1

Substantial role of rare inherited variation in individuals with developmental disorders

Kaitlin E. Samocha^{1,2,3,4*}, V. Kartik Chundru^{4,5}, Jack M. Fu^{1,2,6}, Eugene J. Gardner⁴, Petr Danecek⁴, Emilie M. Wigdor^{4,7}, Daniel S. Malawsky⁴, Sarah J. Lindsay⁴, Patrick Campbell^{4,8}, Tarjinder Singh⁴, Ruth Y. Eberhardt⁴, Giuseppe Gallone⁴, Caroline F. Wright⁵, Hilary C. Martin⁴, Helen V. Firth^{4,8}, Matthew E. Hurles^{4,*}

- 1. Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA
- Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA
- 3. Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA
- 4. Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK
- 5. Department of Clinical and Biomedical Sciences, University of Exeter Medical School, Royal Devon and Exeter Hospital, Exeter, UK
- Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA
- 7. Institute of Developmental and Regenerative Medicine, Department of Paediatrics, University of Oxford, Oxford, UK
- 8. Cambridge University Hospitals Foundation Trust, Addenbrooke's Hospital, Cambridge, UK
- * Corresponding authors: K.E.S. (samocha@broadinstitute.org), M.E.H. (meh@sanger.ac.uk)

1 Abstract

- 2 While the role of *de novo* and recessively-inherited coding variation in risk for rare
- 3 developmental disorders (DDs) has been well established, the contribution of damaging
- 4 variation dominantly-inherited from parents is less explored. Here, we investigated the
- 5 contribution of rare coding variants to DDs by analyzing 13,452 individuals with DDs, 18,613 of
- 6 their family members, and 3,943 controls using a combination of family-based and case/control
- 7 analyses. In line with previous studies of other neuropsychiatric traits, we found a significant
- 8 burden of rare (allele frequency $< 1 \times 10^{-5}$) predicted loss-of-function (pLoF) and damaging
- 9 missense variants, the vast majority of which are inherited from apparently unaffected parents.
- 10 These predominantly inherited burdens are strongest in DD-associated genes or those
- 11 intolerant of pLoF variation in the general population, however we estimate that ~10% of the
- 12 excess of these variants in DD cases is found within the DD-associated genes, implying many
- 13 more risk loci are yet to be identified. We found similar, but attenuated, burdens when

2

14 comparing the unaffected parents of individuals with DDs to controls, indicating that parents 15 have elevated risk of DDs due to these rare variants, which are overtransmitted to their affected 16 children. We estimate that 6-8.5% of the population attributable risk for DDs are due to rare 17 pLoF variants in those genes intolerant of pLoF variation in the general population. Finally, we 18 apply a Bayesian framework to combine evidence from these analyses of rare, mostly-inherited 19 variants with prior *de novo* mutation burden analyses to highlight an additional 25 candidate DD-20 associated genes for further follow up.

22 Introduction

23 Developmental disorders (DDs) have a strong genetic component with many hundreds of genes associated with these disorders via both dominant and recessive mechanisms. 24 However, we are able to find diagnostic genetic variants for only $\sim 40\%$ of cases^{1,2}, leaving over 25 26 half of patients with potential genetic contributions undiagnosed. The majority of known 27 diagnoses in cohorts such as the Deciphering Developmental Disorders (DDD) are from de 28 novo variants (~76% of diagnosed cases in DDD²), including *de novo* single nucleotide variants 29 (SNVs), small insertions and deletions (indels), and structural variants (SVs) such as copy 30 number variants. Recessive diagnoses in DDD are a distant second (~12% of diagnosed cases) and most commonly impact families with consanguinity². Large SVs, particularly at established 31 32 genomic disorder loci, are well-known DD risk factors, but are often observed in unaffected family members and inherited from parents^{3,4}. While the field has long appreciated incomplete 33 penetrance of SVs⁵, there is still much to be learned about how incompletely penetrant SNVs 34 35 and indels that could be inherited from unaffected parents may contribute to overall risk of DDs⁶. 36 We and others have previously shown that inherited common variants influence 37 phenotypic variability^{7–9}, but have not undertaken a systematic analysis of how rare, inherited variants could contribute to DDs outside of ANKRD11¹⁰. Prior studies of DD individuals have 38 39 made the argument that variants inherited from unaffected parents can be diagnostic⁶. 40 Additionally, recent studies of autism, schizophrenia, and bipolar disorder have indicated a role 41 for rare, damaging variants in more common and complex neuropsychiatric disorders^{11–17}. 42 These rare variant burdens are most often concentrated in genes intolerant to loss-of-function 43 variants in the general population ("constrained genes") or those genes previously associated 44 with severe DDs. Damaging variation in these two gene sets has also been tied to reduced educational attainment and cognitive performance in population cohorts such as the UK 45 Biobank^{18–23}, further strengthening evidence of their role in DDs. 46

3

47 Here, we sought to define the contribution of rare, and mostly inherited, variation to DDs 48 in the DDD cohort. We jointly processed the entirety of the DDD cohort (n = 32,065) with 49 relatively healthy UK population samples from the INTERVAL study (n = 3.943), which allowed us to evaluate both within-family and case/control rare variant burdens (Fig 1a). We found a 50 significant burden of rare, damaging variants, particularly concentrated in constrained and 51 52 known DD-associated genes, in both our case/control analysis and family-based analysis. In line with these results, we also found that the parents in the DDD cohort have a significantly 53 higher burden of rare variants compared to controls. Finally, we combine gene-association 54 evidence from our prior *de novo* study²⁴ with the results of these analyses using TADA^{13,25}, and 55 nominate an additional 25 candidate DD genes for further investigation. 56 57

58 Results

Significant burden of rare variants in known DD-associated genes and constrained genes in DD cases

61 We joint-called autosomal SNVs and indels in the DDD cohort with controls from the 62 INTERVAL study. After sample and variant quality control (see Methods), we retained 32,065 63 individuals from the DDD study (both parents and DD cases) of all inferred genetic ancestry 64 groups. Of these individuals, 25,701 were of inferred European genetic ancestry for comparison against 3,943 ancestry-matched individuals from the INTERVAL study (Supplementary Figure 65 66 **1**). Post-quality control, we observed similar rates per exome of rare (allele frequency < 0.001) 67 autosomal variants in DD cases, their parents, and controls (Supplementary Table 1; 68 Supplementary Figure 2). We first compared the rare ($qnomAD^{26}$ allele frequency < $1x10^{-5}$; dataset allele 69 frequency $< 1 \times 10^{-4}$) variant burden between DD cases and control individuals (n = 10,644 cases 70 71 versus 3,943 controls), correcting for sex, the first 20 principal components (PCs), and the 72 number of rare variants per exome, in line with recent work¹². Correcting for the number of rare

- variants per exome leads to a significant odds ratio (OR) < 1 for synonymous variants when
- testing across all genes (OR = 0.974, p = 4.4×10^{-6} , **Fig 1b**); we thus conclude that our ORs for
- 75 damaging variants may be slightly under-estimated. Removing this correction, or correcting only
- for rare nonsynonymous variants, results in an OR significantly > 1 for synonymous variants that
- 77 was not eliminated by various corrections (Supplemental Note).

4

78 We found modest ORs for the burden of rare predicted loss-of-function (pLoF) or damaging missense variants (defined as $CADD^{27} \ge 20$) when evaluating all genes at once 79 80 (LOFTEE²⁶ high-confidence pLoF OR = 1.10 and p = 5.3×10^{-19} , damaging missense OR = 1.02 and $p = 5.3 \times 10^{-3}$; Supplementary Table 2). However, these ORs are much larger in previously 81 82 established monoallelic DD-associated genes^{24,28} (n=666 genes; pLoF OR = 3.24, p = 2.7×10^{-63}) and constrained genes²⁶ (pLI \ge 0.9, n=2,699 genes; pLoF OR = 1.88, p = 1.0x10⁻⁶¹). We used all 83 individuals in these analyses, but found nearly identical results when repeating the analyses 84 using only unrelated individuals (Supplementary Figure 3). These results reinforce the role for 85 86 these two gene sets in DDs, as has been demonstrated before^{23,29}.

We then split DD cases into non-trio cases (n=2,790) and trio cases (n=7,854), which 87 88 were further split based on whether (1) one or both parents were known to have a similar 89 phenotypic presentation and (2) whether the proband had a likely diagnostic de novo variant. All groups have a significant burden of rare variants compared to controls (Fig 1d; Supplementary 90 91 Table 2), with significantly higher ORs for pLoF variants in constrained and DD-associated 92 genes compared to the set with unaffected parents and no de novo diagnosis (n=4,183; pLoF OR = 1.49 and 1.73, respectively) for the non-trio cases (pLoF OR = 2.03 and 3.27; Wald test p 93 94 = 1.4×10^{-6} and 7.2×10^{-8} . respectively), trios with a diagnostic *de novo* variant (n=2.679; pLoF OR = 2.30 and 6.37; Wald test $p = 2.8 \times 10^{-11}$ and $< 1 \times 10^{-20}$, respectively), and trios with one or more 95 96 affected parents (n=1,246; pLoF OR = 2.42 and 3.25; Wald test p = 1.2×10^{-8} and 8.2×10^{-7} . 97 respectively; **Supplementary Table 3**). These significantly higher results were anticipated for the trios with one or more affected parents, where we expected that the affected parent would 98 99 be transmitting risk variants to their children. For the trios with a likely diagnostic *de novo* 100 variant, we might not have expected to see a large exome-wide burden given the presence of a 101 high-impact variant (i.e., the diagnostic *de novo* variant), however this burden could be driven by 102 the de novo variants.

103 Given that *de novo* variants play a large role in DDs and that they were included here, 104 we repeated these regressions after removing the known *de novo* variants reported previously²⁴ and observed an attenuated signal in all trio groups. This attenuation was most notable for the 105 106 individuals that have a diagnostic de novo variant, which no longer have a significant burden of 107 rare damaging variants compared to controls in all genes (pLoF OR = 1.00, p = 0.865; 108 Supplementary Figure 4; Supplementary Table 4). However, this set of de novo diagnosed 109 individuals does retain some signal in constrained (pLoF OR = 1.19, p = 3.9×10^{-4}) and DD-110 associated genes (pLoF OR = 1.58, p = 8.2×10^{-8}), even after removing the *de novo* variants.

5

111 While case/control comparisons are very sensitive to differences in genetic ancestry, 112 within-family analyses are protected against such stratification, allowing us to test all trios. We 113 sought to guantify the burden of rare variation that had been transmitted from parents using the 114 Transmission Disequilibrium Test (TDT). We applied TDT to 9,305 DDD trios (all inferred 115 genetic ancestry groups, but selecting only one sibling per family) and found a significant signal of overtransmission of rare (gnomAD allele frequency $< 1 \times 10^{-5}$; transmitted doubletons 116 117 compared to nontransmitted singletons) damaging variants from parents to their children with 118 DDs (Fig 1c; Supplementary Table 5). We found similar patterns as with the case/control 119 analysis, with the strongest rate ratios (RRs) for pLoF variants in DD-associated and/or constrained genes (RR DD = 1.48 and p = 5.8×10^{-10} ; RR constrained = 1.23 and p = 1.4×10^{-9}). 120 121 Notably, we do not see any overtransmission of rare, synonymous variants in any of the gene 122 sets tested (RR for all genes = 1.01, p = 0.183). 123 As in the case/control analysis, we split the trios by parental affected status and whether 124 they had a diagnostic *de novo* variant. Even though the trios with one or more affected parents 125 (n=1,306) had higher RRs than trios with unaffected parents and no *de novo* diagnosis 126 (n=5,124) for pLoF variants in constrained (affected parents: RR = 1.51, $p = 5.5 \times 10^{-7}$; unaffected parents, no *de novo* diagnosis: RR = 1.28, p = 1.5x10⁻⁷; Supplementary Table 5) and/or DD-127 associated genes (affected parents: RR = 1.87, $p = 1.7 \times 10^{-5}$; unaffected parents, no *de novo* 128 129 diagnosis: RR = 1.68, p = 2.1×10^{-9}), the differences were not significantly different from each 130 other (Supplementary Table 6). When evaluating the trios with a diagnostic *de novo* variant 131 (n=3,155), we found no significant overtransmission of damaging variants in any of the gene 132 sets (e.g., pLoF RR in constrained genes = 1.02, p = 0.800). These attenuated RRs for pLoF 133 variants were nominally significantly lower than what was seen for unaffected parent and no de 134 *novo* diagnosis trios, but not all comparisons survive multiple testing correction (p = 0.05/16) 135 comparisons = 0.003; **Supplementary Table 6**). No differences were significant for missense 136 variants between trios with a diagnostic *de novo* variant and the unaffected parent, no 137 diagnostic de novo variant trios.

138

6



139

140 **Figure 1.** Overview of the dataset and analyses.

a) Schematic representation of the study, showing the joint-calling of 32,065 individuals from the 141 142 DDD study (18,613 parents and 13,432 DD cases) with 3,943 genetic-ancestry matched 143 controls from the INTERVAL study. DD cases can be further subdivided into non-trio cases 144 (n=3,596; 2,790 of inferred European genetic ancestry), trio cases with at least one affected parent (n=1,462; 1,246 of inferred European genetic ancestry), trio cases with a *de novo* 145 146 diagnosis (n=3,215; 2,679 of inferred European genetic ancestry), and trio cases with 147 unaffected parents and no de novo diagnosis (n=5,470; 4,183 of inferred European genetic 148 ancestry). Given the joint-calling, we performed both case versus control burden testing and 149 within-family transmission disequilibrium tests (TDT). Case/control burden (b) and transmission disequilibrium (c) testing for all DD cases (n=10,644 for case/control, 9,305 for TDT). d) 150 151 Case/control regressions split by the type of DD case being compared to controls: trios with a 152 diagnostic de novo variant are shown in circles, trios with one or more affected parents are 153 shown with triangle, non-trio DD cases are shown in squares, and trios with unaffected parents 154 and no de novo diagnosis are shown in diamonds. In b) and d), we report the odds ratios from 155 logistic regressions. For the TDT tests in c), we are displaying the rate ratio between transmitted 156 to nontransmitted variants. Displayed are the results for three mutation classes-LOFTEE high 157 confidence predicted loss-of-function (HC pLoF, dark red), damaging missense with CADD \geq 20 158 (orange), and synonymous (yellow) variants-and four gene sets: all genes (n=18,610), 159 constrained (pLI \geq 0.9, n=2,699), monoallelic DD-associated (n=666), and unconstrained genes with no prior monoallelic DD-association (n=15,667). * $1x10^{-3} \le p < 1x10^{-2}$; ** $1x10^{-4} \le p < 1x10^{-1}$ 160 ³; *** p < 1x10⁻⁴. 161

7

1	62	

163 164 As both the TDT and case/control analyses establish, inherited rare variants significantly 165 contribute to DDs as do common variants (see Huang*, Wigdor* et al.⁹). Given that we saw 166 significant overtransmission of damaging variants from unaffected parents in the TDT results. 167 we wanted to test if this was due to these parents having an excess of rare damaging variants 168 compared to controls versus other potential explanations, such as ascertaining families where 169 there was significant overtransmission of such variants. We therefore compared unaffected 170 parents to controls with a similar logistic regression as above and found significant differences 171 for missense and pLoF variation, albeit lower than what we found for the DD cases (13,861 172 unaffected parents versus 3,943 controls; Supplementary Figures 5 & 6). No significant 173 difference was seen for synonymous variants. As expected, affected parents have higher 174 burdens than unaffected parents, but the enrichments for pLoF variants in DD-associated and 175 constrained genes are significant for both sets of parents (Fig 2a). While we found a few 176 instances where mothers and fathers seemed to have a difference in their genetic burden - for 177 example, the OR of pLoF variants in DD-associated genes in unaffected mothers was higher 178 than in unaffected fathers (mother OR = 1.44 and p = 5.9×10^{-4} , father OR = 1.05 and p = 0.643; 179 Fig 2b) – these differences were not statistically significant (Supplementary Figure 7).

180

8



181

182 **Figure 2.** Parents of DD cases have significant rare variant burdens compared to controls.

183 a) Odds ratio when comparing rare variant burdens to 3,943 controls for individuals of European 184 genetic ancestry: developmental disorder cases in complete trios ("Affected child", circles, 185 n=7,854), affected parents (triangles, n=1,195), and unaffected parents (squares, n=13,861). b) The same plot as a), but with the unaffected parents split by sex of the parent. In all plots, high-186 187 confidence predicted loss-of-function (HC pLoF) variants are in dark red, damaging missense 188 variants with CADD \geq 20 are in orange, and synonymous variants are in yellow. There are four 189 gene sets shown: all genes (n=18,610), constrained (pLI ≥0.9, n=2,699), monoallelic DD-190 associated (n=666), and unconstrained genes with no prior monoallelic DD-association (n=15,667). * $1x10^{-3} \le p < 1x10^{-2}$; ** $1x10^{-4} \le p < 1x10^{-3}$; *** $p < 1x10^{-4}$. 191

192

193 Investigating differences in rare variant burden

194 We wanted to further investigate differential burden in the DD cases, accounting for

195 factors known to influence damaging genetic burden. Of note, neurodevelopmental disorders,

196 particularly autism, have been reported to have a noticeable female protective effect^{30,31}. The

- 197 female protective effect would be expected to lead to patterns where female cases and mothers
- 198 have a higher burden of rare, damaging variants compared to male cases and fathers,

9

199 respectively. Indeed, we previously reported that female cases had a higher rate of damaging 200 *de novo* variants compared to male cases^{24,32}. However, we found no significant differences in 201 the rare autosomal variant burden when comparing: (1) male to female cases (Fig 3a, 202 Supplementary Figure 8), (2), unaffected mothers to unaffected fathers, as mentioned above 203 (Supplementary Figure 7), or (3) the burden inherited from mothers to that inherited from 204 fathers (Supplementary Table 7). 205 Under a liability threshold model, numerous genetic and environmental factors can 206 contribute to the risk of developing a given disorder, and individuals whose accumulated risk 207 factors (liability) cross a threshold are affected by the disorder³³. The female protective effect 208 discussed above can be thought of as one example of a liability threshold model, where, for 209 example, females require a greater liability (e.g., burden of risk factors) than males to be 210 diagnosed. Here, we wanted to explore other factors that could contribute overall to liability and 211 would therefore influence the amount of expected burden from rare variation. For example, 212 individuals who were born prematurely (< 37 weeks gestation) or had other environmental exposures in utero^{34,35} may be expected to have a greater liability coming from environmental 213 214 factors and could have a lower rare variant burden than those without these exposures. In line 215 with this, we have previously reported that this set of individuals has a lower diagnostic rate than 216 those DD cases without prematurity of environmental exposures². Similarly, DD cases who have 217 affected family members have both been reported to have a significantly lower chance of having 218 a *de novo* variant and of receiving a genetic diagnosis^{2,36}. For those DD cases with affected 219 family members, we would anticipate a greater inherited genetic burden of rare and/or common variants, the latter of which is explored in other work⁹, particularly given the lack of a large-220 221 impact diagnostic variant.

222 For environmental exposures, we selected a set of 2,637 probands within DDD who 223 were born prematurely (< 37 weeks gestation), had mothers with diabetes, and/or were exposed 224 to antiepileptic medications in utero, all three of which were shown to have lower diagnostic 225 rates in DDD². Rare variant burdens for DD cases with unaffected parents and no diagnostic de 226 novo variant were similar whether they had or did not have an environmental exposure (Fig 3b. 227 **Supplementary Figure 9**), with no significantly different burdens when directly comparing 228 cases with (n=879) versus without (n=3,304) environmental exposures in a case versus case 229 regression (Supplementary Table 8).

By contrast, when we compared rare variant burdens in unrelated European ancestry cases with versus without affected family members using a case versus case regression as above, we found a significant trend of increasing burden in constrained genes that was

10

233 strongest for those cases with at least one affected parent (Fig 3c). These results are consistent 234 with recent work⁹ that reported that cases with more affected first-degree relatives had lower polygenic scores for educational attainment³⁷ as well as a lower chance of receiving a 235 236 monogenic diagnosis. However, we did not see as strong of a trend for DD-associated genes 237 (Supplementary Figure 10) despite this being the gene set that showed the largest difference from controls (Supplementary Figure 6), nor did we find any significant differences in rare 238 239 variant burdens when the other affected family members were siblings or other first-degree relatives, even though these cases have a lower diagnostic rates and lower polygenic scores for 240 241 educational attainment^{9,36}.

242



243



a) Odds ratios when comparing female cases to female controls (circles, n = 1,651 and 1,940,

respectively) and male cases to male controls (squares, n = 2,532 and 2,003, respectively) for

247 individuals of European genetic ancestry with unaffected parents and no de novo diagnosis. 248 While female cases have higher odds ratio estimates for high-confidence predicted loss-of-249 function (HC pLoF) variants in DD-associated genes, there are no significant differences for any 250 group, b) The rate ratio from the transmission disequilibrium test (TDT) for cases of all genetic 251 ancestries with unaffected parents and no de novo diagnosis (n = 5.124) compared to that set of 252 cases who also had an environmental exposure (i.e., premature birth, maternal diabetes, and/or 253 exposure to antiepileptic medications in utero; n = 1,075). There were no significant differences 254 between the two sets. c) Odds ratios in constrained genes from a regression comparing 255 unrelated European genetic ancestry cases with and without additional affected family 256 members, split by the number and type of relative affected. All comparisons are against the 257 burden in cases that are the only affected family member (n = 6,664). For all panels, LOFTEE 258 high-confidence pLoFs are shown in dark red; missense variants with CADD \geq 20 in orange; 259 and synonymous variants in yellow. For a) and b), there are four gene sets shown: all genes (n=18,610), constrained (pLI \geq 0.9, n=2,699), monoallelic DD-associated (n=666), and 260 261 unconstrained genes with no prior monoallelic DD-association (n=15.667). Panel c) only shows 262 constrained genes. * $1x10^{-3} \le p < 1x10^{-2}$; ** $1x10^{-4} \le p < 1x10^{-3}$; *** $p < 1x10^{-4}$. 263

264 Burden of rare variation primarily resides outside of known DD-associated genes

Given the heterogeneity of DD cases within our cohort, we expected to be underpowered to detect significant burdens on an individual gene basis. When we performed a Fisher's exact test to compare the burden of rare damaging variants in DD cases to those in controls, no individual gene crossed an exome-wide significance threshold ($p < 2.8x10^{-6}$). The most significant gene was *ANKRD11* (pLoF carriers: 50 DD versus 1 control; Fisher's OR = 18.6 and $p = 3.7x10^{-6}$), a well-established DD-associated gene, with at least 20 *de novo* variants included in the DD carrier counts.

272 We therefore sought to estimate the excess burden in terms of the number of additional 273 variants in DD cases in the significant gene lists from above. To do so, we estimated the excess 274 of pLoF and damaging missense variants using the rate of observed variants in controls and 275 corrected for the synonymous burden in each gene group. In the 4,183 DD cases with 276 unaffected parents and no de novo diagnosis, we estimated an excess of approximately 4,000 277 pLoF and damaging missense variants, with two thirds coming from missense variants 278 (Supplementary Figure 11). Only ~11% of the excess was found in the previously DD-279 associated genes, indicating that there are more disease-associated genes to be identified. 280 These excesses are not driven by specific individuals with far more variants than others, but by 281 a minor shift in the distribution of the number of rare variants per individual (**Supplementary** 282 Figure 12). Further, we estimated that the population attributable risk (PAR) from rare pLoF 283 variants in constrained genes was 8.4%. This estimate of PAR was only minimally impacted by

12

a range of population prevalence values for DDs, but drops to 5.9% when using the case/control
 analysis that removed known *de novo* variants (see **Supplemental Note** for caveats). This PAR
 estimate is similar to the 6% PAR reported for inherited pLoF variants in a recent autism study¹⁵.

287 To further investigate the nature of the rare variant exome-wide burden in the DD cases, 288 we calculated s_{het} burden scores¹⁹, a measure of the cumulative burden of rare variants made by combining the s_{het} selection coefficients³⁸ for each autosomal gene impacted by damaging 289 variants (pLoF with CADD \ge 25; missense with MPC³⁹ \ge 2 and CADD \ge 25), for all individuals in 290 291 the study (see Methods). Specifically, we wanted to evaluate if the exome-wide burden of rare 292 inherited damaging variants in DD cases differed from the burden seen in parents. For these 293 analyses, we removed known *de novo* variants from the DD cases when calculating shet burden 294 scores. We found a significant difference in s_{het} burden scores calculated from rare pLoF variants (CADD \geq 25 as in Gardner et al.¹⁹) when comparing DD cases to the parental data 295 296 (Wilcox p = 2.01 x 10^{-9}), but not for s_{het} burden scores calculated using synonymous variants 297 (Wilcox p = 0.121; **Table 1**). This suggests an overtransmission of pLoF s_{het} burden scores from 298 parents to their children with DDs. Notably, we did not find a significant difference for s_{het} burden 299 scores calculated using damaging missense variants as defined by Gardner et al.¹⁹, likely as the 300 class is too rare for there to be large differences (CADD \ge 25 and MPC \ge 2; **Supplementary** 301 **Table 9**). We also found no significant differences between s_{het} burden scores for male versus 302 female DD cases, in line with analyses reported above.

		Synonymou	pLoF		
Group	Effect size N (95% Cl) P-value			Effect size (95% Cl) P-value	
All	9305	0.0038 (-0.0042 – 0.0106)	0.121	0.0045 (0.0027 – 0.0064)	1.88e-09
European	7037	-0.0029 (-0.0108 – 0.0064)	0.732	0.0047 (0.0028 – 0.0067)	6.01e-08
Unaffected parents, no DNM diagnosis	5124	0.0056 (-0.0048 – 0.0151)	0.344	0.0047 (0.0023 – 0.0074)	2.37e-05
DNM diagnosis	3155	0.0052 (-0.0093 – 0.0214)	0.111	0.0022 (-0.0004 – 0.0049)	0.002
Affected parents	1306	-0.0093 (-0.0262 – 0.0092)	0.992	0.0091 (0.0035 – 0.0152)	0.001

Table 1. Differences in the s_{het} burden scores for DD cases compared to their parents.

Shown are s_{het} burden scores made when using rare synonymous variants or rare pLoF variants with CADD \ge 25 for five sets of trios. In the case where a family had multiple affected children,

13

306 only one child (and thereby complete trio) was used. For these analyses, all known *de novo* 307 variants were removed from calculations. Bootstrapping with 1000 replicates was used to 308 determine the median difference in s_{het} burden scores as well as the 95% confidence intervals 309 (CI). P-values are from the Wilcoxon rank sum test.

310

311 We additionally compared the s_{het} burden scores between unrelated European genetic 312 ancestry DD cases (n=8,062) and controls in a logistic regression, which allowed us to estimate 313 the variance explained on the liability scale. Assuming a population prevalence of 1% for DDs, 314 we estimated that the shet burden scores explained 2.6% [0.2 - 7.5%] of the variance on the 315 liability scale, which is lower than the estimated 11.2% due to common variants genome-wide9. 316 Additionally, we found that the variance explained was higher if only using the *de novo* 317 diagnosed individuals (n=1,905 cases; 4.6% of liability [1.1 - 10.1%]) compared to only 318 undiagnosed DD cases (n=6,157; 1.9% [0.4 - 6.3%]). By contrast, the raw counts of rare pLoF 319 and damaging missense variants would only explain 0.06% [0.03 - 0.1%] of the variance on the 320 liability scale.

321 Combining evidence from *de novo* and inherited variants identifies additional candidate 322 DD genes

323 We previously identified nearly 300 genes significantly associated with DDs via a genespecific enrichment of *de novo* variants²⁴, with dozens more genes near the significance 324 325 threshold ("31k analysis"). We wanted to combine evidence from *de novo* variants with the TDT 326 and case/control analyses performed here to improve our ability to identify DD-associated genes. We applied the Transmission and De Novo Association (TADA) framework^{13,25}, a 327 328 Bayesian analytical framework that makes use of priors on the risk of a given variant class in 329 each gene to measure statistical evidence, which has been successfully applied to studies of 330 the genetic basis of autism^{13,25,40}.

At an exome-wide significance threshold of 2.8x10⁻⁶, we identified 269 significant genes 331 332 (Supplementary Table 10). As in the prior autism studies, pLoF variants had the largest 333 combined Bayes Factor (BF) contribution for the variant types tested (66.7%; Fig 4a), and we 334 found that nearly all of the BF contribution came from *de novo* variants (97.2%; Fig 4b) with 335 case/control a distant second (2.5%). Unsurprisingly given the strong *de novo* contribution, the 336 vast majority of these 269 genes were significant in our previous 31k analysis²⁴ (217; 81%). To 337 identify a set of genes with limited prior association to DDs, we focused on the 52 genes that 338 were not significant in the prior study of *de novo* variants. Of those, 27 already had strong or 339 definitive evidence of being monoallelic DD-associated genes according to DDG2P²⁸ as of July

- 340 2023 (Supplementary Figure 13). We selected the remaining 25 genes (2 with moderate, 8
- with limited, and 15 with no evidence of being monoallelic DD-associated genes in DDG2P) for
- 342 deeper investigation (**Fig 4c**). For most of these genes, the primary driver of association was
- 343 the signal from *de novo* pLoF variants, although we saw stronger case/control and inherited
- 344 signals as measured by BF contribution for these 25 genes compared to the 244 significant but
- 345 known genes (**Fig 4d-e**; **Supplementary Table 11**).
- 346



347

Figure 4. Results from a combined *de novo*, case/control, and inherited analysis of rare variants
 in developmental disorder (DD) cases using the Transmission and De Novo Association (TADA)
 framework.

The evidence of association with DD, as measured by Bayes Factors (BF) in TADA, for different

variant types (a) and inheritance types (b) across all 269 significant genes ($p < 2.8 \times 10^{-6}$). In c)

353 we depict the specific Bayes Factors from TADA for 25 genes that were significant in the TADA

16

analysis, but were not significant in our prior analysis of *de novo* variants alone²⁴ and had 354 355 limited or no evidence of DD association in DDG2P. In d) and e) the Bayes Factor percentages 356 are compared for the TADA significant genes with previous DD-associations ("Known 357 significant", n=244 genes) versus those with minimal prior evidence DD-association ("New 358 significant", n=25 genes) for different variant types (d) and inheritance types (e). For all plots, 359 high-confidence predicted loss-of-function (pLoF) variants are depicted in red, missense 360 variants with MPC \geq 2 (misB) in orange, and missense variants with 1 \leq MPC < 2 (misA) in 361 yellow. Similarly, signal from *de novo* variants are depicted in black, case/control in dark gray, 362 and inherited (e.g., transmitted versus nontransmitted) in light gray.

363

364 For these 25 genes with limited evidence of association in DDG2P, we further classified 365 them as having high, medium, or low likelihood of being true DD-associated genes based on manual curation of additional gene-phenotype databases (e.g., OMIM⁴¹, GenCC⁴², PanelApp⁴³), 366 constraint scores^{26,44,45}, literature searches, and evaluation of the Bayes Factors from TADA 367 368 (see Supplemental Note for more details about classification; Supplementary Figures 14-15). 369 Eleven of these 25 genes had strong evidence of being *bona fide* DD-associated genes, 370 primarily due to being considered by other clinical centers to have strong enough evidence to 371 use for diagnosis in closely related phenotypes (e.g., SYNCRIP) or recent publications (e.g., 372 *FOSL2⁴⁶*). Thirteen of the genes were considered to have medium evidence of DD association; 373 the majority of these genes had strong statistical evidence but no additional evidence in the 374 literature. We note that 77% of these genes are highly constrained against loss-of-function 375 variation in the general population (10 of 13 genes are LOEUF < 0.35), a feature that is greatly 376 upweighted in TADA. Finally, one gene, GLYR1, had weaker evidence for true DD-association 377 given that two of the three *de novo* pLoF variants, which were the primary driver of association, 378 were found in patients that could be considered diagnosed by other variants. These findings 379 diminish the strength of association, but as a small subset of DDD cases have dual molecular 380 diagnosis², we cannot fully rule out this gene as a potential DD-associated gene.

381 For the eleven genes with a high likelihood of being true DD genes (BAP1, FOSL2, 382 KCNT2, PSMC3, RFX7, SEMA6B, SRSF1, SYNCRIP, VCP, ZBTB7A, ZFHX3), we investigated 383 how many of individuals harboring damaging variants in these genes were considered 384 diagnosed with other variants (see **Supplemental Note** for details). When using a Bayes Factor 385 threshold of >1.3 to define variant carriers, we found that only 14.1% (12 of 85) of individuals 386 with one of these variants has some other diagnosis, which was significantly lower than the 387 overall diagnostic rate (27.5%; Fisher's exact $p = 4.8 \times 10^{-3}$). Of note, we recently estimated⁴⁷ that 388 a similar percentage (~12.5%) of DDD cases diagnosed with a *de novo* variant could have a

17

389 second *de novo* diagnosis, with most of that burden outside of the known genes. These results 390 emphasize that these eleven genes are more likely to be *bona fide* DD-associated genes. 391 Surprisingly, we found that a known biallelic DD-associated gene, SLC39A8, was 392 significant in the TADA analysis ($p = 1.14 \times 10^{-9}$) with contributions from all three inheritance 393 classes (i.e., de novo, case/control, and inherited). We reported a similar set of findings for 394 KDM5B previously⁴⁸, which also had signal from *de novo*, biallelic, and overtransmission of 395 pLoFs. SLC39A8 shows some depletion of both pLoF and missense variants in population 396 cohorts (gnomAD²⁶ v2: pLoF observed / expected = 0.18 and missense observed / expected = 397 0.8), but does not have particularly strong constraint scores (pLI = 0.67 and LOEUF = 0.47), 398 meaning that the priors in TADA would not be large. Focusing on the contribution from pLoF 399 variants to the association signal for SLC39A8, there were three individuals with a de novo pLoF 400 in SLC39A8, four non-trio individuals that contributed to the case/control analysis (with zero 401 found in controls), and eight instances where a parent transmitted a pLoF variant (as part of an 402 8:0 transmitted:nontransmitted signal). Across these fifteen individuals, only one carried another 403 putatively diagnostic variant – one of the individuals with a de novo pLoF variant in SLC39A8 404 also harbored a de novo missense variant in DDX6, a known DD-associated gene. For those 405 individuals in the DDD cohort (all twelve in the case/control and inherited analysis), we searched 406 for additional rare coding variants in SLC39A8 that could contribute to a recessive diagnosis. As 407 both missense and pLoF variants have been reported as pathogenic or likely pathogenic in 408 ClinVar⁴⁹ for SLC39A8, we included both in our search, but found no additional rare variants in 409 the twelve patients we queried. 410 Additionally, we noted a striking 24:3 transmitted:nontransmitted signal for missense

variants with moderate MPC scores³⁹ ($1 \le MPC < 2$) in SLC39A8. Three of these 24 individuals 411 412 had a partial or full diagnosis from other variants, but the rest were considered undiagnosed 413 (n=21; 87.5%). As above, we searched for other rare coding variants in SLC39A8 and found a 414 second missense variant inherited from the other parent for only one individual, whose 415 phenotypes were consistent with the recessive SLC39A8-Congenital Disorder of Glycosylation⁵⁰. The remaining individuals have no other convincing diagnostic candidate or 416 417 other rare coding variants in SLC39A8. Based on these analyses, we consider SLC39A8 a 418 promising candidate DD-associated gene that can operate via both monoallelic and biallelic 419 mechanisms, like the previously reported KDM5B.

420

18

421 Discussion

422 In this work, we demonstrated a significant contribution of rare, typically inherited, 423 damaging variants to the risk of severe developmental disorders (DD) by comparing both the 424 rare variant burden in DD cases to controls, as well as evaluating within-family overtransmission 425 of these variants from parents to their affected children. The majority of this rare variant burden 426 was found in genes intolerant of loss-of-function variants in reference population datasets (i.e., 427 constrained genes)²⁶ and genes previously associated with DDs via monogenic forms of 428 inheritance, as has also been seen for other neurodevelopmental and psychiatric disorders¹¹⁻ ^{13,17,29}. We estimate that rare pLoF variants in constrained genes, for example, account for 6-429 430 8.5% of the population attributable risk in our cohort. While the burden of overtransmission was 431 stronger from affected parents, cohort-wide nearly all of the burden was transmitted from 432 unaffected parents (e.g., 92.5% of transmitted rare pLoFs and missense variants across all 433 genes are from unaffected parents). In a set of DD cases with unaffected parents, we found an 434 excess of thousands of rare pLoF and damaging missense variants spread across many genes 435 and concentrated outside the previously DD-associated genes, indicating that there are 436 additional DD risk loci to be identified. Finally, we applied the Transmission and De Novo Association (TADA) framework¹³ to combine the evidence of association from *de novo* variants, 437 438 case/control analyses, and the overtransmission of variants within families, and identified 269 439 significant genes, of which 25 had limited to no prior evidence of DD-association.

440 When removing de novo variants, which are known to be large contributors to DD risk²⁴, 441 from our case/control analyses, we still found significant burdens of rare pLoF and, to a lesser 442 extent, damaging missense variants in constrained and DD-associated genes across all DD 443 cases (Supplementary Figure 4), even in those with a likely diagnostic *de novo* variant. In 444 recent work, we report that DD cases with a monogenic diagnosis have higher common, polygenic risk than controls, but significantly less risk than undiagnosed cases⁹. However, this 445 446 increased polygenic risk was driven nearly entirely by affected parents in the diagnosed cases. 447 When repeating the case/control analyses for trios with a diagnostic de novo variant and 448 removing those with an affected parent (n=2,425 DD cases remaining), we found significant 449 burdens of rare pLoF and missense variants (**Supplementary Table 12**). Similarly, recent work 450 in DDD has shown via burden analysis that ~12.5% of probands who currently have a single 451 molecular diagnosis may also have an additional *de novo* variant (mostly outside of known 452 genes) contributing to their condition⁴⁷. These results show that rare, inherited variants may be 453 contributing both to DD risk and to phenotypic presentations for DD cases both with and without 454 a monogenic diagnosis.

19

455 The significant overtransmission of rare pLoF and damaging missense variants from 456 parents to their children with DDs could arise via multiple mechanisms, including skewed 457 transmission of deleterious variants and/or an overall excess of such variants in the parents 458 themselves. When directly testing the latter, we found that parents of DD cases-even those 459 apparently unaffected by DD-related phenotypes-were significantly different from ancestry-460 matched controls in terms of their rare autosomal variant burden (Fig 2). The ORs were 461 attenuated compared to the DD cases, but were still significant for pLoF variants in constrained 462 and/or DD-associated genes (Supplementary Figure 5). While we did not see a significant 463 difference between unaffected mothers and unaffected fathers (Supplementary Figure 7), as 464 we might have predicted based on previous work on sex differences in reproductive success¹⁹, 465 we had incomplete power to detect modest differences. Beyond suggestive differences in the 466 ORs (e.g., in Fig 3a comparing male and female cases), we could not find any significant sex-467 specific differences in rare variant burdens in this cohort, in line with the recent findings from a 468 smaller cohort of intellectual disability²⁹. Similarly, while we have reported a higher burden of *de* novo variants in female cases previously^{24,51}, this excess was only found in well established DD-469 470 associated genes with no significant differences in the *de novo* burden between male and 471 female cases in genes not associated with DDs or in undiagnosed cases²⁴.

472 Beyond diagnostic status and sex, we investigated other factors that could be associated 473 with differential rare variant burden within our cohort. While we found no differential burden for 474 DD cases with an environmental exposure, we found a relationship between genetic burden and 475 the number of family members with similar clinical phenotypes as in recent work on polygenic 476 burden⁹. Here, we observed stronger rare variant burdens in constrained genes (**Fig 3c**) as the 477 number of affected relatives increases, which is consistent with recent findings from Huang*. 478 Wigdor* et al. when comparing polygenic scores for traits related to DD-risk, such as 479 educational attainment^{9,37}. However, we did not see this same trend for DD-associated genes 480 (Supplementary Figure 10), in contrast to their work and that from Urpa and colleagues²⁹. 481 We estimated that there was an excess of ~4,000 rare pLoF and damaging missense 482 variants in 4.183 DD cases with unaffected parents and no diagnostic *de novo* variant, and that 483 only ~10% of this excess could be found within the known monoallelic DD-associated genes 484 (Supplementary Figure 11), which prompted us to search for additional DD risk genes. Given 485 our lack of significance for individual genes in a per-gene burden test and the knowledge that de 486 novo variants play a substantial role in DDs, we applied the TADA Bayesian framework to 487 combine association evidence across multiple variant classes (e.g., *de novo* and inherited), 488 following work from studies of autism¹³. Nearly all of the 269 significant genes from TADA were

20

489 previously tied to DDs either via our prior study on *de novo* variants²⁴ or the manually-curated 490 DDG2P²⁸ list. Of these significant genes, 25 had minimal evidence for prior association as of 491 July 2023 and eleven have additional evidence supporting their association with DD based on 492 recent publications or their addition to other DD-related gene lists. These 25 genes, however, 493 only represent ~1.5% of the excess of rare pLoF and missense variants reported above, again 494 reinforcing that there are additional DD-associated risk loci to be discovered.

495 In the TADA analysis, we found that the major driver of association was the *de novo* 496 variant contribution (Fig 4b). Indeed, ~80% of the significant genes in the TADA analysis were 497 significant in our prior study of *de novo* variants alone²⁴ and, of those that were not significant, nearly all had suggestive evidence of association in the prior study (e.g., 20/25 with p-values < 498 499 0.001). For these genes with suggestive evidence previously, the inclusion of case/control and 500 inherited data provided enough additional statistical evidence to cross an exome-wide 501 significance threshold. In fact, the 25 TADA significant genes with minimal prior DD-association 502 evidence have a larger contribution from case/control or inherited analyses in the TADA 503 framework (Fig 4d-e; Supplementary Table 11) compared to the previously DD-associated 504 genes. For the eleven genes that were determined to have a high likelihood of being bona fide 505 DD-associated genes, we found that carriers of damaging variants in these genes were clinically diagnosed at a significantly lower rate than the cohort overall (14% versus 27.5%, $p = 4.8 \times 10^{-3}$). 506 507 In some of the genes with mixed evidence of being *bona fide* DD risk genes (e.g., ZNF644), the 508 association signal seemed to be driven primarily by the genes' low LOEUF scores (indicating 509 severe selective constraint; **Supplementary Figure 15**), which are strongly upweighted in the 510 TADA framework (Supplementary Table 13). A deeper analysis of the phenotypic profiles of 511 individuals with damaging genetic variation in these genes would provide more confidence in 512 their association with DDs.

513 Beyond the limitations of sample size-which is most notable when the cohort is 514 subdivided by diagnostic status and parental affected status-another limitation of this study is 515 the inability to know which variants are truly having a large functional impact. We aimed to 516 reduce this limitation by using strict allele frequency thresholds and filtering only to LOFTEE 517 high-confidence pLoF variants and missense variants with some evidence of being deleterious 518 (e.g., CADD \geq 20). However, we know that both rarity and *in silico* score deleteriousness 519 predictions are not sufficient to establish true variant impacts. For example, manual curation of 520 variants from the gnomAD database has shown that a large fraction of pLoF variants in 521 haploinsufficient genes associated with severe phenotypes have evidence that they are not true loss-of-function variants^{52,53}. Generally, there is a need to better understand the incomplete 522

21

penetrance of DD-risk variants, and analyses like those presented here would be improved by
the creation of approaches to estimate penetrance of risk variants from population cohorts with
longitudinal health data. Finally, we did not consider combinations of rare variants or the
interaction between rare and common variants, but a recent paper by Urpa et al. found evidence
of additivity between rare and common variants when studying individuals with intellectual
disability²⁹. A more integrative model that could account for the contributions of rare variants,
polygenic risk, and environmental factors would further improve assessment of DD risk.

530 While this study has established a role for rare, inherited variants in DD risk, it will do 531 little to improve the diagnostic rate of patients with DD primarily because these risk-increasing 532 variants are often inherited from apparently unaffected parents. The majority of diagnostic 533 pipelines do not consider variants inherited from unaffected parents, even with increasing evidence that incompletely penetrant variants can contribute both to diagnoses^{6,10} and to related 534 535 phenotypes²³. Diagnostic pipelines could be modified to allow for the consideration of these 536 lower penetrance variants, but such a change would need to balance the inclusion of these 537 potential risk variants with additional clinical curation time to evaluate the many other variants 538 that would also qualify for diagnostic consideration (e.g., ~19% of the trios have inherited at 539 least one rare pLoF variant in a constrained gene from an unaffected parent).

540 Potentially deeper phenotyping of the parents-including collecting data on their 541 educational history and phenotypes throughout life-would reveal sub-clinical phenotypes or 542 phenotypes that were stronger earlier in life that are not readily apparent in routine interactions 543 with clinicians when enrolling their children in studies such as DDD. Indeed, concurrent work in 544 birth cohorts (Malawsky et al. in prep) has shown that rare damaging variants in constrained 545 genes have a larger impact on cognitive ability early than later in life, implying that the parents 546 may have had learning difficulties in early childhood that subsequently improved. Considering 547 sub-clinical or earlier life phenotypes of parents could aid in identifying or prioritizing inherited 548 risk variation in DD cases, but would also necessitate additional genetic counseling 549 considerations, such as the impact on recurrence risk and the potential disclosure of these 550 incompletely penetrant variants to the parents.

551 There is also the possibility that parents appear to be able to tolerate apparently 552 damaging genetic variation due to other genetic, environmental, or stochastic factors that either 553 protect the parents or increase susceptibility in their children. For genetic factors, it has been 554 suggested that protective polygenic backgrounds (e.g., higher overall educational attainment 555 polygenic scores) or *cis*-regulatory variation that reduces the penetrance of the damaging 556 genetic variant⁵⁴ could contribute. While Kingdom et al.²³ reported evidence of educational

22

attainment polygenic scores protecting against the phenotypic impacts of carrying rare,
damaging genetic variants, we found no evidence of this in a study of 11,573 DD cases⁹. We
also found limited evidence for the protective effect of *cis*-regulatory variation in a set of 1700
trios where DD cases inherited rare, putatively damaging variants from unaffected parents⁵⁵.

561 Taken together, our findings indicate that most, if not all, DD cases in our cohort have 562 genetic contributions to risk from rare inherited variation, even those individuals with an 563 established diagnosis from a presumably large-effect *de novo* variant. We found stronger rare 564 variant burdens in undiagnosed cases and those with affected family members, as anticipated 565 given concurrent work on the contribution of polygenic scores⁹ to DDs and prior study of the factors that impact diagnostic rates². However, there is still much work to be done to both 566 567 identify additional DD-associated genes and to understand how these rare variants are 568 increasing risk for DD, including how they may interact with polygenic risk and environmental 569 risk factors. Larger sample sizes, particularly those with access to longitudinal and 570 comprehensive phenotype or health record information for the parents, will improve our 571 understanding of the genetic architecture of DDs and will provide more insight into the 572 mechanisms of incomplete penetrance for rare, inherited damaging genetic variation.

573

574 Acknowledgements

575 We thank the families and their clinicians for their participation and engagement, and our 576 colleagues who assisted in the generation and processing of data. The DDD study presents 577 independent research commissioned by the Health Innovation Challenge Fund (grant number 578 HICF-1009-003). The full acknowledgements can be found at www.ddduk.org/access.html. We 579 additionally thank Rachel Hobson and Rosemary Kelsell for DDD project management, and the 580 Human Genetics Informatics (HGI) group at Sanger and Aleksejs Sazonovs for fruitful 581 discussions and support during quality control of the data. This study makes use of DECIPHER, which is funded by the Wellcome Trust. This research was funded in part by Wellcome (grant 582 583 no. 220540/Z/20/A, "Wellcome Sanger Institute Quinguennial Review 2021–2026"). For the 584 purpose of open access, the authors have applied a CC-BY public copyright license to any 585 author accepted manuscript version arising from this submission.

586 Author Contributions

587 K.E.S. performed analyses and figure generation. K.E.S., V.K.C., E.J.G., P.D., E.M.W., S.J.L.,

588 T.S., R.Y.E., and G.G. were involved with data generation and quality control. V.K.C., J.M.F.,

23

- 589 D.S.M, S.J.L., and P.C. contributed to code, methods, or additional data. The DDD study is
- 590 supervised by C.F.W., H.C.M., H.V.F., and M.E.H. M.E.H. supervised this study. The primary
- 591 writing was completed by K.E.S. with input from V.K.C., C.F.W., H.C.M., and M.E.H. All authors
- 592 approved the final manuscript.

593 Competing interests

- 594 K.E.S. has received support from Microsoft for work related to rare disease diagnostics. E.J.G.
- is an employee of and holds shares in Insmed Incorporated. M.E.H. is a co-founder of,
- 596 consultant to and holds shares in Congenica, a genetics diagnostic company. The remaining
- 597 authors declare no competing interests.

598 Data Availability

- 599 Sequence and variant-level data and phenotype data from the DDD study data are available on
- 600 the European Genome-phenome Archive (EGA; https://www.ebi.ac.uk/ega/) with study ID
- 601 EGAS00001000775. Exome sequencing for the INTERVAL cohort is also available on EGA with
- 602 study ID EGAD00001002221. Previously described databases were from the Genome
- 603 Aggregation Database (gnomAD v2.1.1; https://gnomad.broadinstitute.org/downloads) and the
- 604 Developmental Disorders Genotype-Phenotype Database (DDG2P;
- 605 <u>https://www.ebi.ac.uk/gene2phenotype/downloads</u>).

606 Code Availability

- 607 Analyses were primarily performed with Python and R (version 4.2.0). The R code used to
- 608 generate the TADA results are available at: <u>https://github.com/talkowski-lab/TADA_2022</u>

609 References

- Clark, M. M. *et al.* Meta-analysis of the diagnostic and clinical utility of genome and exome
 sequencing and chromosomal microarray in children with suspected genetic diseases. *NPJ Genom Med* 3, 16 (2018).
- Wright, C. F. *et al.* Genomic Diagnosis of Rare Pediatric Disease in the United Kingdom
 and Ireland. *N. Engl. J. Med.* 388, 1559–1571 (2023).
- Girirajan, S. & Eichler, E. E. Phenotypic variability and genetic susceptibility to genomic disorders. *Hum. Mol. Genet.* **19**, R176–87 (2010).
- 617 4. Kirov, G. *et al.* The penetrance of copy number variations for schizophrenia and developmental delay. *Biol. Psychiatry* **75**, 378–385 (2014).
- 619 5. Männik, K. *et al.* Copy number variations and cognitive phenotypes in unselected
 620 populations. *JAMA* 313, 2044–2054 (2015).
- 621 6. de Masfrand, S. et al. Penetrance, variable expressivity and monogenic

600		nourodouclonmental disorders Fur / Mad Canat CO 104022 (2024)
622	7	Niomi M E K at al Common genetic variante contribute to rick of rare sovere
624	1.	nuerii, M. E. K. <i>et al.</i> Common genetic valiants commodel to fisk of falle severe
625	0	Kurki M L at a Contribution of rare and common variants to intellectual disability in a sub
626	о.	isolate of Northern Finland, Not Commun 10 , 410 (2010)
627	0	Solate of Northern Finland. Nat. Commun. 10 , 410 (2019).
620	9.	nually, Q. Q. et al. Dissecting the contribution of common variants to fisk of fate
620	10	Aitkon S. et al. Finding Diagnostically Llosful Datterns in Quantitative Datterns i
620	10.	Alken, S. et al. Finding Diagnostically Oseiul Patients in Quantitative Phenotypic Data. Am.
030		J. Hulli. Genet. 103, 933–940 (2019).
031	11.	Singh, T. et al. Rate could variants in ten genes conter substantial risk for schizophrenia.
03Z	10	Nature 604 , 509–516 (2022). Relmer D. S. et al. Exemple acquencing in bipolar disorder identifies AKAR11 as a rick gapa
624	12.	charad with achizantronia. Not Const E4 , 541, 547 (2022)
034 625	10	Shared with Schizophienia. <i>Nat. Genet.</i> 34 , 541–547 (2022).
626	15.	ru, J. W. et al. Rate couling variation provides insight into the genetic architecture and
627	11	Milfort A P. et al. Depent ultre rare inherited variants implicate new autiem condidete rick
620	14.	villeri, A. D. <i>et al.</i> Recent una-rare innenieu variants implicate new autism candidate risk
620	15	Zhou X at al Integrating do novo and inherited variants in 42 607 outiom appear identifies
640	15.	$2100, \Lambda$. <i>et al.</i> Integrating de 1000 and interned variants in 42,007 autism cases identifies
6/1	16	Eni25 Collaborativo, Electronic address: s borkovic@unimolb.odu au & Eni25 Collaborativo
642	10.	Liltra-Rare Cenetic Variation in the Enilensies: A Whole-Exome Sequencing Study of
643		17 606 Individuals Am I Hum Genet 105 267–282 (2010)
644	17	Eni25 Collaborative Chen S. Neale B. M. & Berkovic S. F. Shared and distinct ultra-rare
645		cenetic risk for diverse epilepsies: A whole-exome sequencing study of 54 423 individuals
646		across multiple genetic ancestries medRyiv (2023) doi:10.1101/2023.02.22.23286310
647	18	Ganna A et al Ultra-rare disruptive and damaging mutations influence educational
648	10.	attainment in the general population. Nat. Neurosci 19 1563–1565 (2016)
649	19	Gardner, F. J. <i>et al.</i> Reduced reproductive success is associated with selective constraint
650	10.	on human genes. Nature 603, 858–863 (2022)
651	20.	Chen, CY. et al. The impact of rare protein coding genetic variation on adult cognitive
652		function. <i>Nat. Genet.</i> 55 , 927–938 (2023).
653	21.	Rolland, T. et al. Phenotypic effects of genetic variants associated with autism. Nat. Med.
654		29 . 1671–1680 (2023).
655	22.	Fenner, E. et al. Rare coding variants in schizophrenia-associated genes affect generalised
656		cognition in the UK Biobank. <i>bioRxiv</i> (2023) doi:10.1101/2023.08.14.23294074.
657	23.	Kingdom, R., Beaumont, R. N., Wood, A. R., Weedon, M. N. & Wright, C. F. Genetic
658		modifiers of rare variants in monogenic developmental disorder loci. Nat. Genet. 56. 861-
659		868 (2024).
660	24.	Kaplanis, J. et al. Evidence for 28 genetic disorders discovered by combining healthcare
661		and research data. Nature 586, 757–762 (2020).
662	25.	He, X. et al. Integrated model of de novo and inherited genetic variants yields greater
663		power to identify risk genes. PLoS Genet. 9, e1003671 (2013).
664	26.	Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in
665		141,456 humans. Nature 581, 434–443 (2020).
666	27.	Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the
667		deleteriousness of variants throughout the human genome. Nucleic Acids Res. 47, D886-
668		D894 (2019).
669	28.	Thormann, A. <i>et al.</i> Flexible and scalable diagnostic filtering of genomic variants using G2P
670		with Ensembl VEP. Nat. Commun. 10, 2373 (2019).
671	29.	Urpa, L. et al. Evidence for the additivity of rare and common variant burden throughout the
672		spectrum of intellectual disability. Eur. J. Hum. Genet. 32, 576–583 (2024).

- 25
- 30. Jacquemont, S. *et al.* A higher mutational burden in females supports a 'female protective model' in neurodevelopmental disorders. *Am. J. Hum. Genet.* **94**, 415–425 (2014).
- Wigdor, E. M. *et al.* The female protective effect against autism spectrum disorder. *Cell Genom* 2, 100134 (2022).
- Beciphering Developmental Disorders Study. Prevalence and architecture of de novo
 mutations in developmental disorders. *Nature* 542, 433–438 (2017).
- Antaki, D. *et al.* A phenotypic spectrum of autism is attributable to the combined effects of
 rare variants, polygenic risk and sex. *Nat. Genet.* 54, 1284–1292 (2022).
- Blencowe, H. *et al.* Preterm birth-associated neurodevelopmental impairment estimates at
 regional and global levels for 2010. *Pediatr. Res.* **74 Suppl 1**, 17–34 (2013).
- 683 35. Coste, J. *et al.* Risk of early neurodevelopmental disorders associated with in utero
 684 exposure to valproate and other antiepileptic drugs: a nationwide cohort study in France.
 685 Sci. Rep. 10, 17362 (2020).
- 36. Deciphering Developmental Disorders Study. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* 542, 433–438 (2017).
- 37. Lee, J. J. *et al.* Gene discovery and polygenic prediction from a genome-wide association
 study of educational attainment in 1.1 million individuals. *Nat. Genet.* 50, 1112–1121
 (2018).
- Agarwal, I., Fuller, Z. L., Myers, S. R. & Przeworski, M. Relating pathogenic loss-of-function
 mutations in humans to their evolutionary fitness costs. *Elife* 12, (2023).
- Samocha, K. E. *et al.* Regional missense constraint improves variant deleteriousness
 prediction. *bioRxiv* 148353 (2017) doi:10.1101/148353.
- 40. Satterstrom, F. K. *et al.* Large-Scale Exome Sequencing Study Implicates Both
 Developmental and Functional Changes in the Neurobiology of Autism. *Cell* 180, 568–
 584.e23 (2020).
- 698 41. Home OMIM. https://omim.org/.
- 42. DiStefano, M. T. *et al.* The Gene Curation Coalition: A global effort to harmonize genedisease evidence resources. *Genet. Med.* 24, 1732–1742 (2022).
- 43. Martin, A. R. *et al.* PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nat. Genet.* **51**, 1560–1565 (2019).
- 44. Weghorn, D. *et al.* Applicability of the Mutation-Selection Balance Model to Population
 Genetics of Heterozygous Protein-Truncating Variants in Humans. *Mol. Biol. Evol.* 36,
 1701–1710 (2019).
- 45. Collins, R. L. *et al.* A cross-disorder dosage sensitivity map of the human genome. *Cell*185, 3041–3055.e25 (2022).
- 46. Cospain, A. *et al.* FOSL2 truncating variants in the last exon cause a neurodevelopmental disorder with scalp and enamel defects. *Genet. Med.* 24, 2475–2486 (2022).
- 47. Chundru, V. K. *et al.* Federated analysis of the contribution of recessive coding variants to
 29,745 developmental disorder patients from diverse populations. *bioRxiv* (2023)
 doi:10.1101/2023.07.24.23293070.
- 48. Martin, H. C. *et al.* Quantifying the contribution of recessive coding variation to developmental disorders. *Science* 362, 1161–1164 (2018).
- 49. Landrum, M. J. *et al.* ClinVar: improvements to accessing data. *Nucleic Acids Res.* 48, D835–D844 (2020).
- 717 50. Park, J. H. SLC39A8-CDG. in *GeneReviews®* [Internet] (University of Washington, Seattle, 2023).
- 51. Deciphering Developmental Disorders Study. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* 542, 433–438 (2017).
- 52. Singer-Berk, M. *et al.* Advanced variant classification framework reduces the false positive
 rate of predicted loss-of-function variants in population sequencing data. *Am. J. Hum. Genet.* **110**, 1496–1508 (2023).

26

- 53. Gudmundsson, S. *et al.* Exploring penetrance of clinically relevant variants in over 800,000
 humans from the Genome Aggregation Database. *bioRxiv* (2024)
 doi:10.1101/2024.06.12.593113.
- 54. Castel, S. E. *et al.* Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. *Nat. Genet.* **50**, 1327–1334 (2018).
- 55. Wigdor, E. M. *et al.* Investigating the role of common cis-regulatory variants in modifying
 penetrance of putatively damaging, inherited variants in severe neurodevelopmental
 disorders. *Sci. Rep.* 14, 8708 (2024).
- 56. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler
 transform. *Bioinformatics* 25, 1754–1760 (2009).
- 57. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
 (2013) doi:10.48550/ARXIV.1303.3997.
- 736 58. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples.
 bioRxiv 201178 (2018) doi:10.1101/201178.
- 59. Moore, C. *et al.* The INTERVAL trial to determine whether intervals between blood
 donations can be safely and acceptably decreased to optimise blood supply: study protocol
 for a randomised controlled trial. *Trials* 15, 363 (2014).
- Di Angelantonio, E. *et al.* Efficiency and safety of varying the frequency of whole blood
 donation (INTERVAL): a randomised trial of 45 000 donors. *Lancet* **390**, 2360–2371 (2017).
- 51. Sifrim, A. *et al.* Distinct genetic architectures for syndromic and nonsyndromic congenital
 heart defects identified by exome sequencing. *Nat. Genet.* 48, 1060–1065 (2016).
- Singh, T. *et al.* Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders. *Nat. Neurosci.* **19**, 571–577 (2016).
- 63. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, (2021).
- 64. McLaren, W. et al. The Ensembl Variant Effect Predictor. Genome Biol. 17, 122 (2016).
- 65. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* 26, 2867–2873 (2010).
- Price, A. L. *et al.* Long-range LD can confound genome scans in admixed populations. *Am. J. Hum. Genet.* 83, 132–5; author reply 135–9 (2008).
- 753 67. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation.
 754 *Nature* 526, 68–74 (2015).
- 755 68. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.*756 2, e190 (2006).
- Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide
 association studies. *Nat. Genet.* 38, 904–909 (2006).
- 759 70. Wright, C. F. *et al.* Making new genetic diagnoses with old data: iterative reanalysis and
 760 reporting from genome-wide data in 1,133 families with developmental disorders. *Genet.* 761 *Med.* 20, 1216–1223 (2018).
- 762
- 763 Malawsky D. S., Musa M. K., Danacek P., Wootton O., Wade E., Huang W., Huang Q., Arden
- R., Lindsay S., Hurles M. E., Martin H. C. The differential effects of common and rare genetic
- 765 variants on cognitive performance across development. *In preparation*.

766 Methods

767 Samples included in analyses

768 Deciphering Developmental Disorders (DDD)

769 Patients with severe, undiagnosed developmental disorders were recruited from 24 regional 770 genetics services within the United Kingdom National Health Service and the Republic of Ireland 771 between 2011 and 2015. Families gave informed consent to participate, and the study was 772 approved by the UK Research Ethics Committee (10/H0305/83 granted by the Cambridge 773 South Research Ethics Committee, and GEN/284/12 granted by the Republic of Ireland 774 Research Ethics Committee). The inclusion criteria included neurodevelopmental conditions, 775 congenital, growth or behavioral abnormalities, and dysmorphic features. Additional details on 776 sample collection, exome sequencing, alignment, variant calling (inherited and *de novo*) and variant annotation have been described previously^{24,51}. In brief, exome capture was carried out 777 778 with either Agilent SureSelect Human All Exon V3 or V5 baits. Reads were aligned to the

- GRCh37 1000 Genomes Project phase 2 reference (hs37d5) using BWA aln and BWA
- 780 mem^{56,57}. Variants were called using GATK's HaplotypeCaller, CombineGVCFs, and
- 781 GenotypeGVCFs (GATK version 3.5.0)⁵⁸, and then restricted to merged bait regions from the
- two capture kits plus 100 base pairs of padding on either side.
- 783

After removing samples that had withdrawn consent for research, these analyses involve 13,452

- individuals with developmental disorders: 9,856 individuals with complete trios from 9,305
- families, and 3,596 non-trio individuals. *De novo* and inherited variants from a subset of these
- individuals have been published previously^{24,48}.

788 Controls from the INTERVAL study

789 The INTERVAL study⁵⁹ was a randomized controlled trial of the safety and efficacy of varying the duration between blood donations⁶⁰. As part of this work, 50,000 presumed healthy adults 790 791 (18 years or older) were consented and recruited from NHS Blood and Transplant blood 792 donation centers across England. DNA extraction and genotyping was funded by the National 793 Institute of Health Research (NIHR), the NIHR BioResource (http://bioresource.nihr.ac.uk/) and 794 the NIHR Cambridge Biomedical Research Centre (www.cambridge-brc.org.uk). The academic 795 coordinating center for INTERVAL was supported by core funding from: NIHR Blood and 796 Transplant Research Unit in Donor Health and Genomics, UK Medical Research Council 797 (G0800270), British Heart Foundation (SP/09/002) and NIHR Research Cambridge Biomedical 798 Research Centre. A complete list of the investigators and contributors to the INTERVAL trial is provided in Moore et al.⁵⁹ and at www.intervalstudy.org.uk/about-the-study/whos-799 800 involved/interval-contributors/.

- 801
- 802 For a subset of the INTERVAL cohort, exome sequencing was performed as described
- 803 previously^{61,62} using the Agilent SureSelect Human All Exon V5 baits. In this study, we had
- access to exome sequencing data from 4,502 individuals. As detailed below, 3,943 of these

28

individuals were retained for analyses after QC and selecting for inferred northwestern

806 European genetic ancestry.

807 Data quality control

808 Variant quality control

809 We performed quality control on autosomal single nucleotide variants (SNVs) and

- insertions/deletions (indels). We tested a range of values for both genotype-level (e.g., genotype
 guality) and variant-level metrics (e.g., VQSLOD).
- 812

For SNVs, we selected our quality control filters by evaluating (1) the transmission rate of rare synonymous variants from a parent to their child (allele count [AC] = 2, only seen in parent and their child) and (2) sensitivity to retaining known *de novo* variants²⁴. The final filters for SNVs were:

817 • VQSLOD ≥ -2

- 818 Genotype quality $(GQ) \ge 20$
- 819 Depth (DP) ≥ 7
- p-value for sampling the observed allele balance under a binomial model, assuming an allele balance of 0.5 for heterozygous sites > 1x10⁻³
- 822 823
- Fraction of non-missing genotypes passing genotype-level QC thresholds > 0.5

Using the above thresholds, our transmission rate of singleton synonymous variants was 0.500
(e.g., perfect balance of transmitted to nontransmitted variants) and our sensitivity to recovering
known *de novo* variants was 88%. Analyses were performed with bcftools⁶³.

827

For indels, we selected our quality control filters by evaluating (1) the transmission rate of rare inframe variants in unconstrained (pLI < 0.9), non-monoallelic DD-associated genes from a parent to their child, (2) the sensitivity to retaining known *de novo* variants, and (3) the frameshift:nonsense ratio, which we expected to be 1-1.2. The final filters for indels were:

- VQSLOD ≥ -2
 - GQ ≥ 25
- DP ≥ 10
 - Allele balance (AB) > 0.3
- 835 836

832

833

834

Using the above thresholds, our transmission rate of rare inframe variants in unconstrained,
non-monoallelic DD-associated genes was 0.490 (not significantly different from the expected
0.5), our sensitivity to recover known *de novo* indels was 72%, and our frameshift:nonsense
ratio was 1.10. Finally, we removed indels found in the same gene and sample, which
represented ~4% of all indels with minor allele frequency (MAF) < 1%. These indels were often
part of a complex mutational event that would require haplotype-aware annotations to resolve.

843 Variants annotation and filtering

Variants were annotated with VEP⁶⁴, including the LOFTEE plugin²⁶. We additionally annotated
 all variants with CADD v1.4²⁷ and MPC³⁹.

29

846

848

849

847 Finally, variants in the joint VCF were filtered using the following criteria:

- gnomAD v2.1 raw allele frequency < 0.001
- Coding consequences based on VEP annotations in protein-coding genes
- On autosomes
- 850 851
- 852 Note that nearly all analyses in this work use a lower gnomAD allele frequency, namely MAF <
- $1x10^{-5}$, as well as a filter on dataset allele frequency. Additionally, we used the worst
- 854 consequence in protein-coding transcripts as the mutational consequence for variants.

855 Sample relatedness

856 KING⁶⁵ was run to determine relatedness between samples in this joint-called cohort. We used

857 common variants (MAF > 1%) that passed the SNV filters listed above and, after applying those

filters, had low missingness (< 5%). Related individuals were defined as those with a kinship

- 859 coefficient > 0.04419417, which is the lower bound cut-off for third-degree relatives. A list of
- unrelated parents (n = 18,494) and probands (n = 10,613) was created to maximize the number
- 861 of samples retained.

862 Determining a set of individuals with inferred European genetic ancestry

863 For analyses comparing cases to controls, we needed to identify a subset of individuals who

- 864 had similar inferred genetic ancestries. Given that both DDD and INTERVAL were primarily
- 865 collected in the United Kingdom, our largest genetic ancestry group for comparison was
- 866 individuals with inferred northwestern European genetic ancestry (e.g., matching ancestry
- historically tied to the British Isles).
- 868
- After splitting multiallelic variants, we selected common (MAF > 1%) single nucleotide polymorphisms (SNPs) that passed QC filters (GQ \ge 20 and DP \ge 7) and had < 10% genotype missingness. Ambiguous SNPs and indels were removed as were variants in 24 long-range
- 872 linkage disequilibrium (LD) regions, such as the HLA⁶⁶. Overlapping SNPs from the 1000
- 673 Genomes Phase 3 individuals⁶⁷ were identified and filtered to those with MAF > 1% and
- genotype missingness > 10%. SNPs were then merged between our DDD/INTERVAL joint-call
- and the 1000 Genomes, and LD-pruned ($r^2 < 0.2$), leaving 32,413 variants. Principal component
- analysis (PCA) was performed on the 1000 Genomes samples using the smartpca function from
- 877 EIGENSOFT^{68,69}, with the DDD and INTERVAL samples projected onto the resulting PCs.
- 878 Further dimensionality reduction was done using UMAP on the first 20 PCs, and the individuals
- of European genetic ancestry were defined as those overlapping with the labeled European
- ancestry individuals from the 1000 Genomes Project.
- 881

We needed to generate PCs within the inferred European genetic ancestry subgroup to use as covariate in the case/control regressions. To do this, we extracted variants from a set of

- unrelated parents and controls as defined by KING⁶⁵ with inferred European genetic ancestry.
- 885 We performed similar variant filtering as above (e.g., MAF > 1%, passing QC, not in long-range
- LD regions) and additionally removed variants if they had > 5% missingness in any of three
- groups: (1) parents sequenced using the V3 exome capture kit, (2) parents sequenced using the

30

V5 exome capture kit, or (3) controls. Samples that were outliers for differential missingness or
 differential MAF were excluded. PCs were generated with smartpca. The remaining samples
 (e.g., the DDD children) were projected onto these PCs (Supplementary Figure 1).

891 **Definitions of case subsets**

892 Affected parents

Parents were defined as potentially "affected" if clinicians listed HPO terms for them and/or noted that they had similar phenotypes as their affected children. These analyses were only done for complete parent-child trios. In total, 1,462 trios had at least one affected parent, of which 1,246 were of inferred European genetic ancestry for use in case/control analyses and 1,306 were used for TDT analyses.

898 Diagnosed with a de novo variant

DD cases in complete parent-child trios were considered "diagnosed" with a *de novo* variant if

900 (1) a clinician had annotated a *de novo* variant as pathogenic or likely pathogenic, (2) they were

901 considered diagnosed with a *de novo* variant as part of an iterative analysis of the first ~1k

902 cases included DDD⁷⁰, and/or (3) they had a *de novo* loss-of-function variant in a monoallelic

DD-gene that had yet to be interpreted by a clinician. In total, 3,215 trios had a child with a *de*

novo diagnosis, of which 2,679 were of inferred European genetic ancestry for use in the

905 case/control analyses and 3,155 were used for TDT analyses.

906 Unaffected parents with no de novo diagnosis

907 These were trios that did not meet either criteria for having an affected parent or being

908 diagnosed with a *de novo* variant. In total, 4,183 trios of inferred European genetic ancestry

909 qualified for use in the case/control analyses and 5,124 were used for TDT analyses.

910 Environmental exposures

911 We used a list of 2,637 DD cases with environmental exposures as defined in Wright et al.².

912 Specifically, these DD cases were born prematurely (< 37 weeks gestation), had mothers with

913 diabetes, and/or were exposed to antiepileptic medications in utero. Of these, we focused only

914 on those with unaffected parents and no *de novo* diagnosis; for the case/control analysis, this

915 was 879 individuals, and for the TDT analysis it was 1,075 individuals.

916 Affected family members

- 917 We used a list of DD cases with affected family members as defined in Wright et al. and
- 918 Huang^{*}, Wigdor^{*} et al.^{2,9}. DD cases were split in 5 categories: (1) both parents are affected and
- 0 or more siblings are affected (n=87); (2) one parent is affected and at least one sibling is
- 920 affected (n=187); (3) one parent is affected and no siblings are affected (n=345); (4) neither
- parent is affected but at least one sibling is affected (n=645); (5) a non-first degree family

922 member is affected (n=469). All numbers reported were for individuals of inferred European

923 genetic ancestry.

31

924 Definition of gene sets

925 Constrained

- 926 We defined constrained genes as those with a pLI (probability of being loss-of-function
- 927 intolerant) score ≥ 0.9 as defined in gnomAD v2.1²⁶ (n=2,699 autosomal genes).

928 DD-associated

- 929 We defined monoallelic DD-associated genes using DDG2P²⁸, downloaded on June 29, 2023.
- 930 Specifically, we selected genes with an allelic requirement of "monoallelic autosomal" with
- 931 either "definitive" or "strong" for the confidence category (n=666 genes).

932 Unconstrained not DD-associated

- 933 Genes that had pLI < 0.9 and were not considered monoallelic DD-associated genes were
- 934 defined as unconstrained and not DD-associated (n=15,667 autosomal genes).

935 Other gene sets evaluated

- 936 Most analyses in the main text focused on all genes and the three gene lists described above.
- However, we also tested (1) all unconstrained genes (pLI < 0.9, n=15,911 autosomal genes); (2)
- all genes that were not considered monoallelic DD-associated genes (n=17,944 autosomal
- genes); (3) genes that were considered both constrained and monoallelic DD-associated genes
- 940 (n=422 autosomal genes); and (4) genes in the top decile of MOEUF (missense observed /
- 941 expected upper bound fraction, n=1,686 autosomal genes) from gnomAD v2.1²⁶.

942 Case versus control regressions

- 943 We included all DD cases of inferred European genetic ancestry in these analyses, including 944 those that were not in complete parent-child trios (n=10,644 cases versus 3,943 controls). We
- focused on rare variants (gnomAD allele frequency $\leq 10^{-5}$ and allele count in the dataset < 7,
- equivalent to an allele frequency $\sim 10^{-4}$) and, for each individual, determined the count of
- 947 alternative alleles, split by mutational consequence (e.g., loss-of-function, missense) and gene
- set (e.g., constrained genes, known DD genes). For all tested mutation and gene set
- 949 combinations, we used a logistic regression that corrected for sex, the first 20 principal
- 950 components from the inferred European genetic ancestry group only analysis, and the number
- 951 of rare autosomal variants per person. The final correction has been used in recent
- 952 publications¹², and should be a conservative correction. Specifically, we ran:
- 953 *is.case* ~ *variant.count* + *sex* + *PC*1 + *PC*2...+*PC*20 + $n_{rarevar}$ 954 where *is.case* is 0 for controls and 1 for cases, *variant.count* is the number of rare variants in a 955 given gene set, *sex* is the sex of the individual, *PC1* through *PC20* are the principal 956 components, and $n_{rarevar}$ is the number of rare autosomal variants. Intercepts from these 957 regressions were transformed into odds ratios using the exponential function.
- 958
- See Supplemental Note for more details about modifications to these regressions, includingremoving the correction for the number of rare variants per person.

32

961 Testing significance between two regressions

To determine if the results of various regressions were different from each other (as in

963 **Supplementary Table 3**), a Wald test was performed to obtain Z-scores that were transformed 964 into p-values. Specifically,

965

$$Z = (b_1 - b_2) / \sqrt{(se_1^2 + se_2^2)}$$

966 where b_1 is the estimate and se_1 is the standard error from the logistic regression for the first 967 group, respectively, and b_2 is the estimate and se_2 is the standard error from the logistic 968 regression for the second group, respectively.

969 *Removing known* de novo variants

970 We repeated the regressions above after removing known *de novo* variants²⁴ from the trio case

971 counts (**Supplementary Figure 4**). We note that this is not a fair comparison, given that the

972 control individuals will also have a background rate of *de novo* variants, but since we do not

have parental information, we do not know which variants to remove.

974 Transmission disequilibrium test (TDT)

975 We included all DD cases, regardless of inferred genetic ancestry, that were in complete parent-

- 976 child trios. For families with multiple affected children, we randomly selected one child as the
- 977 representative trio for the family to avoid double counting, which resulted in 9,305 trios for
- analysis. We focused on rare variants (gnomAD allele frequency $\leq 10^{-5}$), specifically transmitted doubletons (AC=2 seen in a parent and their child) versus nontransmitted singletons (AC=1
- 980 seen only in a parent). The rate ratio reported was the number of transmitted variants to the
- 981 number of nontransmitted variants with the confidence intervals determined using the
- 982 rateratio.test() function in R. P-values were calculated using a χ^2 test.
- 983 Testing significance between two TDT rate ratios
- 984 To determine if the TDT results were different between two groups (as in **Supplementary Table**
- 985 **6**), we compared the transmitted to nontransmitted counts of each group in a 2x2 table and
- 986 performed a χ^2 test to obtain a p-value.

987 Other regressions

988 For all other regressions in this work, such as parents versus controls or male cases versus

- female cases, we used rare variants as defined above (e.g., gnomAD allele frequency $\leq 10^{-5}$ and allele count in the dataset < 7) in individuals of inferred European genetic ancestry. These
- and allele count in the dataset < 7) in individuals of inferred European genetic ancestry. These
 were tested in a logistic regression that corrected for sex (except for sex-specific comparisons),
- the first 20 principal components from the inferred European genetic ancestry group only
- analysis, and the number of rare autosomal variants per person. We note that for the affected
- family member analyses, we further restricted to only unrelated individuals of inferred European
- ancestry to match similar work for polygenic scores⁹, but report that overall restricting to only
- 996 unrelated individuals has a minimal impact on the results (Supplementary Figure 3).

997 Per gene burden testing

- 998 For every gene, we tallied the number of individuals of inferred European genetic ancestry who
- carried a rare (gnomAD allele frequency $\leq 10^{-5}$ and allele count in the dataset < 7) variant in the 999
- 1000 gene, split by mutational consequence. We then performed a Fisher's exact test to compare the
- 1001 rate of carriers in cases (n=10,644) versus controls (n=3,943) for pLoF variants, damaging
- 1002 missense (CADD \geq 20) variants, pLoF and missense variants, and synonymous variants. We
- 1003 used an exome-wide significance threshold of 2.8x10⁻⁶.

1004 **Estimating excess variants**

1005 To determine the excess number of variants per mutation consequence and gene list, we 1006 calculated the expected number of variants in cases by using the rate seen in controls and 1007 further corrected for the synonymous case versus control rate for the given gene list.

1008 Specifically, this was estimated as:

1009
$$excess_{case} = nvar_{case} - \left(\frac{nvar_{control}}{n_{control}}\right) \times (n_{case}) \times (syn_{case/control\,rate})$$

1010 where *nvar_{case}* is the number of variants in cases, *nvar_{control}* is the number of variants in controls,

1011 n_{case} is the number of cases, $n_{control}$ is the number of controls, and $syn_{case/control rate}$ is the rate of

1012 synonymous variants in cases divided by the rate of synonymous variants in controls.

1013 **Population attributable risk**

- 1014 To estimate the population attributable risk (PAR), we used the following formula:
- p * (RR 1)....

1015
$$PAR = \frac{1}{p * (RR - 1) + 1}$$

where p is the probability of inheriting a damaging rare variant from an unaffected parent and 1016 1017 *RR* is the relative risk for DDs for those that inherited the given rare variants.

- 1018
- 1019 To estimate the RR, we used an estimated population prevalence of DDs of 1% and converted
- 1020 the odds ratio from the case/control regression of trios with unaffected parents and no 1021 diagnostic *de novo* variant for pLoF variants in constrained genes with the following formula:
- OR 1022 $RR = \frac{1}{1 - p(DD \mid no \text{ inherited rare variants}) + (p(DD \mid no \text{ inherited rare variants}) * OR)}$ 1023

1024 To estimate p, we took the rate of controls who carry a rare pLoF variant in a constrained gene 1025 (~ 0.22) and used it in the following formula:

- 1026 $p = p(one \ parent \ carrier) * p(inheriting \ rare \ variant \ | \ one \ parent \ carrier) +$
- 1027 p(both parents carriers) * p(inheriting rare variant | both parents carriers)

1028 Evaluating the shet burden

1029 In line with previous work¹⁹, we sought to determine a per person cumulative burden of rare 1030 variants by combining the shet selection coefficients of each gene affected by these variants via 1031 the following equation (taken from Gardner et al):

$$s_{het[i,v]} = 1 - \prod_{g} \square (1 - s_{het[i,v,g]})$$

- 1033 where $s_{het[i,v]}$ indicates individual *i*'s s_{het} burden for variant class *v* and $s_{het[i,v,g]}$ indicates the s_{het} 1034 score for gene *g* with a qualifying annotation for variant class *v* in individual *i*. 1035
- 1036 Here, qualifying variants were defined as:
 - gnomAD allele frequency $\leq 10^{-5}$ and allele count in the dataset < 7
 - pLoF: LOFTEE high-confidence & CADD ≥ 25
 - Missense: MPC ≥ 2 & CADD ≥ 25
 - Synonymous: no additional filters
- 1040 1041

1037

1038

1039

1042 Variants that fell into a gene without an s_{het} score were not included, and individuals with no 1043 qualifying variants were given an s_{het} burden score of 0. The specific variant filters were chosen 1044 to match what was done previously¹⁹. Additionally, compared to Gardner et al., we updated the 1045 selective coefficients scores to be those from Agarwal et al.³⁸ (called *hs* in their work, but s_{het} 1046 here for consistency with Gardner et al.).

1047

1048 We did two comparisons with these scores – within family and case versus control. For within 1049 family, we removed all known *de novo* variants²⁴ to ensure that we were comparing only 1050 inherited variants. These s_{het} burden scores were calculated for all parents and children included 1051 in the trios used for TDT analyses (n=9,305). We compared the s_{het} burden score in children to 1052 the scores seen in parents with a Wilcoxon rank sum test. Bootstrapping with 1000 replicates 1053 was used to determine the median difference in s_{het} burden scores as well as the 95% 1054 confidence intervals.

1055

For the case versus control comparison, we included known *de novo* variants for the DD cases
as we had no ability to remove such variants from the controls. We specifically focused on a set
of 8,062 unrelated DD cases of inferred European genetic ancestry to compare to the 3,943

- 1059 controls. Here, we used a probit regression of case status by scaled *s*_{het} burden scores,
- 1060 corrected as above for sex, the first 20 principal components from the inferred European genetic 1061 ancestry group only analysis, and the number of rare autosomal variants per person.
- 1062 $is.case \sim scaled.s_{het}burden + sex + PC1 + PC2... + PC20 + n_{rarevar}$ 1063 In the regression, controls were given a weight of 1 and cases were given a weight of

1064
$$\frac{(1-P) * K}{P * (1-K)}$$

where *P* is the fraction of cases in the regression and *K* is an estimate of the population
prevalence (here, 1%). This regression was also repeated to compare only unrelated DD cases
with a *de novo* diagnosis (n=1,905) to controls and unrelated, undiagnosed DD cases (n=6,157)
to controls.

35

1069 Determining variance explained on the liability scale

1070 For both the s_{het} burden scores and the rare variant counts, we wanted to determine the

1071 variance explained on the liability scale. For the s_{het} burden scores, we used the estimate from 1072 the probit regression above and transformed it into a percent variance explained on the liability 1073 scale via the following equation:

- 1074 $\frac{\beta^2}{1+\beta^2}$
- 1075 where β is the beta from the regression.

1076

1077 For the rare variant counts, we ran a probit regression with the same weights as mentioned 1078 above, replacing the scaled s_{het} burden scores with the total counts of rare pLoF and damaging 1079 missense variants across all genes.

1080 Transmission and De Novo Association (TADA) test

TADA is a Bayesian framework that incorporates per-gene mutation rates, sample size, and a
prior on the risk of a given variant in each gene to determine a Bayes Factor (BF) to measure
the statistical evidence of association of each gene to the condition being tested. It has been
used extensively to identify genes associated with autism^{13,40}. We used the TADA code from Fu
et al. (<u>https://github.com/talkowski-lab/TADA_2022</u>) to analyze our data.

1086

1087 We used mutation rates from gnomAD v2.1 and recalculated the mutation rate expected for 1088 missense variants with MPC \ge 2 ("misB") and 1 \le MPC < 2 ("misA") by dividing the total 1089 missense mutation rate by the fraction of such variants possible in the gene. These missense 1090 categories matched those that were previously used¹³.

1091

1095

1092 We additionally retrained the priors (or weights) used in TADA using the following data

- 1093 (displayed in Supplementary Figure 16):
 1094 De novo analysis: de novo variants figure 16
 - De novo analysis: de novo variants from 31,058 individuals with DD²⁴, which includes all DDD samples used in other analyses in this work
- 1096Case versus control analysis: rare (gnomAD allele frequency $\leq 10^{-5}$ and allele count in1097the dataset < 7) variant counts from 2,790 non-trio DD cases and from 3,943 controls of</td>1098inferred European genetic ancestry
- Inherited analysis: rare (gnomAD allele frequency ≤ 10⁻⁵ and allele count in the dataset <
 7) variants that were either transmitted from a parent to their child or not transmitted
 from parent to child in 9,305 trios
- 1102

1103 A comparison of the priors used in this analysis versus the Fu et al. paper is available in 1104 **Supplementary Table 13**, although we note that using the mutation rates and priors from the 1105 previous paper gives very similar results (**Supplementary Figure 17**). All results from the TADA 1106 analysis, including mutation rates, priors, Bayes Factors, and counts per gene are available in 1107 **Supplementary Table 10**.