 33 Wang, H., Copeland, N. G., Gilbert, D. J., Jenkins, N. A. & Tessier-Lavigne, M. Netrin-3, a mouse
581 homolog of human NTN2L, is highly expressed in sensory ganglia and shows differential binding

- 582 to netrin receptors. *J Neurosci* **19**, 4938-4947 (1999). [https://doi.org/10.1523/JNEUROSCI.19-](https://doi.org/10.1523/JNEUROSCI.19-12-04938.1999)
583 [12-04938.1999](https://doi.org/10.1523/JNEUROSCI.19-12-04938.1999)
- 584 34 Meng, Y., Sun, S., Cao, S. & Shi, B. Netrin-1: A Serum Marker Predicting Cognitive Impairment
585 after Spinal Cord Injury. *Dis Markers* **2022**, 1033197 (2022).
586 <https://doi.org/10.1155/2022/1033197>
- 587 35 Ju, T. *et al.* Decreased Netrin-1 in Mild Cognitive Impairment and Alzheimer's Disease Patients.
588 *Front Aging Neurosci* **13**, 762649 (2021). <https://doi.org/10.3389/fnagi.2021.762649>
- 589 36 Bai, B. *et al.* Deep Multilayer Brain Proteomics Identifies Molecular Networks in Alzheimer's
590 Disease Progression. *Neuron* **105**, 975-991 e977 (2020).
591 <https://doi.org/10.1016/j.neuron.2019.12.015>
- 592 37 Weinfeld, M., Mani, R. S., Abdou, I., Aceytuno, R. D. & Glover, J. N. Tidying up loose ends: the
593 role of polynucleotide kinase/phosphatase in DNA strand break repair. *Trends Biochem Sci* **36**,
594 262-271 (2011). <https://doi.org/10.1016/j.tibs.2011.01.006>
- 595 38 Dumitrache, L. C. & McKinnon, P. J. Polynucleotide kinase-phosphatase (PNKP) mutations and
596 neurologic disease. *Mech Ageing Dev* **161**, 121-129 (2017).
597 <https://doi.org/10.1016/j.mad.2016.04.009>
- 598 39 McCorkindale, A. N., Patrick, E., Duce, J. A., Guennewig, B. & Sutherland, G. T. The Key Factors
599 Predicting Dementia in Individuals With Alzheimer's Disease-Type Pathology. *Front Aging*
600 *Neurosci* **14**, 831967 (2022). <https://doi.org/10.3389/fnagi.2022.831967>
- 601 40 Hehir-Kwa, J. Y. *et al.* A high-quality human reference panel reveals the complexity and
602 distribution of genomic structural variants. *Nat Commun* **7**, 12989 (2016).
603 <https://doi.org/10.1038/ncomms12989>
- 604 41 Chen, L. *et al.* Association of structural variation with cardiometabolic traits in Finns. *Am J Hum*
605 *Genet* **108**, 583-596 (2021). <https://doi.org/10.1016/j.ajhg.2021.03.008>
- 606 42 Beecham, G. W. *et al.* The Alzheimer's Disease Sequencing Project: Study design and sample
607 selection. *Neurol Genet* **3**, e194 (2017). <https://doi.org/10.1212/NXG.0000000000000194>
- 608 43 Kwong, A. M. *et al.* Robust, flexible, and scalable tests for Hardy-Weinberg equilibrium across
609 diverse ancestries. *Genetics* **218** (2021). <https://doi.org/10.1093/genetics/iyab044>
- 610 44 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform.
611 *Bioinformatics* **25**, 1754-1760 (2009). <https://doi.org/10.1093/bioinformatics/btp324>
- 612 45 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-
613 generation DNA sequencing data. *Genome Res* **20**, 1297-1303 (2010).
614 <https://doi.org/10.1101/gr.107524.110>
- 615 46 Naj, A. C. *et al.* Quality control and integration of genotypes from two calling pipelines for
616 whole genome sequence data in the Alzheimer's disease sequencing project. *Genomics* **111**,
617 808-818 (2019). <https://doi.org/10.1016/j.ygeno.2018.05.004>
- 618 47 GitHub-DecodeGenetics/svimmer. Structural Variant Merging Tool. (2021).
- 619 48 Eggertsson, H. P. *et al.* GraphTyper2 enables population-scale genotyping of structural variation
620 using pangenome graphs. *Nat Commun* **10**, 5402 (2019). [https://doi.org/10.1038/s41467-019-](https://doi.org/10.1038/s41467-019-13341-9)
621 [13341-9](https://doi.org/10.1038/s41467-019-13341-9)
- 622 49 Delaneau, O., Zagury, J. F., Robinson, M. R., Marchini, J. L. & Dermitzakis, E. T. Accurate,
623 scalable and integrative haplotype estimation. *Nat Commun* **10**, 5436 (2019).
624 <https://doi.org/10.1038/s41467-019-13225-y>

- 625 50 Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat Genet* **48**, 1284-
626 1287 (2016). <https://doi.org/10.1038/ng.3656>
- 627 51 Zhu, W. *et al.* IMMerge: merging imputation data at scale. *Bioinformatics* **39** (2023).
628 <https://doi.org/10.1093/bioinformatics/btac750>
- 629 52 Gogarten, S. M. *et al.* Genetic association testing using the GENESIS R/Bioconductor package.
630 *Bioinformatics* **35**, 5346-5348 (2019). <https://doi.org/10.1093/bioinformatics/btz567>
- 631 53 Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10** (2021).
632 <https://doi.org/10.1093/gigascience/giab008>
- 633 54 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features.
634 *Bioinformatics* **26**, 841-842 (2010). <https://doi.org/10.1093/bioinformatics/btq033>
635