Contribution of Retrotransposition to Developmental Disorders

Eugene J. Gardner[1], Elena Prigmore[1], Giuseppe Gallone[1], Petr Danecek[1], Kaitlin E. Samocha[1], Juliet Handsaker[1], Sebastian S. Gerety[1], Holly Ironfield[1], Patrick J. Short[1], Alejandro Sifrim[2], Tarjinder Singh[1], Kate E. Chandler[3], Emma Clement[4], Katherine L. Lachlan[5,6], Katrina Prescott[7], Elisabeth Rosser[4], David R. FitzPatrick[8], Helen V. Firth[1,9], and Matthew E. Hurles[1,a] on behalf of the Deciphering Developmental Disorders study

[1]Wellcome Sanger Institute, Hinxton, Cambridge, United Kingdom

[2]Center of Human Genetics, KU Leuven, Leuven, Belgium

[3]Department of Genetic Medicine, St Mary's Hospital, Central Manchester Foundation Trust, Manchester, United Kingdom

[4]Department of Clinical Genetics, North East Thames Regional Genetics Service, Great Ormond Street Hospital for Children NHS Trust, London, United Kingdom

[5]Wessex Clinical Genetics Service, Southampton University Hospitals NHS Foundation Trust, Princess Anne Hospital, Southampton, UK

[6]Faculty of Medicine, Human Development and Health, University of Southampton, Southampton, UK

[7]Department of Clinical Genetics, Yorkshire Regional Genetics Service, Leeds, United Kingdom

[8]MRC Human Genetics Unit, MRC IGMM, University of Edinburgh, WGH, Edinburgh, United Kingdom

[9]East Anglian Medical Genetics Service, Cambridge University Hospitals NHS Foundation Trust, Cambridge, United Kingdom

[a]To whom correspondence should be addressed: meh@sanger.ac.uk

Abstract

Mobile genetic Elements (MEs) are segments of DNA which, through an RNA intermediate, can generate new copies of themselves and other transcribed sequences through the process of retrotransposition (RT). In humans several disorders have been attributed to RT, but the role of RT in severe developmental disorders (DD) has not yet been explored. As such, we have identified RT-derived events in 9,738 exome sequenced trios with DD-affected probands as part of the Deciphering Developmental Disorders (DDD) study. We have ascertained 9 *de novo* MEs, 4 of which are likely causative of the patient's symptoms (0.04% of probands), as well as 2 *de novo* gene retroduplications. Beyond identifying likely diagnostic RT events, we have estimated genome-wide germline ME mutagenesis and constraint and demonstrated that coding RT events have signatures of purifying selection equivalent to those of truncating mutations. Overall, our analysis represents a comprehensive interrogation of the impact of retrotransposition on protein coding genes and a framework for future evolutionary and disease studies.

41 # Main

42 In humans, three classes of Mobile genetic Elements (MEs) – *Alu*, long interspersed
43 nuclear element 1 (L1), and SINE-VNTR-*Alu* (SVA) – are still active and can generate new
44 copies, known as Mobile Element Insertions (MEIs), throughout their host genome[1]. The L1
45 replicative machinery can also facilitate the duplication of non-ME transcripts, typically
46 protein-coding genes, through the mechanism of retroduplication to generate processed
47 pseudogenes (PPGs)[2]. Combined, these two processes constitute retrotransposition (RT) in
48 the human genome, with new (*de novo*) MEI variants previously estimated to occur in every
49 1 out of 18.4 to 26.0 births[3]. On a population level, each individual human genome harbors
50 ~1,200 polymorphic variants, with the smallest ME, *Alu*, generally contributing 75% of total
51 RT polymorphisms[4-6].

52 To date roughly 130 pathogenic variants caused by RT activity have been
53 documented[7]; however, the majority of these deleterious events have been discovered in
54 isolated cases. Neither MEIs nor PPGs are analyzed as part of routine clinical sequencing
55 and thus represent a largely unassessed category of genetic variation in many disorders.
56 Furthermore, of the clinically relevant RT-attributable cases thus identified, few (~14/123;
57 11.4%) are caused by new mutational events and are instead typically attributable to rare
58 inherited polymorphisms[7]. Additionally, of the large disease-focused whole genome
59 sequencing (WGS) projects which have ascertained MEIs, all have focused on autism[8,9] and
60 have failed to identify likely causative RT-derived variants. In fact, in the largest and most
61 recent WGS study investigating the role of large structural variants in the genetic
62 architecture of autism, the authors failed to identify a single *de novo* MEI in a coding exon,
63 deleterious or otherwise, in 829 families[9]. This finding is likely a result of several factors,
64 predominant among them the low frequency of cases attributable to gene disruption by MEIs
65 in autism[10], due in part to a low ME mutation rate[3] and lack of a sufficiently large sample
66 size[8,9,11]. As such, it is not precisely known at what rate *de novo* ME variants are generated
67 in the human genome, the functional consequences of such variants, the role that they play

68 in the etiology of rare disease, and if routine clinical sequencing should assess patient
69 genomes for deleterious RT events.

70 We analyzed the whole exome sequencing (WES) data produced by the Deciphering
71 Developmental Disorders (DDD) study to systematically assess the role of RT in severe
72 developmental disorders (DDs). The DDD data have already been investigated for
73 pathogenic single nucleotide variants (SNVs), small insertions and deletions (InDels), large
74 copy number variants (CNVs), and other classes of structural variation[12-18]. Approximately
75 24% of DDD cases harbour a pathogenic *de novo* mutation in a gene known to be
76 associated with developmental disorders[12]. The DDD cohort should thus be relatively
77 enriched for highly penetrant *de novo* RT events in comparison to recent studies on
78 autism[8,9]. With a cohort of 9,738 trios (n = 28,132 individuals) whole exome sequenced, the
79 DDD study presents a powerful opportunity to identify, and ascertain the role in DD of,
80 pathogenic *de novo* RT events that impact coding sequences.

81 ## Results

82 ### Generation of a genome-wide dataset of RT variants

83 To assess 9,738 DDD study trios for RT events we utilized two separate
84 computational approaches to identify both MEIs and PPGs. First, we used the Mobile
85 Element Locator Tool (MELT)[5] to identify *Alu*, L1, and SVA variants located within the WES
86 bait regions (Methods). The second is a new bespoke tool developed to identify PPGs from
87 WES data (Methods, Supplemental Fig. 1). Due to cross-hybridisation between a PPG and
88 the exome baits targeting the donor gene, we anticipated that we should be able to detect
89 PPGs genome-wide, not just the subset that insert within the WES bait regions. Our PPG
90 detection tool ascertained putative PPGs by identifying multiple discordant read pairs
91 mapping to different exons of the same transcript, before then typing all individuals for the
92 presence/absence of the PPG using discordant read-pairs and split reads. The tool was
93 optimized by comparing against previously described PPG polymorphisms in the 1000
94 genomes project (1KGP; see below).

95

	Total Sites Cohort-Wide	**Mean Sites Per Unaffected Parent**	**Total de novo Sites**
Alu	917	23.6±4.2	7
LINE-1	167	2.8±1.5	2
SVA	45	0.2±0.5	0
Total - MEI	*1,129*	*26.6±4.7*	*9*
Processed Pseudogenes (PPGs)	576	6.9±2.6	2
Total – MEI + PPG	*1,705*	*33.5±7.3*	*11*

96 *Table 1: RT variant discovery in the DDD*

97 Quantification of the four different classes of retrotransposons discovered as part of this study.
98 Grey-highlighted rows indicate totals across the classes listed above.

99 As our study is the first to discover MEIs directly from WES on a large scale with
100 MELT, we first utilized matched sample WGS data to determine if MELT could ascertain MEI
101 variants reliably from WES data. We compared MEI variants identified by MELT within the
102 DDD WES data to both WGS data generated for 90 overlapping DDD individuals and
103 population MEI data previously generated from the 1000 Genomes Project Phase 3
104 (1KPG)[4-6] WGS data. The latter comparison was to ensure that the number of exonic MEIs
105 identified within DDD WES data was concordant with expectations at the individual and
106 population level. When comparing our WES genotypes to WGS in the same individuals, we
107 had a genotype concordance rate of 94.46% (93.93% *Alu*, 97.29% L1, 98.25% SVA) among
108 calls with at least 10X coverage in our WES data. In total, we were able to re-identify 1,355
109 (1,289 *Alu*, 160 L1, 1 SVA) MEI genotypes, or 84.5% of all heterozygous or homozygous
110 genotypes identifiable with WGS in WES bait regions (Supplemental Table 1). Based on
111 these findings we were confident that MELT was appropriately calibrated to ascertain MEIs
112 in WES data.
113 We identified 1,129 MEI variants and 576 polymorphic PPGs, with each individual's
114 exome containing on average 33.5±7.3 variants (Table 1). All MEIs were genotyped across
115 all individuals to form a comprehensive catalogue of RT-derived variation within and
116 adjacent to (±50 bp) sequences targeted in the WES assay (Methods), including coding
117 exons and targeted non-coding elements (Table 1; Fig. 1). The average time to assess a
118 single family for RT-derived events was approximately 15 minutes and the rate of false
119 findings was low (1 incorrect candidate *de novo* variant per every 295 patients; either a false
120 positive variant [1 per 649] or false negative genotype [1 per 541] in at least one parent). We
121 investigated whether false negative parental genotypes could be mosaic. Using a long-PCR
122 coupled to long-read sequencing assay (see methods), we were able to phase four of our
123 false-negative *Alu* variants to nearby SNVs. With our approach, we could demonstrate two
124 insertions were mosaic (chr3:g.101378585_101378586ins*Alu*(1_280),
125 chr15:g.91145646_91145647ins*Alu*(1_281)), one was part of a complex event of unknown
126 mosaicism (*FCGBP*:c.4743_4744ins*Alu*(1_110)), and another constitutive. Therefore these

127 false apparent *de novo* calls can be generated by a variety of causes, including parental
128 mosaicism (Supplemental Fig. 2).

129 As expected, the total number of variants per individual for each RT class (Fig. 1a-d)
130 as well as combined number of RT events (Fig. 1e) approximated a Poisson distribution (see
131 methods; Supplemental Fig. 3). The vast majority of variants are rare (allele frequency <
132 $1x10^{-4}$; Fig. 1f), with >65% of *Alu* and L1 variants identified in fewer than 4 unrelated
133 individuals. SVA and PPGs appear to be moderately under ascertained compared to *Alu* and
134 L1 at lower AFs, with >50% of variants identified in the lowest AF bin and length estimates
135 for the three MEI classes largely fit the findings of previous studies (Fig. 1g)[4,5].

136 We next sought to ensure that our total number of ascertained RT variants, both on a
137 population and individual basis, accorded with previously published WGS data[4,5]. On a
138 population level, WES did not appreciably limit our overall sensitivity compared to WGS
139 sampled data. When we compared a downsampled version of our call set to the 1KGP, our
140 total number of *Alu* and SVA variants fell within the expected distribution, while L1 was close
141 to expectation (Supplemental Fig. 4).

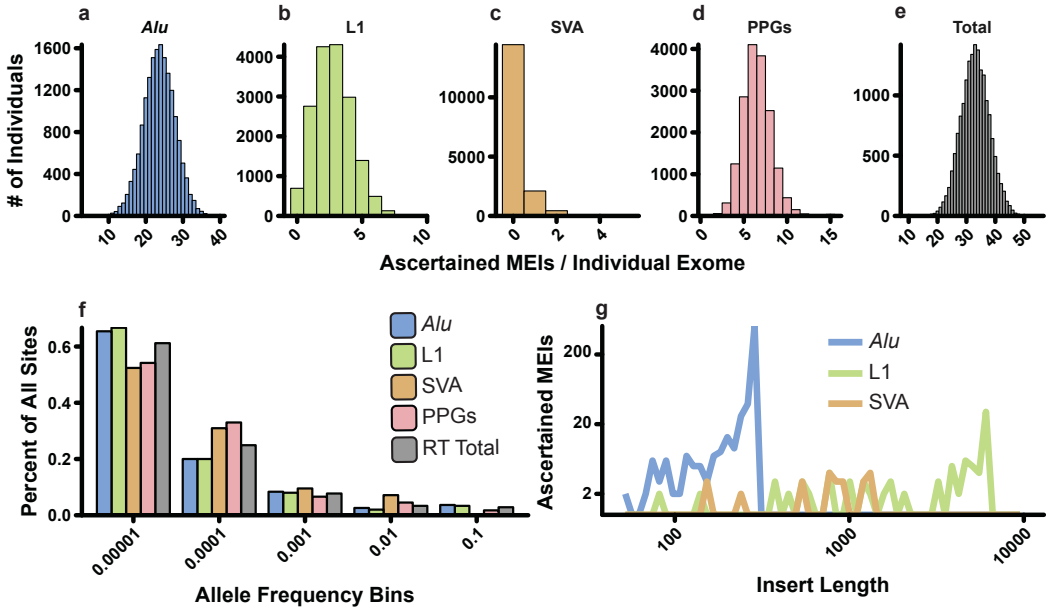

Figure 1: The DDD RT call set

(**a-e**) Histograms of total number of variants per individual for the four classes of RT events identified in the DDD cohort (*Alu* – blue; L1 – green; SVA – orange; PPGs – red; combined RT events – grey) in size one bins. (**f**) Allele frequency distributions for the RT classes depicted in **a-e** in $\log_{10}$ allele frequency bins. (**g**) Insert size estimates provided by MELT for the MEI classes ascertained in this study in $\log_{10}$ insert size bins. All plots only include variants from unaffected parents.

151 To assess the quality of the PPG call set, we compared PPG allele counts (i.e. total
152 number of individuals with a retro-duplication of a given gene) to a recent assessment of
153 PPGs in samples sequenced as part of the 1KGP[6]. Generally, PPGs identified in both data
154 sets shared similar relative allele counts ($r^2 = 0.64$) and variants identified in this study but
155 missing from Zhang et. al.[6] are typically rare (Supplemental Fig. 5). To further validate our
156 approach and ensure that the identified PPG donor genes fit with previously identified
157 patterns of germline PPG formation[2,19], we assessed each donor gene for both functional
158 annotation and expression across 30 tissue types analyzed by the GTEx consortium[20]. The
159 major functional cluster (DAVID[21] enrichment score 8.82) belonged to genes involved in the
160 ribosomal and translational machinery, consistent with previous findings involving fixed
161 PPGs in the human genome[2]. Our expression analysis likewise confirmed previous
162 findings[19], and shows that donor genes that give rise to PPGs are more highly expressed in
163 a large number of tissues compared to non-retroposed genes (Wilcoxon rank sum $p < 1x10^{-3}$
164 for all tissues; Supplemental Fig. 6). Additionally, while it could be assumed that increased
165 germ-line expression of a gene may play a role in increased probability of PPG generation,
166 when we compared PPG donor gene expression in the testis and ovary to that in other
167 tissues, the majority of tissues (20/28, identical tissues for ovary and testis) showed
168 statistically identical patterns of donor gene expression (Wilcoxon rank sum $p > 1x10^{-3}$;
169 Supplemental Fig. 6).
170

171 ## Coding RT burden and constraint

172 As expected for WES, the vast majority (84.9%) of detected MEIs impacted the
173 coding or intronic sequence of a protein-coding gene or a regulatory element targeted in the
174 augmented WES assay described in Short et. al.[15] (Fig. 2a). Due to the large numbers of
175 individuals with WES data in this study, we have ascertained over five-fold more exonic
176 variants than the largest previously published study (Supplemental Fig. 7)[4,5]. Nonetheless,
177 the number of MEIs identified in this study, based on the proportion of the genome assayed,
178 likely represent only 2.2% of MEI variants genome-wide in these individuals.

179 Our large collection of coding variants allowed us to examine the evolutionary forces
180 acting on coding MEI variation (Fig. 2b). To examine selective constraint, we utilized two
181 common measures: the proportion of variants observed in only one individual (e.g.
182 singletons)[22] and the proportion of variants found in genes likely to be intolerant of loss of
183 function (LoF) as determined by the pLI score[23]. To avoid issues of relatedness and the
184 potential for clinical ascertainment bias for pathogenic MEIs in individual DD patients, only
185 the 17,032 unaffected parents sequenced as part of DDD were included in our analysis.
186 MEIs which directly impact exons are under strong selective constraint, indistinguishable
187 from that of both nonsense and essential splice site SNVs (Fig. 2b). Interestingly, we did not
188 find any sign of selection acting on intronic MEIs as they appear to be constrained similarly
189 to synonymous SNVs. In contrast to previous studies[24,25], we did not find a statistically
190 significant (χ^2 $p < 0.05$) bias towards intronic MEIs inserted in the antisense orientation of
191 the gene in which they are found (Supplemental Fig. 8). This is likely not a repudiation of
192 such work, but attributable to the relatively small number of intronic events we identified as
193 part of our analysis compared to WGS[4,24] or reference genome-based[25] studies. To put our
194 findings on exonic MEI constraint into perspective with other forms of variation, every human
195 genome will harbor approximately one (0.76±0.62 per individual) MEI which directly impacts
196 protein-coding sequence. Since MEIs are similar to nonsense SNVs in terms of
197 deleteriousness (Fig. 2), MEIs thus make up 0.6%[22,26] of all coding protein truncating
198 variants (PTVs; among SNVs, InDels, and large CNVs) in each individual human genome.

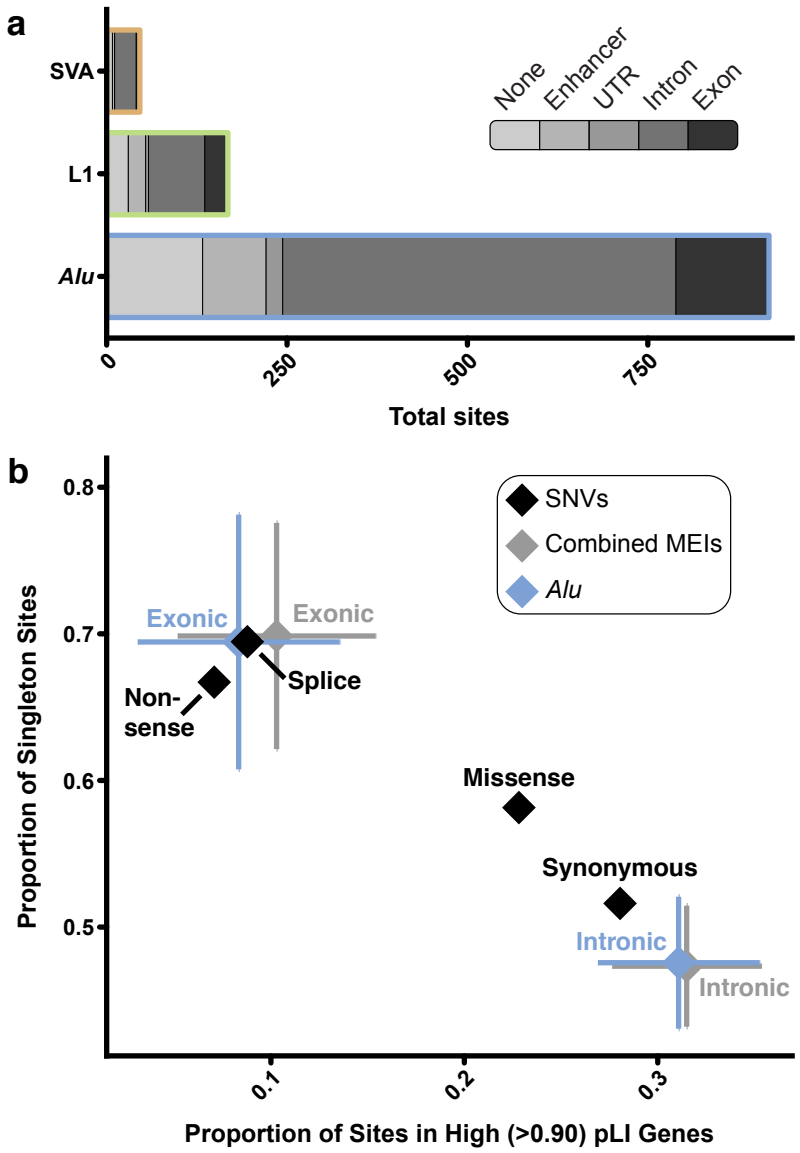

Figure 2: Coding constraint on MEIs

(**a**) Cumulative consequence annotations for *Alu*, L1, and SVA MEIs. The majority of variants identified in this study fell within the non-coding space (either an enhancer or intron) (**b**) Comparison of constraint between MEIs and SNVs in unaffected parents. To compare the impact of exonic and intronic *Alu* (blue) and all MEIs (grey) to varying classes of SNVs (black), we used two metrics: the proportion of variants in genes that have been identified as LoF intolerant as gauged by pLI-score[22] (x-axis) and the proportion of variants identified in only one individual (i.e. singletons; y-axis). Error bars indicate 95% confidence intervals based on population proportion; confidence intervals were calculated for SNVs, but are too small to appear at the resolution displayed in this figure.

210 While we were unable to perform similar population genetic analyses for PPG
211 events, due to the difficulty of resolving the putative insertion site with WES data and thus
212 distinguishing between different PPGs for the same donor gene, we were able to assess the
213 propensity for specific genes to give rise to PPGs based on their selective constraint. We
214 observed that PPG donor genes were significantly enriched for genes that are highly
215 intolerant of loss of function variation (pLI > 0.9). High pLI genes make up 25.3% of donor
216 genes, compared to 17.6% of all protein-coding genes (χ^2 p = 2.4×10^{-6})[22]. This observation is
217 likely driven by loss of function intolerant genes being more likely to be highly expressed in
218 multiple tissues[22], similar to genes known to have been retroduplicated (Supplemental Fig.
219 6)[19]. This observation implies that PPG events rarely strongly perturb the function of their
220 donor gene – despite several previously documented instances of PPGs impacting
221 expression or functionality of their donor gene[27].
222

223 ## Discovery and clinical annotation of *de novo* RT variants in DD

224 Using the computational approaches outlined above we identified a total of 11 germ-
225 line *de novo* RT variants (Table 2). Our findings include coding, noncoding, pathogenic and
226 benign variants, as well as, to our knowledge, the first *de novo* MEI identified in a pair of
227 monozygotic twins (Supplemental Fig. 9). All *de novo* RT variants were confirmed via a PCR
228 assay specific to the RT class (Fig. 3; Supplemental Fig. 9; Supplemental Table 2) and,
229 where possible, inspected for poly(A) tail and target site duplication – hallmarks of *bona fide*
230 RT activity[28]. We identified no *de novo* RT variants which localized to the non-coding
231 elements included on the WES capture, which falls in line with expectations based on
232 mutation rate estimates (Fig. 4b). We also attempted to determine the parental origin of each
233 RT event using SNVs located on sequencing reads which support the RT insertion (Table 2).
234 Of the 11 *de novo* RT events, we were able to phase three variants, all to the father. While
235 this finding is not statistically significant (χ^2 p = 0.083), it fits with previous findings that the

236 majority of *de novo* structural variants[9], and indeed most variant classes[29], are attributable to
237 paternal origin.

238 Nine of our validated *de novo* mutations were MEIs (7 *Alu*, 2 L1), or a rate of
239 approximately one *de novo* event per every 1,000 patient exomes sequenced (9/9,738). As
240 expected, based on both the total number of polymorphisms[3-5] and mutation rate (Table 1;
241 Supplemental Table 4), we identified more *Alu de novo* variants than the other RT classes.
242 We also identified 2 PPG germ-line *de novo* variants, or approximately one new PPG per
243 every 5,000 patient whole genomes sequenced (2/9,738). As a further quality control for
244 PPGs, we capillary sequenced all resulting PCR products to confirm the gene of origin
245 (Supplemental Data 1) and performed WGS to identify the PPG insertion site. We were able
246 to localize the *SERINC5* PPG to an ~50Kbp intron of the gene *CLIC4* and the *SLC35F2*
247 event to an intergenic region between the genes *MAK* and *GCM2* (Fig. 3e). Neither of the
248 events directly impacted coding sequence and *CLIC4* is neither under strong selective
249 constraint nor known to have any link with DD.

Insertion Coord.	RT Type	Genomic Compartment	ENSEMBL Gene ID	HGNC Gene ID	pLI	DDG2P Annotation	Decipher ID[30]	Diagnostic?	Parental Origin	Notes
chr3:9495459	*Alu*	Exonic	ENSG00000168137	*SETD5*	1.000	confirmed,monoallelic	280818	True	Father	
chr5:176638159	*Alu*	Exonic	ENSG00000165671	*NSD1*	1.000	confirmed,monoallelic	259118	True	Unknown	Included in Wright et. al.[31]
chr6:159190834	*Alu*	Exonic	ENSG00000092820	*EZR*	0.988	None	300984	False	Unknown	
chr7:77552086	*Alu*	Exonic	ENSG00000006576	*PHTF2*	0.024	None	271388	False	Father	
chr3:135913800	*Alu*	Intronic	ENSG00000174579	*MSL2*	0.890	None	292325	False	Unknown	
chr3:148614204	*Alu*	Intronic	ENSG00000163751	*CPA3*	<0.001	None	270426; 270428	False	Unknown	Monozygotic twins
chr3:172480619	*Alu*	Intronic	ENSG00000114346	*ECT2*	<0.001	None	307591	False	Unknown	
chr12:46246325	L1	Exonic	ENSG00000189079	*ARID2*	1.000	probable,monoallelic	264759	True	Unknown	
chr5:88100580	L1	Exonic	ENSG00000081189	*MEF2C*	0.004	confirmed,monoallelic	285645	True	Unknown	
chr6:10847968	Retrogene-SLC35F2	Intergenic	N/A	N/A	N/A	N/A	291670	False	Unknown	
chr1:25074202	Retrogene-SERINC5	Intronic	ENSG00000169504	*CLIC4*	0.009	None	301168	False	Father	

Table 2: Confirmed germ-line de novo variants in the DDD study

Relevant clinical and annotation information for MEI and PPG de *novo* variants identified as part of this study. Location of the insertion event is given in hg19 reference coordinates (Insertion Coord.). A “True” value in the “Diagnostic” column indicates, at the time of publication, that this variant intersected a known DD gene and was deemed likely to be involved in the patient’s phenotype by the referring clinician. “False” does not indicate whether or not, with additional future evidence, the gene may become associated with DD and the variant thus deemed diagnostically relevant. If applicable, ENSEMBL[32] gene IDs indicate the gene impacted, not the gene from which the event is derived (i.e. for PPGs).

256 Each *de novo* mutation was then compared to known DD-associated genes (using
257 the Developmental Disorders Genotype-to-Phenotype database – DDG2P) to identify
258 potentially pathogenic variants (Table 2). Of the mutations identified, four directly inserted
259 into coding exons of DD-associated and fetal brain-expressed (see methods; Supplemental
260 Table 5)[33] genes (Fig. 3, Table 2) with all four found in genes statistically enriched for PTVs[12]
261 and therefore likely to operate by a LoF mechanism. We did not identify any intronic *de novo*
262 mutations likely to be pathogenic (Fig. 3a-d; Supplemental Fig. 9). An additional mutation
263 inserted into the coding sequence of a strongly LoF-intolerant gene, *EZR* (pLI = 0.99;
264 Supplemental Fig. 9), but we could not directly attribute it to the patient's phenotype due to
265 lack of significant enrichment for PTVs, although there is prior evidence for a role in a
266 familial DD syndrome[34]. The four mutations in DD-associated genes were reported to the
267 referring clinician for clinical interpretation based on both initially reported and updated
268 phenotypes (Supplemental Table 3). Three out of four reported mutations (*NSD1*, *MEF2C*,
269 *ARID2*) were subsequently deemed to be likely causative of the patient's phenotype
270 (Supplemental Table 3) by the referring clinician. To determine if our patient with a likely
271 pathogenic *NSD1* variant phenocopies the already described aberrant blood DNA
272 methylation signature typical of Sotos syndrome patients[35], we compared our patient's
273 genome-wide methylation status to several already published Sotos syndrome cases (see
274 methods)[35]. Our patient strongly clusters with other confirmed Sotos syndrome patients and
275 confirms our clinical assessment and diagnosis of this particular patient (Supplemental Fig.
276 10). The fourth patient, with an *Alu* insertion in *SETD5* (Fig. 3a), has clinical features
277 (polydactyly and truncal obesity; Supplemental Table 3) more suggestive of a ciliopathy. As
278 such, the identified MEI is unlikely to be the sole cause for the patient's DD but may
279 contribute to a composite phenotype. Accordingly, we did investigate this patient's exome for
280 other potentially diagnostic variants, but did not identify additional *de novo* or inherited
281 variation that could plausibly be associated with this patient's phenotype.

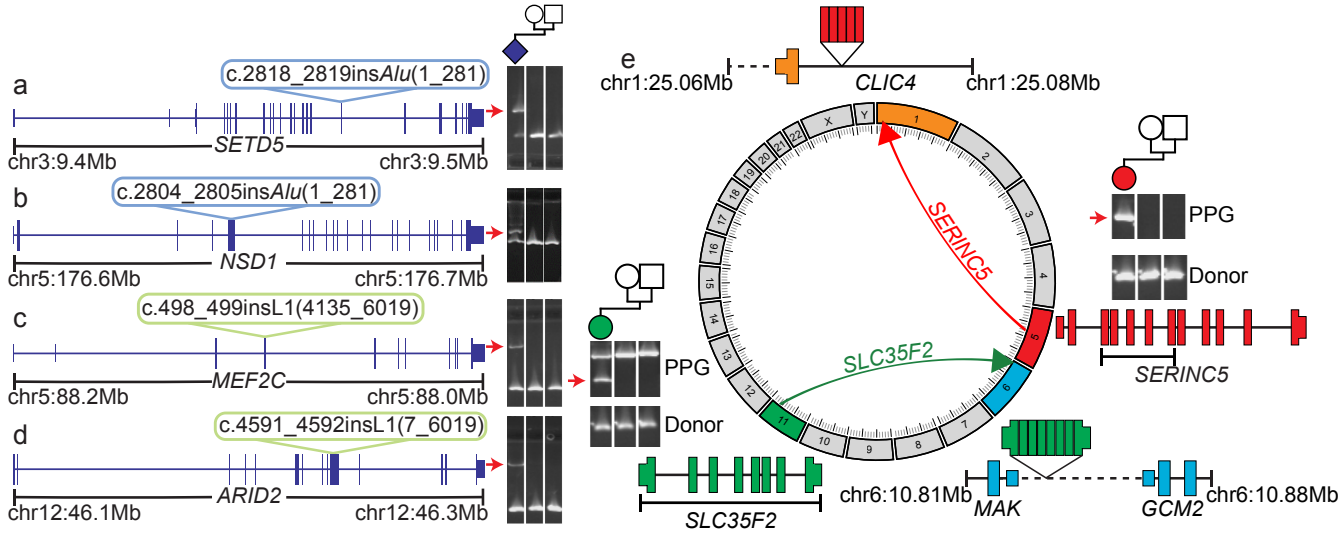

Figure 3: RT-derived de novos in the DDD

We identified a total of nine *de novo* MEIs, four of which disrupted the protein-coding sequence of a known DD gene: (**a**) *SETD5*, (**b**) *MEF2C*, (**c**) *ARID2*, and (**d**) *NSD1*. Shown in each panel is a diagram of the affected gene (blue model) with the relevant insertion indicated with a colored bubble. To the right are PCR validations confirming the *de novo* status of each mutation; a positive result is indicated by a raised secondary band present only in the proband sample (red arrow). (**e**) Circos diagram and PCR results for two identified germ-line *de novo* PPGs. For each *de novo* PPG shown is a diagram of the donor gene (gene model), location of duplication as PPG (directional arrow), and new insertion site. Exons from the donor gene included in the PPG are indicated by brackets underneath the donor gene model. To confirm PPG presence, PCR was performed (Methods) on proband, paternal, and maternal gDNA (sample in each lane is shown by pedigree). The band which represents the PPG is marked with a red arrow and was confirmed via capillary sequencing (Supplemental Data 1). Dashed lines indicate intergenic regions, all genes models are shown in sense orientation, and PPG gene diagrams are not to scale.

297 We also examined our dataset for inherited rare pathogenic RT variants. We
298 evaluated variants inherited from an affected parent, bi-allelic inheritance (either a
299 homozygous MEI or a heterozygous MEI paired with another variant class), and X-linked
300 variants maternally inherited by affected males. We did not identify any rare MEI variants
301 inherited from an affected parent nor any compound heterozygous individuals with a rare
302 MEI and a non-MEI PTV (e.g. SNV/InDel) impacting the same gene. We did identify a single
303 proband-specific homozygous MEI inserted into an exon of *PAN2* which was unique to a
304 single family. This gene was recently identified as nominally significant (genome-wide p =
305 4.2×10^{-4}) in a study investigating the role of recessive variants in DD[13], although more data
306 are required to be confident of its association to DD. We also identified a total of 22 (14 *Alu*,
307 7 L1, 1 SVA) polymorphic MEIs on the X chromosome, of which 4 (3 *Alu*, 1 L1) directly
308 impacted protein-coding sequence. Of these variants, none were at a low enough allele
309 frequency to be reasonably DD-associated, were located within a gene associated with DD,
310 nor fit an inheritance pattern consistent with X-linked disease.
311

312 ## MEI mutation rate and enrichment of deleterious RT events in DDD

313 Based on our findings, in the coding and peri-coding portion of the genome, one out
314 of every 2,434 DD cases (0.04%±0.04; 95% CI) is directly attributable to RT-derived
315 mutagenesis. To determine both if our observed number of *de novo* variants meets
316 expectation and if our patient cohort is enriched for causal *de novo* RT events, we estimated
317 the population mutation parameter, Θ[36], from the unaffected parents in the DDD study and
318 from the 1KGP[4,5] (Supplemental Table 4). The resulting calculation gives very similar
319 estimates of MEI mutation rate (combined across *Alu*, L1, SVA) of between 1.4×10^{-11} (1KGP)
320 and 1.2×10^{-11} (DDD) variants per bp per generation (μ), or ~1 new MEI genome-wide per
321 every 12 to 14 births – largely concordant with prior estimates from smaller WGS
322 datasets[3,37].
323

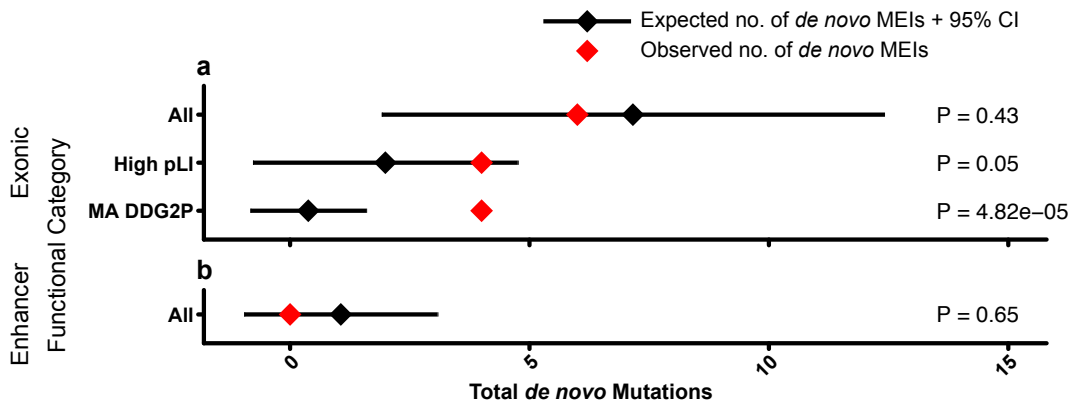

Figure 4: Estimating enrichment of deleterious MEIs

Depicted are total number of expected (black) and observed (red) *de novo* mutations observed in exons (**a**) and enhancers (**b)** for all, high pLI (pLI > 0.9), and known monoallelic DD (MA DDG2P) genes. Expectation is based on the Poisson distribution of 100 simulations utilizing the neutral mutation rate (1.2×10^{-11} μ). P-values are based on the Poisson distribution, and used to determine statistical deviation of observed to expected *de novo* counts for exons and enhancers (right).

Using this genome-wide mutation rate, we estimated the number of expected mutations in various genomic compartments, including within genes intolerant to PTVs and within DD-associated genes (Fig. 4; Methods). We identified a significant enrichment of *de novo* MEIs in dominant DD-associated genes ($p = 4.82 \times 10^{-5}$), but not in the much larger set of LoF intolerant genes ($p = 0.05$). To ensure that this finding was not due to inaccurate estimation of the genome-wide mutation rate, we also assessed the probability that four out of six exonic *de novo* MEIs would fall within exons of dominant DD-associated genes by chance, based on the proportion of the exome represented by these genes (and assuming known DD-associated genes have the same MEI mutation rate as other genes) and likewise found a significant enrichment ($p = 4.3 \times 10^{-5}$).

342 # Discussion

343 Here we have described the development, validation and exemplification at scale of
344 an analytical pipeline for the rapid assessment of patient genomes for RT variants.
345 Alongside Torene et. al.[38], we have used these approaches to present the largest study
346 examining the coding genome for RT-derived variation to date (Table 1; Fig. 1). With this
347 dataset, we first demonstrated that exonic MEIs (regardless of insertion length) are under
348 selective constraint on par with protein-truncating SNVs (Fig. 2). We identified four likely
349 pathogenic RT mutations, two *Alu* and two L1 insertions (Fig. 3), all of which arose *de novo*
350 in known haploinsufficient DD-associated genes (Fig. 3a-d), implying that dominant loss-of-
351 function is the major mode of pathogenic exonic RT variation. Finally, we estimated the
352 genome-wide MEI mutation rate and used it to determine that DDD probands are enriched
353 for damaging RT variation within exons of dominant DD-associated genes (Fig. 4a).

354 The total number of polymorphic, exonic RT variants identified in DDD is concordant
355 with previous studies characterizing MEI variation[3,5,39]. Pathogenic MEIs make up 0.04% of
356 diagnoses in the DDD study (4/9,738 probands), a small yet individually significant collection
357 of diagnostic variants. Reassuringly, our proportion of diagnostic variants in DDD is
358 statistically identical to the 7/11,011 (0.06%) diagnostic rate for neurodevelopmental disorder
359 patients as determined by Torene et. al.[38] (Fisher's exact test $p = 0.56$). Unlike Torene et.
360 al.[38], we did not identify a causative inherited MEI, although this difference is not statistically
361 significant (Fisher's exact test $p = 0.51$). We infer that despite making up a significant
362 proportion of reported MEI variants in the clinical literature[7], bi-allelic or X-linked MEI events
363 are a less frequent class of pathogenic variant in developmental disorders. This is in keeping
364 with recent estimates[13] that in a largely outbred clinical population, such as in the UK,
365 recessive disorders caused by coding variants account for a much smaller fraction of
366 patients than dominant disorders. We also attempted to investigate both the parental origin
367 of *de novo* variants (Table 2) and the mosaicism of MEIs with low variant allele frequency
368 (Supplemental Fig. 2). In both cases, additional work incorporating long-read technology is

369 needed to ascertain parental contribution and, in the case of parental mosaic MEIs, potential
370 recurrence risk.
371 Interestingly, it appears that the contribution of diagnostic RT variants may vary
372 among diseases. Wimmer et. al.[40] reported a total of 13 diagnostic, exonic MEI variants in
373 4,500 neurofibromatosis type I patients (0.3% of patients). This rate is seven times higher
374 than that observed in DDD or Torene et. al.[38] and was attributed to a potential RT mutation
375 "hotspot" associated with the canonical L1 endonuclease cleavage site of 3'-AA/TTTT-5'[41]
376 within the neurofibromatosis-associated gene, *NF1*. A recent study using a cell culture
377 retrotransposition assay to study MEI site preference was unable to identify a mutation
378 hotspot in *NF1*[42] and therefore suggests further work is needed to investigate the role of
379 sequence context in determining the overall genomic landscape of RT-mediated disease.
380 Analogously, inclusion of sequence context into the SNV mutation model noticeably
381 improved the ability to determine enrichment/depletion of deleterious SNVs within genes[22,23].
382 Our study is limited in that we only identified ~2% of the RT variants in each
383 individual human genome[4,5]. Despite a number of known disease-associated intronic MEIs in
384 the literature, we did not identify a pathogenic intronic MEI. As such, it remains an open
385 question as to what contribution RT mutations in the noncoding genome plays in the etiology
386 of DD. While it appears that the contribution of regulatory elements to DD is relatively small,
387 as defined by this (Fig. 4b) and other studies[15], previous work has identified a significant
388 signature of purifying selection against MEI events within 100bps of exons[25] – variants which
389 our study could potentially identify. As our data suggests that the majority of DD cases with
390 pathogenic coding MEIs are due to *de novo* insertions (Table 2; Fig. 3), we conjecture that
391 most additional DD-associated MEIs may be located in the introns of known DD-causing
392 genes and disrupt splicing – a known disease mechanism attributable to RT-derived
393 mutagenesis[7,40,43]. Simulations suggest that under a null genome-wide mutation model we
394 should expect to observe 12.5 (5.5-19.4, 95% CI) *de novo* intronic RT mutations in dominant
395 DD-associated genes in a population sample of 9,738 individuals. As such, a WGS study of

396 a clinical population of similar size to that analyzed here should be well powered to estimate
397 the pathogenic contribution of intronic MEIs.
398 *De novo* MEIs are typically readily interpretable with modest informatics expertise,
399 and represent a clinically relevant class of variation to assay in clinical bioinformatics
400 pipelines. While we ultimately find that the overall burden of RT-attributable disease is
401 relatively low in the human population, it is nonetheless an important consideration when
402 elucidating the genetic basis of disease in individual patients.

403 # Online Methods

404 ## Patient recruitment and sequencing

405 A total of 13,462 patients were recruited from 24 clinical genetics centers from
406 throughout the United Kingdom and the Republic of Ireland as previously described[44].
407 Informed consent was obtained for all families and the study was approved by the UK
408 Research Ethics Committee (10/H0305/83, granted by the Cambridge South Research
409 Ethics Committee and GEN/284/12, granted by the Republic of Ireland Research Ethics
410 Committee). For the purposes of this study, individuals that were not recruited as part of a
411 trio (e.g. individual patients or patients with just one parent), were included on the DDD
412 sample blacklist, or failed to meet MELT QC requirements[5] were excluded from downstream
413 analysis (leaving n = 9,738 probands; 28,132 individuals). Sequencing of families and
414 alignment to version hg19 of the human reference genome with bwa[45] was performed as
415 previously described[12].
416

417 ## Processed pseudogene pipeline development

418 PPGs, particularly young polymorphic events, share highly homologous sequence
419 with the source gene from which they are derived. Consequently, the WES bait capture
420 method will capture both DNA from the original “donor” gene and the new “daughter” copy.
421 This allows, compared with our approach for MEI discovery, for ascertainment of PPGs
422 genome-wide. While this approach does come with limitations, such as difficulty in
423 identifying insertion variants, we can still determine events per individual.
424 Our discovery pipeline functions in two steps: first we collect read evidence on an
425 individual level to determine which genes have been retroduplicated in that individual
426 (Supplemental Fig. 1). Second, we determine presence/absence of each PPG in every
427 individual in the DDD cohort based on the gene models built in the first step. In step one, we
428 iterate over all genes in the ENSEMBL gene database which have a determined pLI score[22]
429 and collect discordant read pairs (DRPs) which map between exons and have an insert size

430 >99.5% of all other reads in the sample. If more than four reads linking two exons are found,
431 the gene is considered to be retroduplicated elsewhere in the genome. In step two, for each
432 gene identified in step one, all evidence across all PPG positive individuals are pooled to
433 make a model of the PPG. This model is then used to check for DRP and split read pair
434 (SRP) evidence in all genomes. If an individual has at least 5 total read pairs of supporting
435 evidence with at least one SRP and one DRP, an individual is considered positive for the
436 given PPG. All genes and individuals were combined into a flat file listing presence or
437 absence of a given PPG in each individual. Source code and more information is available
438 online at github: https://github.com/eugenegardner/Retrogene.git.
439

440 ## MEI call set generation and consequence annotation

441 To identify MEIs in the DDD WES data we utilized the previously published Mobile
442 Element Locator Tool (MELT)[5]. MELT was run with default parameters (except the '-exome'
443 flag during IndivAnalysis) using 'Split' mode to generate a final unified VCF-format file[46] of all
444 28,132 unfiltered individuals independently for each MEI type (*Alu*, L1, SVA). Following initial
445 data set generation, we found that a subset of variants internal or adjacent to (±50bp) low
446 complexity repeats (defined here as a run of sequence ≥ 15bp composed of two or fewer
447 nucleotides) were likely false positive. As such, we added an additional filter to the final
448 MELT VCF, lc (low complexity), which removes such false positives from downstream
449 analysis. Variants that could not be genotyped in at least 25% of individuals, had ≤ 2 split
450 reads, had MELT ASSESS score < 3, or had any value in the VCF FILTER column other
451 than PASS or rSD were filtered.
452 To determine if variants per individual approximated a Poisson distribution, we
453 calculated the mean number of sites per unrelated individual (λ) independently for each RT
454 class and for all RT classes combined as displayed in Fig. 1a-e. Our calculated λ was then
455 supplied to the function *rpois* in R to generate a random deviate and plotted alongside
456 histograms identical to Figs 1a-e in Supplemental Fig. 3.

457 To generate consequences plotted in Fig. 2a, all MEIs were annotated using Variant
458 Effect Predictor v88 (VEP)[47] and intersected with bedtools intersect[48] to enhancers (one of
459 heart[49], VISTA[50], or highly evolutionarily conserved[51]) included on the DDD WES capture[15].
460 Only a single consequence was retained for each variant, with priority given to enhancer
461 annotation. Primary transcript as determined by VEP was used for all gene-based
462 consequences, pLI score[22] annotation, and DDG2P disease association (Table 2).
463

464 ## Assessment of Potentially Mosaic MEIs

465 To determine the mosaicmism of variants that were identified as "false-negative" *de*
466 *novo* during initial MELT analysis (i.e. genotyped as heterozygous in a single proband and
467 homozygous reference in both parents), we randomly selected 1,000 trios and determined
468 the total number of reads supporting both the insertion and reference allele (i.e. allelic
469 proportion) at all 917 *Alu* loci identified in this study. Alongside this data, we plotted the
470 parental allelic proportion at the 15 "false negative" *Alu* loci described above (Supplemental
471 Fig. 2a). To formalize this result and ensure that these specific loci do not have higher
472 genotyping or sequencing error than other loci we performed three analyses. First, we
473 compared the combined allelic proportion of all heterozygous variants to the combined allelic
474 proportion of both false negative parents and the associated heterozygous probands
475 (Supplemental Fig. 2b). Second, we examined the allelic proportion for all homozygous
476 reference genotypes at each false negative locus for evidence of higher error rates based on
477 higher average allelic proportion at those loci using a z-test (Supplemental Fig. 2c). Finally,
478 to assess if these loci had similar read start distributions in both parent and proband, we
479 retrived all read start sites within 250 bps of each insert site and used a Kolmogorov-
480 Smirnov test to compare the two distributions. Only one locus,
481 chr11:g.95523918_95523919ins*Alu*(1_280), showed a significant difference at a threshold of
482 $p = 0.05$ ($p = 2.70 \times 10^{-4}$).
483

484 Since these sites are not likely to be genotyping or sequencing artifact, we next
485 attempted to clarify the nature of a subset of variants with a custom long-range PCR and
486 Pacific Biosciences sequencing approach. The goal of this methodology was to attempt to
487 both validate and phase candidate mosaic MEIs to nearby (±5Kbps) SNVs. As above, we
488 limited our analysis to *Alu* MEIs as the large insert length of L1 (n = 1) and repeat heavy
489 SVA (n = 2) elements would likely prove refractory to long PCR with enough flanking
490 sequence to phase to nearby SNVs. To design locus specific primers, the genomic region
491 surrounding each target MEI was entered into Primer-BLAST
492 (www.ncbi.nlm.nih.gov/tools/primer-blast/) using default parameters. Each locus was
493 targeted with 3 primer pairs: 1 pair amplifying a 10kb region with the MEI lying centrally, and
494 two pairs of primers to amplify 5kb fragments extending up and downstream of the MEI.
495 10kb PCRs were performed using AccuTaq LA DNA Polymerase (D8045, Sigma-Aldrich)
496 with the following parameters: 98 degree denaturation for 1 minute, followed by 35 cycles of
497 94 degrees (30 seconds)/54 degrees (30 seconds)/68 degrees (11 minutes), with a final
498 extension of 18 minutes at 68 degrees. 5kb PCRs were performed as above, but with a
499 reduced 6 minute extension time at 68 degrees, and a final extension time of 10 minutes.
500 After quality checking the PCRs on control DNA, 100-200ng of parental DNA was amplified
501 in a 50 microlitre PCR reaction using the same parameters.
502 After confirming amplification by agarose gel, PCR products were isolated using
503 SPRI bead purification (Ampure-XP, Agilent), quantified, and mixed in equimolar amounts.
504 We then performed a standard PacBio library prep using SMRTbell Template Prep Kit 1.0
505 (part no. 100-991-900) without size selection. Sequencing was performed on the Pacific
506 Biosciences Sequel platform using Sequel Sequencing Kit 3.0 (101-597-800) and Sequel
507 SMRT cell 1M v3 (101-531-000). For error correction, raw sequencing reads were then
508 provided as input to pbccs v3.4.1 (https://github.com/PacificBiosciences/unanimity) with
509 default parameters other than minLength=100. Corrected reads were then aligned to the
510 human genome using blasr v5.3.3[52] and downsampled to between 2,000-3,000x coverage
511 per locus assessed.

512 MEIs were then phased to the nearest heterozygous SNV on a per-read basis using
513 a custom java pipeline. After dropping sites due to PCR errors or phasing ambiguity, we
514 were able to examine a total of 4 *Alu* MEIs (Supplemental Table 6). Total number of reads
515 supporting each of four possible phases (MEI+SNV+, MEI+SNV-, MEI-SNV+, MEI-SNV-)
516 were then quantified. As PCR can induce template switching errors resulting in variants
517 being sequenced on the same read despite in truth not being on the same haplotype (e.g.
518 MEI+SNV-)[53], we then determined if total number of reads supporting both possible true
519 phases (MEI+SNV+ and MEI-SNV+) were significantly different from error (MEI+SNV-) using
520 a Fisher's exact test. MEIs where both true phases were significantly different from error
521 were considered true mosaic. Primers and quantification of results are available in
522 Supplemental Table 6.
523

524 ## Quality control of RT data using WGS and 1KGP

525 To determine if our MEI WES call set was biased compared to WGS data, we
526 performed two independent comparisons: 1) to high coverage (>30x) WGS data generated
527 for a subset of DDD trios and 2) to a published collection of MEIs from 1KGP phase III[5].
528 For WGS quality-control, we used a subset of 30 DDD trios (n = 90 individuals) which
529 were previously whole genome sequenced. MEI discovery using MELT[5] on all 90 individuals
530 was performed and filtered identically to WES data. Genotypes identified in the WGS data
531 but not in WES were then separated based on coverage in the corresponding WES.
532 Genotypes in low coverage areas (<10x) were considered not possible (n.p), while variants
533 where coverage was greater than 10x are considered not detected (n.d). All remaining
534 genotypes were than compared for identity between WGS and WES results (Supplemental
535 Table 1).
536 To compare the DDD MEI call set to the 1KGP, we first filtered 1KGP calls to
537 variants with mean >10x coverage in 1,000 randomly sampled DDD WES individuals
538 (leaving 318 *Alu*, 81 L1, 26 SVA). We then randomly selected 2,453 DDD parents 1,000

539 times, retaining only loci present in downsampled individuals[4,5]. The resulting distribution
540 was then compared to the observed number of variants in the 1KGP-masked data to
541 generate z-scores independently for all three MEI types (Supplemental Fig. 4).
542 To compare our PPG dataset to Zhang et. al.[6], we downloaded provided
543 supplemental tables. We then summed the total number of unique events per person and
544 determined "allele counts" for each gene reported. Genes were then matched between our
545 call set and Zhang et. al.[6] using ENSEMBL gene identifiers and allele counts between each
546 data set were plotted to create Supplemental Fig. 5.
547

548 ## GTEx annotation of processed pseudogenes

549 To determine RNA expression levels of donor genes which gave rise to PPGs
550 identified in this study, we queried transcript per kilobase per megabase of sequencing
551 (TPM) scores for all genes in 30 tissues assessed by the current GTEx v7 release (available
552 at https://gtexportal.org/home/datasets). Only the 18,225 protein-coding genes which were
553 assessed for gene PPGs by our project were retained for subsequent analysis. TPM values
554 were then averaged across all GTEx individuals for a given tissue to generate a mean TPM
555 value as plotted in Supplemental Fig. 6. Nonparametric Wilcoxon rank-sum tests were
556 performed using the wilcox.test function in R with default parameters to generate p values
557 for both within tissue and between tissue comparisons.
558

559 ## SNV calling and quality control

560 To call SNVs from all DDD individuals we utilized GATK 3.5-0-ge91472d[54] in three
561 steps using default settings. First, we called variants in individual samples using
562 HaplotypeCaller. Next, individual VCF files were processed in 200 individual batches using
563 CombineGVCFs. Finally, all batched VCFs were passed to GenotypeGVCF to generate a
564 final joint-called VCF file. This file was then annotated using VEP v88[47] based on canonical

565 transcript. Unaffected parents (n = 17,032 individuals) were then extracted from this VCF
566 and only variants with an allele count greater than 1 in these individuals were retained.
567 For initial filtering, we removed SNVs with a VQSLOD < -2.0, read depth < 8, and
568 genotype quality < 20. We next performed more extensive QC using a 'missingness' score
569 identical to the method described in Martin et. al.[13]. In short, each genotype at a given
570 variant was assessed for genotype quality (GQ), depth (DP), and a binomial test for allelic
571 depth (i.e. number of alternate versus reference supporting reads; AD). If a given genotype
572 had GQ < 20, DP < 8, or AD p-value < 0.001 it was considered 'missing'. If more than 50%
573 of genotypes for a given variant were missing, the variant was subsequently filtered from
574 final analysis. Allele frequencies were recalculated based on included individuals while
575 accounting for missing genotypes.
576

577 ## Selective constraint of SNVs and MEIs

578 As sensitivity of variant discovery can bias our results, we generated an “accessibility
579 mask” of the DDD WES data where we expect our variant ascertainment sensitivity to be
580 >95% (Supplemental Fig. 11)[5]. Our mask thus includes only regions of the genome that
581 contain at least 10X average coverage in a mean cohort of 1,000 randomly selected
582 individuals for a total of 74.2Mbp, or ~2.3% of the genome (Supplemental Table 4). Using
583 this mask, we filtered our original 1,129 variants down to 828 (660 *Alu*, 109 L1, 31 SVA)
584 variants in unaffected parents (n = 17,032 individuals). Parents were determined to be
585 affected either by the referring clinician or, where ambiguous, through manual curation of
586 HPO terms for a matching parent-offspring phenotype.
587 Using this mask subset of variants, we determined genomic constraint as shown in
588 Figure 2b. Allele frequency values were recalculated for all variants, and a pLI score[22] for
589 each MEI was added as described above. MEIs which did not insert into a gene or inserted
590 into a gene without a calculated pLI score[22] were excluded from subsequent analysis. We
591 then calculated proportion of singleton variants and proportion of variants in high pLI genes

592 independently for *Alu* and, due to low overall numbers of the other MEI subtypes, for a
593 combined set of *Alu*, L1, and SVA. SNVs annotated as nonsense, missense, synonymous,
594 or splice acceptor/donor (splice in Fig. 2b) as determined by VEP v88[47] were extracted from
595 the SNV VCF files described above and used to calculate singleton and pLI proportion
596 identically to MEIs.
597

598 ## Fetal Brain Expression Data

599 To assess if genes disrupted by RT variants were expressed in the fetal brain, we
600 downloaded per-gene expression data generated by the BrainSpan consortium (file:
601 genes_matrix_csv.zip; http://www.brainspan.org/static/download.html)[33]. From this file, we
602 extracted columns representing data generated from fetuses 15 to 37 weeks post-
603 conception. This data was then provided as input to a custom python script which calculated
604 the median RPKM of all genes provided for the file. Genes with a median RPKM ≥ 1 were
605 considered expressed in the fetal brain (n = 15,474). Likely causal RT variants (n = 4) were
606 then intersected with this list to determine gene expression status. As RPKM ≥ 1 is a rather
607 liberal cutoff, we also ranked these genes compared to other genes in our expression list;
608 these values are reported in Supplemental Table 5.
609

610 ## MEI validation by PCR

611 To validate all 9 *de novo* MEI variants (Table 2) and the homozygous insertion in
612 *PAN2* we used the following PCR protocol: primers were designed using Primer3 to make
613 products spanning the predicted insertion site (Supplemental Table 2). PCR was carried out
614 using Platinum™ Taq DNA Polymerase High Fidelity (Invitrogen); 20ng of genomic DNA
615 extracted from blood or saliva was amplified in the presence of 0.2 μM of each primer and 1
616 unit of Platinum™ Taq. Amplification was carried out using the following cycling conditions;
617 for Alu insertions: 2 min at 94°C, followed by 36 cycles of (30 sec at 94°C, 30 sec at 60°C
618 and 1 min at 68°C); for LINE1 insertions: 2 min at 94°C, followed by 36 cycles of (30 sec at

619 94°C, 30 sec at 60°C and 7 min at 68°C). PCR products were visualized using a 2% agarose
620 E-Gel® (Invitrogen).
621

622 Processed pseudogene validation by PCR and capillary sequencing

623 To validate the 2 *de novo* PPG variants (Table 2) we used the following PCR
624 protocol: primers were designed using Primer3 to make products within the exons of each
625 gene. Forward and reverse primers were then paired between exons to amplify across the
626 excised intronic regions (Supplemental Table 2). PCR was carried out using either
627 Platinum™ Taq DNA Polymerase High Fidelity (Invitrogen) or Thermo-Start Taq DNA
628 Polymerase (Thermo Scientific). Platinum™ Taq assay: 20ng of genomic DNA extracted
629 from blood or saliva was amplified in the presence of 0.2 μM of each primer and 1 unit of
630 Platinum™ Taq. Amplification was carried out using the following cycling conditions; 2 min at
631 94°C, followed by 36 cycles of (30 sec at 94°C, 30 sec at 60°C and 1 min at 68°C). Thermo-
632 Start Taq DNA Polymerase assay: 40 ng genomic DNA was amplified in the presence of 0.2
633 μM of each primer and 0.42 units of Thermo-Start Taq. Cycling conditions were as follows: 5
634 min at 95°C, 6 cycles of (30 sec at 95°C, 30 sec at 64°C and 1 min at 72°C), 6 cycles of (30
635 sec at 95°C, 30 sec at 62°C and 1 min at 72°C), 6 cycles of (30 sec at 95°C, 30 sec at 60°C
636 and 1 min at 72°C) followed by 36 cycles of (30 sec at 95°C, 30 sec at 58°C and 1 min at
637 72°C) with a final elongation of 10 min at 72°C. PCR products were visualized using a 2%
638 agarose E-Gel® (Invitrogen). PCR products were sequenced using either the forward or
639 reverse primer used in the amplification protocol by Eurofins GATC Biotech GmbH.
640 Sequence traces were aligned using SeqMan Pro 15 (Lasergene 15) and reads were
641 aligned to the human genome (hg19) using BLAT (UCSC)[55].
642

643 WGS of probands with *de novo* processed pseudogenes

644 To validate and determine the insertion site of the two identified *de novo* PPGs
645 (Table 2), we performed Illumina WGS on all individuals of each trio in which the *de novo*

646 event was identified (n = 6 individuals). Samples were first quantified with Biotium Accuclear
647 Ultra high sensitivity dsDNA Quantitative kit using Mosquito LV liquid platform, Bravo WS
648 and BMG FLUOstar Omega plate reader and cherrypicked to 500ng / 120ul using Tecan
649 liquid handling platform. Cherrypicked plates are then sheared to 450bp using a Covaris
650 LE220 instrument and subsequently purified using SPRI Select beads on Agilent Bravo WS.
651 Library construction (ER, A-tailing and ligation) was performed using 'NEB Ultra II custom kit'
652 on an Agilent Bravo WS automation system. Samples were then tagged using NextFLEX
653 Unique Dual Indexed adapter 1-96 barcodes at the ligation stage. Libraries were then
654 quantified by qPCR using Kapa Illumina ABI Sanger custom qPCR kits using a Mosquito LV
655 liquid handling platform, Bravo WS, and Roche Lightcycler. Libraries are then pooled in
656 equimolar amounts on a Beckman BioMek NX-8 liquid handling platform and normalised to
657 2.4nM for cluster generation on a c-BOT and then sequenced on the Illumina TenX
658 sequencing platform. Following sequencing, reads were aligned with BWA mem[45] (with
659 settings -t 16 -p -Y -K 100000000) to version hg19 of the human reference genome. Reads
660 were then manually inspected using the Integrative Genomics Viewer (IGV)[56] to confirm
661 presence, *de novo* status, and parent of origin of each PPG.
662

663 ## Methylation analysis

664 Methylation analyses were performed in accordance with the methods detailed in
665 Choufani, et al. [35]. In short, we downloaded publicly available methylation data from GEO
666 (accession GSE74432) generated by Choufani, et al. [35] and combined it with Illumina EPIC
667 methylation array data generated in-house for our *Alu NSD1* loss of function patient (Fig. 3b;
668 Table 2). Following data curation, we randomly sampled a subset of 15 cases (defined by
669 Choufani, et al. [35] as a carrier of an *NSD1* loss of function variant) and 15 controls to identify
670 a set of differentially methylated "signature" CpGs based on a 20% effect size cut-off and
671 0.01 FDR-corrected p-value which separated case and control individuals. Once a signature
672 was established, we used Pearson correlation to assess each sample's similarity to a mean

673 case and mean control based on differential methylation at signature CpG sites. Confirmed
674 pathogenic *NSD1* variants and control samples used to build the signature were not included
675 in the correlation analysis (Supplemental Fig. 10).
676

677 ## MEI mutation rate and burden

678 To determine the mutation rate independently for each MEI type (*Alu*, L1, SVA), we
679 utilized data generated by both DDD and the 1KGP[5]. For DDD data we filtered sites as
680 above based on our >10X coverage accessibility mask. For the 1KGP data[5], we created a
681 combined mask from three different data sources: 1) the pilot accessibility mask generated
682 by the 1KGP project phase III[57], which removes regions of the genome inaccessible to
683 variant calling, 2) reference ME sequences as identified by repeatmasker[58], as MELT is
684 unable to accurately ascertain MEIs in these regions, and 3) All sequence ±10Kbp from the
685 5' and 3' terminus of all protein-coding genes from RefSeq[59]. This mask was generated
686 separately for *Alu* and L1 and did not filter 1,113.0Mbp or 959.9Mbp of the genome,
687 respectively. The *Alu* mask was used for filtering SVA and both masks excluded both
688 allosomes. On masking the 1KGP data, we were left with a total of 10,930 autosomal MEIs
689 (8,554 *Alu*, 2,047 L1, 329 SVA). Following filtering of the DDD and 1KGP sets with their
690 corresponding masks, we used the Watterson estimator with an effective population size of
691 10,000 for all calculations to estimate the population mutation parameter, Θ[36], and mutation
692 rate, μ (Supplemental Table 4).

693 We next used our estimate of μ to determine the expected number of *de novo* MEI
694 mutations in exons, enhancers, and introns genome-wide. To calculate this value, we
695 extrapolated μ for 9,738 births (i.e. number of probands in our study), resulting in an
696 estimate of 686 *de novo* MEIs genome-wide. We then simulated this number of variants 100
697 times genome-wide using bedtools random[48] and annotated the resulting "variants"
698 identically to real variants reported in this study (e.g. as in Fig. 2a). The total number of
699 expected variants in the three categories depicted in Fig. 4 were then averaged across all

700 simulations to determine Poisson λ of *de novo* variants under neutral mutation rate. These
701 distributions were then statistically compared to the number of observed variants in DDD
702 using the *ppois* function in R (Fig. 4).
703

704 ## Data Availability

705 Sequencing, phenotype data, and variant calls for all data in this paper are accessible via
706 the European Genome-phenome Archive (EGA) under study number EGAS00001000775
707 (https://www.ebi.ac.uk/ega/studies/EGAS00001000775). Within this study, WES files of all
708 DDD families are provided as part of dataset EGAD00001004390 and RT variant calls and
709 WGS data of the two individuals with *de novo* PPG variants are provided as part of dataset
710 EGAD00001004586.
711

712 # Author Contributions

713 E.J.G performed RT variant calling and annotation, PPG algorithm design, constraint and
714 burden testing, computational analysis of mosaic variants, and initial clinical annotation and
715 together with M.E.H. designed experiments, oversaw the study, and wrote the manuscript.
716 E. P. designed and performed PCR experiments. G.G. curated and prepared DDD
717 sequencing data. S.S.G and H.I designed and conducted molecular assays to assess
718 mosaic MEIs. P.J.S. assisted in estimating genetic burden of deleterious MEIs in the human
719 population. A.S. assisted with the design of the PPG discovery algorithm. P.D., K.E.S.,
720 E.J.G., and T.S. performed variant calling and quality control of SNVs. J.H. performed
721 methylation analysis of the *NSD1* variant patient. K.E.C, E.C., K.L.L., K.P., E.R., D.R.F, and
722 H.V.F prepared clinical assessments of patients and confirmation of molecular diagnoses as
723 they relate to patient phenotype.
724

Acknowledgements

The authors wish to thank the Wellcome Sanger Institute sequencing facility staff for their assistance in preparing samples and performing sequencing experiments, all members of the DDD study for providing valuable comments during data analysis and manuscript preparation, and the DDD families – this work would not be possible without their confidence and support. We also thank Panayiotis Constantinou for helping to curate known MEI-associated cases and for annotation of affected parents as well as Hilary Martin for constructive comments during manuscript preparation. We also wish to acknowledge Jeffrey Barrett and Caroline Wright for their leadership of the DDD. The DDD study presents independent research commissioned by the Health Innovation Challenge Fund [grant number HICF-1009-003], a parallel funding partnership between Wellcome and the Department of Health, and the Wellcome Sanger Institute [grant number WT098051]. The views expressed in this publication are those of the author(s) and not necessarily those of Wellcome or the Department of Health. The study has UK Research Ethics Committee approval (10/H0305/83, granted by the Cambridge South REC, and GEN/284/12 granted by the Republic of Ireland REC). The research team acknowledges the support of the National Institute for Health Research, through the Comprehensive Clinical Research Network. This study makes use of DECIPHER (http://decipher.sanger.ac.uk), which is funded by the Wellcome.

Competing Interests

M.E.H. is a co-founder of, consultant to, and holds shares in, Congenica Ltd, a genetics diagnostic company.

750 # References

751 1. Mills, R.E., Bennett, E.A., Iskow, R.C. & Devine, S.E. Which transposable elements
752 are active in the human genome? *Trends Genet* **23**, 183-91 (2007).
753 2. Zhang, Z., Harrison, P.M., Liu, Y. & Gerstein, M. Millions of years of evolution
754 preserved: a comprehensive catalog of the processed pseudogenes in the human
755 genome. *Genome Res* **13**, 2541-58 (2003).
756 3. Stewart, C. *et al.* A comprehensive map of mobile element insertion polymorphisms
757 in humans. *PLoS Genet* **7**, e1002236 (2011).
758 4. Sudmant, P.H. *et al.* An integrated map of structural variation in 2,504 human
759 genomes. *Nature* **526**, 75-81 (2015).
760 5. Gardner, E.J. *et al.* The Mobile Element Locator Tool (MELT): population-scale
761 mobile element discovery and biology. *Genome Res* **27**, 1916-1929 (2017).
762 6. Zhang, Y., Li, S., Abyzov, A. & Gerstein, M.B. Landscape and variation of novel
763 retroduplications in 26 human populations. *PLoS Comput Biol* **13**, e1005567 (2017).
764 7. Hancks, D.C. & Kazazian, H.H., Jr. Roles for retrotransposon insertions in human
765 disease. *Mob DNA* **7**, 9 (2016).
766 8. Brandler, W.M. *et al.* Frequency and Complexity of De Novo Structural Mutation in
767 Autism. *Am J Hum Genet* **98**, 667-79 (2016).
768 9. Brandler, W.M. *et al.* Paternally inherited cis-regulatory structural variants are
769 associated with autism. *Science* **360**, 327-331 (2018).
770 10. Werling, D.M. *et al.* An analytical framework for whole-genome sequence association
771 studies and its implications for autism spectrum disorder. *Nat Genet* **50**, 727-736
772 (2018).
773 11. Hehir-Kwa, J.Y. *et al.* A high-quality human reference panel reveals the complexity
774 and distribution of genomic structural variants. *Nat Commun* **7**, 12989 (2016).
775 12. Deciphering Developmental Disorders Study. Prevalence and architecture of de novo
776 mutations in developmental disorders. *Nature* **542**, 433-438 (2017).

777 13. Martin, H.C. *et al.* Quantifying the contribution of recessive coding variation to
778 developmental disorders. *Science* (2018).
779 14. King, D.A. *et al.* Detection of structural mosaicism from targeted and whole-genome
780 sequencing data. *Genome Res* **27**, 1704-1714 (2017).
781 15. Short, P.J. *et al.* De novo mutations in regulatory elements in neurodevelopmental
782 disorders. *Nature* **555**, 611-616 (2018).
783 16. Lord, J. *et al.* The contribution of non-canonical splicing mutations to severe
784 dominant developmental disorders. *bioRxiv* (2018).
785 17. Kaplanis, J. *et al.* Mutational origins and pathogenic consequences of multinucleotide
786 mutations in 6,688 trios with developmental disorders. *bioRxiv* (2018).
787 18. Niemi, M.E.K. *et al.* Common genetic variants contribute to risk of rare severe
788 neurodevelopmental disorders. *Nature* **562**, 268-271 (2018).
789 19. Goncalves, I., Duret, L. & Mouchiroud, D. Nature and structure of human genes that
790 generate retropseudogenes. *Genome Res* **10**, 672-8 (2000).
791 20. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot
792 analysis: multitissue gene regulation in humans. *Science* **348**, 648-60 (2015).
793 21. Huang da, W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis
794 of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57 (2009).
795 22. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature*
796 **536**, 285 (2016).
797 23. Samocha, K.E. *et al.* A framework for the interpretation of de novo mutation in human
798 disease. *Nat Genet* **46**, 944-50 (2014).
799 24. Hormozdiari, F. *et al.* Rates and patterns of great ape retrotransposition. *Proc Natl*
800 *Acad Sci U S A* **110**, 13457-62 (2013).
801 25. Zhang, Y., Romanish, M.T. & Mager, D.L. Distributions of transposable elements
802 reveal hazardous zones in mammalian introns. *PLoS Comput Biol* **7**, e1002046
803 (2011).

804 26. MacArthur, D.G. *et al.* A systematic survey of loss-of-function variants in human
805 protein-coding genes. *Science* **335**, 823-8 (2012).
806 27. Kubiak, M.R. & Makalowska, I. Protein-Coding Genes' Retrocopies and Their
807 Functions. *Viruses* **9**(2017).
808 28. Gilbert, N., Lutz, S., Morrish, T.A. & Moran, J.V. Multiple fates of L1
809 retrotransposition intermediates in cultured human cells. *Mol Cell Biol* **25**, 7780-95
810 (2005).
811 29. Jonsson, H. *et al.* Parental influence on human germline de novo mutations in 1,548
812 trios from Iceland. *Nature* **549**, 519-522 (2017).
813 30. Firth, H.V. *et al.* DECIPHER: Database of Chromosomal Imbalance and Phenotype
814 in Humans Using Ensembl Resources. *Am J Hum Genet* **84**, 524-33 (2009).
815 31. Wright, C.F. *et al.* Making new genetic diagnoses with old data: iterative reanalysis
816 and reporting from genome-wide data in 1,133 families with developmental disorders.
817 *Genet Med* (2018).
818 32. Kersey, P.J. *et al.* Ensembl Genomes 2016: more genomes, more complexity.
819 *Nucleic Acids Res* **44**, D574-80 (2016).
820 33. Li, M. *et al.* Integrative functional genomic analysis of human brain development and
821 neuropsychiatric risks. *Science* **362**(2018).
822 34. Riecken, L.B. *et al.* Inhibition of RAS activation due to a homozygous ezrin variant in
823 patients with profound intellectual disability. *Hum Mutat* **36**, 270-8 (2015).
824 35. Choufani, S. *et al.* NSD1 mutations generate a genome-wide DNA methylation
825 signature. *Nat Commun* **6**, 10207 (2015).
826 36. Watterson, G.A. On the number of segregating sites in genetical models without
827 recombination. *Theor Popul Biol* **7**, 256-76 (1975).
828 37. Ewing, A.D. & Kazazian, H.H., Jr. High-throughput sequencing reveals extensive
829 variation in human-specific L1 content in individual human genomes. *Genome Res*
830 **20**, 1262-70 (2010).

831 38. Torene, R.I. *et al.* Mobile elements are a significant source of pathogenic variants in
832 rare human diseases. *To be submitted* (2018).
833 39. Witherspoon, D.J. *et al.* Mobile element scanning (ME-Scan) identifies thousands of
834 novel Alu insertions in diverse human populations. *Genome Res* **23**, 1170-81 (2013).
835 40. Wimmer, K., Callens, T., Wernstedt, A. & Messiaen, L. The NF1 gene contains
836 hotspots for L1 endonuclease-dependent de novo insertion. *PLoS Genet* **7**,
837 e1002371 (2011).
838 41. Jurka, J. Sequence patterns indicate an enzymatic involvement in integration of
839 mammalian retroposons. *Proc Natl Acad Sci U S A* **94**, 1872-7 (1997).
840 42. Sultana, T. *et al.* The Landscape of L1 Retrotransposons in the Human Genome Is
841 Shaped by Pre-insertion Sequence Biases and Post-insertion Selection. *Mol Cell*
842 (2019).
843 43. Aneichyk, T. *et al.* Dissecting the Causal Mechanism of X-Linked Dystonia-
844 Parkinsonism by Integrating Genome and Transcriptome Assembly. *Cell* **172**, 897-
845 909.e21 (2018).
846 44. Wright, C.F. *et al.* Genetic diagnosis of developmental disorders in the DDD study: a
847 scalable analysis of genome-wide research data. *Lancet* **385**, 1305-14 (2015).
848 45. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler
849 transform. *Bioinformatics* **26**, 589-95 (2010).
850 46. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-8
851 (2011).
852 47. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122
853 (2016).
854 48. Quinlan, A.R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr*
855 *Protoc Bioinformatics* **47**, 11.12.1-11.12.34 (2014).
856 49. May, D. *et al.* Large-scale discovery of enhancers from human heart tissue. *Nat*
857 *Genet* **44**, 89-93 (2011).

858 50. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L.A. VISTA Enhancer Browser--a
859 database of tissue-specific human enhancers. *Nucleic Acids Res* **35**, D88-92 (2007).
860 51. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and
861 yeast genomes. *Genome Res* **15**, 1034-50 (2005).
862 52. Chaisson, M.J. & Tesler, G. Mapping single molecule sequencing reads using basic
863 local alignment with successive refinement (BLASR): application and theory. *BMC*
864 *Bioinformatics* **13**, 238 (2012).
865 53. Potapov, V. & Ong, J.L. Examining Sources of Error in PCR by Single-Molecule
866 Sequencing. *PLoS One* **12**, e0169774 (2017).
867 54. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for
868 analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-303 (2010).
869 55. Tyner, C. *et al.* The UCSC Genome Browser database: 2017 update. *Nucleic Acids*
870 *Res* (2016).
871 56. Thorvaldsdottir, H., Robinson, J.T. & Mesirov, J.P. Integrative Genomics Viewer
872 (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform*
873 **14**, 178-92 (2013).
874 57. 1000 Genomes Project Consortium. A global reference for human genetic variation.
875 *Nature* **526**, 68-74 (2015).
876 58. Smit AFA, H.R., Green P. RepeatMasker Open-3.0. (1996-2010).
877 59. O'Leary, N.A. *et al.* Reference sequence (RefSeq) database at NCBI: current status,
878 taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733-45
879 (2016).
880