Contribution of Retrotransposition to Developmental Disorders

Eugene J. Gardner[1], Elena Prigmore[1], Giuseppe Gallone[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 whole exome sequencing (WES) 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 the single largest interrogation of the impact of RT activity on the coding genome to date.

39 # Main

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

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

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

79 ## Results

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

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

93

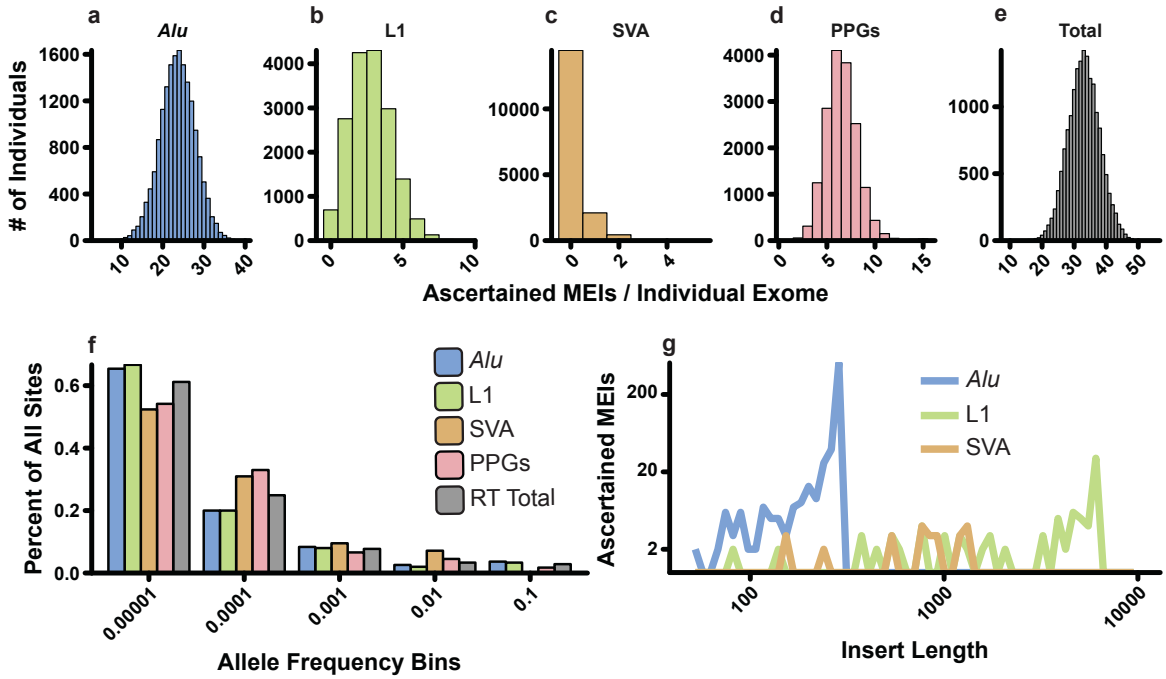
	Total Sites	**Mean Sites Per Individual**	**Total de novo Sites**
Alu	917	23.6±4.2	7
LINE-1	167	2.8±1.5	2
SVA	45	0.2±0.5	0
Total - MEI	*1,129*	*26.6±4.7*	*9*
Processed Pseudogenes (PPGs)	576	6.9±2.6	2
Total – MEI + PPG	*1,705*	*33.5±7.3*	*11*

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

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

97 As our study is the first to discover MEIs directly from WES on a large scale, we first
98 utilized matched sample WGS data to determine if MELT could ascertain MEI variants
99 reliably from WES data. We compared MEI variants identified by MELT within the DDD WES
100 data to both WGS data generated on the same individuals and population MEI data
101 previously generated from the 1000 Genomes Project Phase 3 (1KPG)[4-6] WGS data. The
102 latter comparison was to ensure that the number of exonic MEIs identified within DDD WES
103 data was concordant with expectations at the individual and population level. When
104 comparing our WES genotypes to WGS in identical individuals, we had a genotype
105 concordance rate of 94.46% (93.93% *Alu*, 97.29% L1, 98.25% SVA) among calls with at
106 least 10X coverage in our WES data. In total, we were able to re-identify 1,355 (1,289 *Alu*,
107 160 L1, 1 SVA) MEI genotypes, or 84.5% of all heterozygous or homozygous genotypes
108 identifiable with WGS in WES bait regions (Supplemental Table 1). Based on these findings
109 we were confident that MELT was appropriately calibrated to ascertain MEIs in WES data.
110 We identified 1,129 MEI variants and 576 polymorphic PPGs, with each individual's
111 exome containing on average 33.5±7.3 variants. All MEIs were genotyped across all
112 individuals to form a comprehensive catalogue of RT-derived variation within and adjacent to
113 (±50 bp) sequences targeted in the WES assay (Methods), including coding exons and
114 targeted non-coding elements (Table 1; Fig. 1). The average time to assess a single family
115 for RT-derived events was approximately 15 minutes and the rate of false findings was low
116 (1 incorrect *de novo* variant per every 320 patients; either a false positive variant or false
117 negative genotype in at least one parent). As expected, the total number of variants per
118 individual for each RT class (Fig. 1a-d) as well as combined number of RT events (Fig. 1e)
119 approximated a Poisson distribution. The vast majority of variants are rare (AF < $1x10^{-4}$; Fig.
120 1f), with >65% of *Alu* and L1 variants identified in fewer than 4 unrelated individuals. SVA
121 and PPGs appear to be moderately under ascertained compared to *Alu* and L1 at lower AFs,
122 with >50% of variants identified in the lowest AF bin. The length estimates for the three MEI
123 classes largely fit the findings of previous studies (Fig. 1g)[4,5], except in the case of full-length
124 L1 elements (i.e. L1s >6kbp in length). In our study, we identified a total of 26 full-length L1

125 MEIs (16.0% of measured variants), while in previous studies ~30% of all L1 MEIs are full-
126 length. As MELT was previously validated for MEI length measurement[5], our conclusion is
127 that we have lower sensitivity for ascertainment of longer L1s from WES.



128

129 *Figure 1: The DDD RT call set*

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

136

137 We next sought to ensure that our total number of ascertained RT variants, both on a
138 population and individual basis, accorded with previously published WGS data[4,5]. On a
139 population level, WES did not appreciably limit our overall sensitivity compared to WGS
140 sampled data. When we compared a downsampled version of our call set to the 1KGP, our
141 total number of *Alu* and SVA variants fell within the expected distribution, while L1 was close
142 to expectation (Supplemental Fig. 2). To assess the quality of the PPG call set, we

143 compared PPG allele counts (i.e. total number of individuals with a retro-duplication of a
144 given gene) to a recent assessment of PPGs in samples sequenced as part of the 1KGP[6].
145 Generally, PPGs identified in both data sets shared similar relative allele counts ($r^2 = 0.64$)
146 and variants identified in this study but missing from Zhang et. al.[6] are typically rare
147 (Supplemental Fig. 3). To further validate our approach and ensure that the identified PPG
148 donor genes fit with previously identified patterns of germline PPG formation[2,19], we
149 assessed each donor gene for both functional annotation and expression across 30 tissue
150 types analyzed by the GTEx consortium[20]. The major functional cluster (DAVID[21] enrichment
151 score 8.82) belonged to genes involved in the ribosomal and translational machinery,
152 consistent with previous findings involving fixed PPGs in the human genome[2]. Our
153 expression analysis likewise confirmed previous findings[19], and shows that donor genes that
154 give rise to PPGs are more highly expressed in a large number of tissues compared to non-
155 retroposed genes (Wilcoxon rank sum $p < 1x10^{-3}$ for all tissues; Supplemental Fig. 4).
156 Additionally, while it could be assumed that increased germ-line expression of a gene may
157 play a role in increased probability of PPG generation, when we compared PPG donor gene
158 expression in the testis and ovary to that in other tissues, the majority of tissues (20/29,
159 identical tissues for ovary and testis) showed statistically identical patterns of donor gene
160 expression (Wilcoxon rank sum $p > 1x10^{-3}$; Supplemental Fig. 4).
161

162 ## Coding RT burden and constraint

163 As expected for WES, the vast majority (84.9%) of MEIs impacted the coding or
164 intronic sequence of a protein-coding gene or a regulatory element targeted in the
165 augmented WES assay described in Short et. al.[15] (Fig. 2a). While the number of MEIs
166 identified in this study, based on the proportion of the genome assayed, represent only 2.2%
167 of genome-wide MEI variants, we have ascertained over five-fold more variants that directly
168 impact exons than the next largest study (Supplemental Fig. 5)[4,5].
169 Our large collection of coding variants allowed us to examine the evolutionary forces
170 acting on coding MEI variation (Fig. 2b). To examine selective constraint, we utilized two

171 common measures: the proportion of variants observed in only one individual (e.g.
172 singletons)[22] and the proportion of variants found in genes likely to be intolerant of loss of
173 function (LoF) as determined by the pLI score[23]. To avoid issues of relatedness and the
174 potential for clinical ascertainment bias for pathogenic MEIs in individual DD patients, only
175 the 17,032 unaffected parents sequenced as part of DDD were included in our analysis.
176 MEIs which directly impact exons are under strong selective constraint, indistinguishable
177 from that of both nonsense and essential splice site SNVs (Fig. 2b). Interestingly, we did not
178 find any sign of selection acting on intronic MEIs as they appear to be constrained similarly
179 to synonymous SNVs. In contrast to previous studies[24,25], we did not find a statistically
180 significant (χ^2 $p < 0.05$) bias towards intronic MEIs inserted in the antisense orientation of
181 the gene in which they are found (Supplemental Fig. 6). This is likely not a repudiation of
182 such work, but attributable to the relatively small number of intronic events we identified as
183 part of our analysis compared to WGS[4,24] or reference genome-based[25] studies. To put our
184 findings on exonic MEI constraint into perspective with other forms of variation, every human
185 genome will harbor approximately one (0.76±0.62 per individual) MEI which directly impacts
186 protein-coding sequence. Since MEIs are similar to nonsense SNVs in terms of
187 deleteriousness (Fig. 2), MEIs thus make up roughly 1%[22,26] of all coding PTVs (among
188 SNVs, InDels, and large CNVs) in each individual human genome.

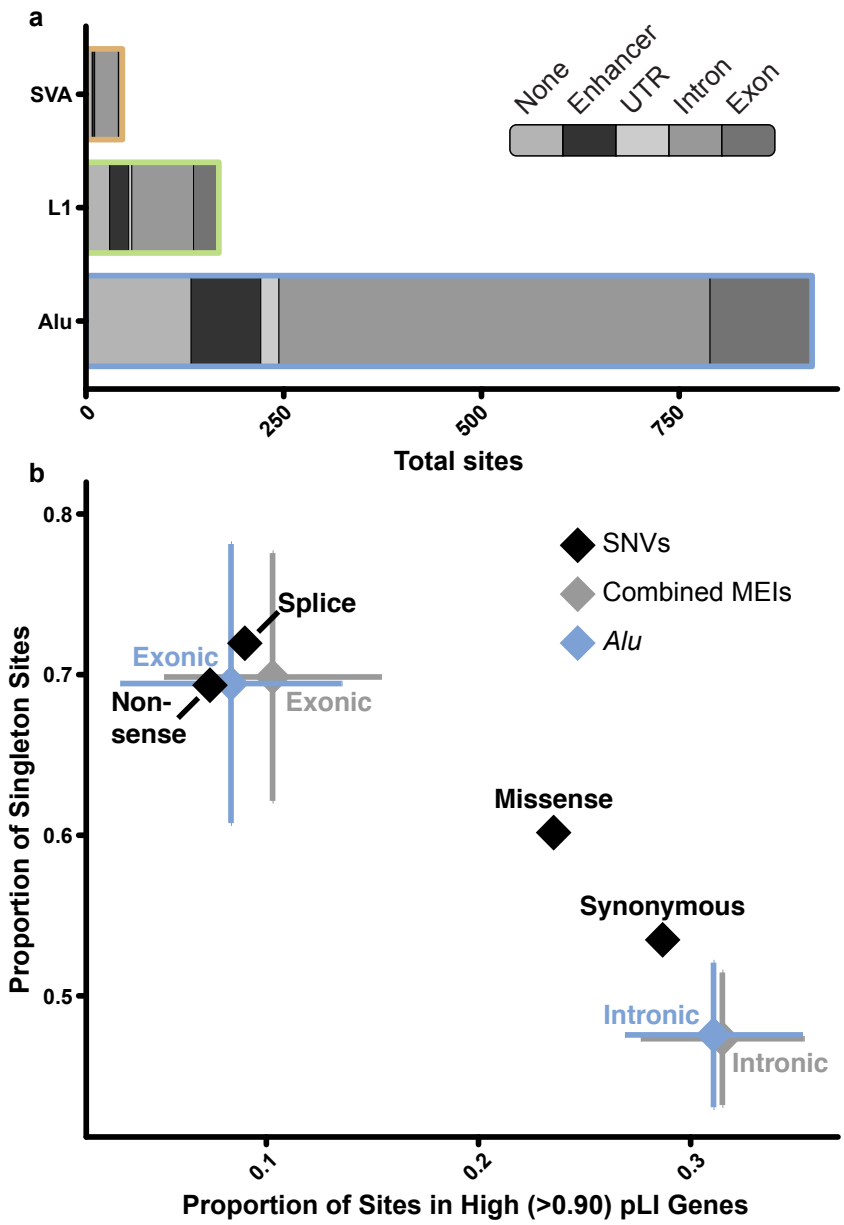


Figure 2: Coding constraint on MEIs

(**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.

199 While we were unable to perform similar population genetic analyses for PPG
200 events, due to the difficulty of resolving the putative insertion site with WES data and thus
201 distinguishing between different PPGs for the same donor gene, we were able to assess the
202 propensity for specific genes to give rise to PPGs based on their selective constraint. We
203 observed that PPG donor genes were significantly enriched for genes that are highly
204 intolerant of loss of function variation (pLI > 0.9). High pLI genes make up 25.3% of donor
205 genes, compared to 17.6% of all protein-coding genes (χ^2 p = 2.4×10^{-6})[22]. This observation is
206 likely driven by loss of function intolerant genes being more likely to be highly expressed in
207 multiple tissues[22], similar to genes known to have been retroduplicated (Supplemental Fig.
208 4)[19]. This observation implies that PPG events rarely strongly perturb the function of their
209 donor gene – despite several previously documented instances of PPGs impacting
210 expression or functionality of their donor gene[27].
211

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

213 Using the computational approaches outlined above we identified a total of 11 germ-
214 line *de novo* RT variants (Table 2). Our findings include coding, noncoding, pathogenic and
215 benign variants, as well as, to our knowledge, the first *de novo* MEI identified in a pair of
216 monozygotic twins (Supplemental Fig. 7). All *de novo* RT variants were confirmed via a PCR
217 assay specific to the RT class (Fig. 3; Supplemental Fig. 7; Supplemental Table 2) and,
218 where possible, inspected for poly(A) tail and target site duplication – hallmarks of *bona fide*
219 RT activity[28]. We identified no *de novo* RT variants which localized to the non-coding
220 elements included on the WES capture, which falls in line with expectations based on
221 mutation rate estimates (Fig. 4b). We also attempted to determine the parental origin of each
222 RT event using SNVs located on sequencing reads which support the RT insertion (Table 2).
223 Of the 11 *de novo* RT events, we were able to phase three variants, all to the father. While
224 this finding is not statistically significant (χ^2 p = 0.083), it fits with previous findings that the

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

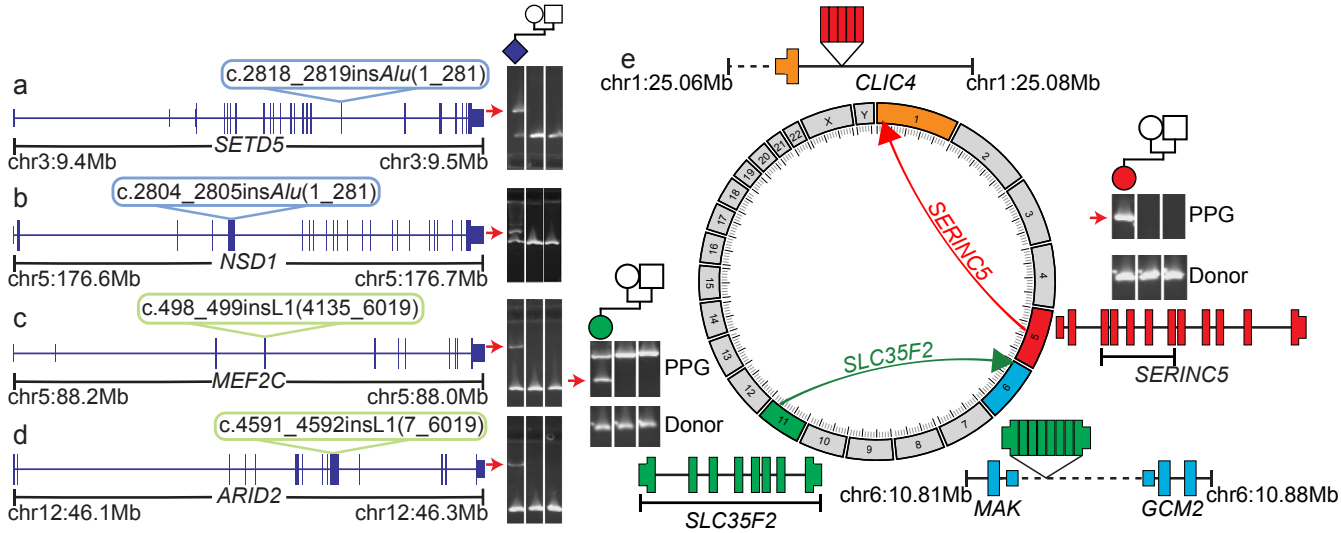
227 Nine of our validated *de novo* mutations were MEIs (7 *Alu*, 2 L1), or a rate of
228 approximately one *de novo* event per every 1,000 patient exomes sequenced (9/9,738). As
229 expected, based on both the total number of polymorphisms[3-5] and mutation rate (Table 1;
230 Supplemental Table 4), we identified more *Alu de novo* variants than the other RT classes.
231 We also identified 2 PPG germ-line *de novo* variants, or approximately one new PPG per
232 every 5,000 patient whole genomes sequenced (2/9,738). As a further quality control for
233 PPGs, we capillary sequenced all resulting PCR products to confirm the gene of origin
234 (Supplemental Data 1) and performed WGS to identify the PPG insertion site. We were able
235 to localize the *SERINC5* PPG to an ~50Kbp intron of the gene *CLIC4* and the *SLC35F2*
236 event to an intergenic region between the genes *MAK* and *GCM2* (Fig. 3e). Neither of the
237 events directly impacted coding sequence and *CLIC4* is neither under strong selective
238 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	1.00E+00	confirmed,monoallelic	280818	True	Father	
chr5:176638159	*Alu*	Exonic	ENSG00000165671	NSD1	1.00E+00	confirmed,monoallelic	259118	True	Unknown	Included in Wright et. al.[31]
chr6:159190834	*Alu*	Exonic	ENSG00000092820	EZR	9.88E-01	None	300984	False	Unknown	
chr7:77552086	*Alu*	Exonic	ENSG00000006576	PHTF2	2.49E-02	None	271388	False	Father	
chr3:135913800	*Alu*	Intronic	ENSG00000174579	MSL2	8.90E-01	None	292325	False	Unknown	
chr3:148614204	*Alu*	Intronic	ENSG00000163751	CPA3	1.28E-12	None	270426; 270428	False	Unknown	Monozygotic twins
chr3:172480619	*Alu*	Intronic	ENSG00000114346	ECT2	2.56E-05	None	307591	False	Unknown	
chr12:46246325	L1	Exonic	ENSG00000189079	ARID2	1.00E+00	probable,monoallelic	264759	True	Unknown	
chr5:88100580	L1	Exonic	ENSG00000081189	MEF2C	4.25E-03	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	9.46E-03	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).

245 Each *de novo* mutation was then compared to known DD-associated genes (using
246 the Developmental Disorders Genotype-to-Phenotype database – DDG2P) to identify
247 potentially pathogenic variants (Table 2). Of the mutations identified, four directly inserted
248 into coding exons of DD-associated genes (Fig. 3, Table 2) with all four found in genes
249 statistically enriched for PTVs[12] and therefore likely to operate by a LoF mechanism. We did
250 not identify any intronic *de novo* mutations likely to be pathogenic (Fig. 3a-d; Supplemental
251 Fig. 7). An additional mutation inserted into the coding sequence of a strongly LoF-intolerant
252 gene, *EZR* (pLI = 0.99; Supplemental Fig. 7), but we could not directly attribute it to the
253 patient's phenotype due to lack of significant enrichment for PTVs, although there is prior
254 evidence for a role in a familial DD syndrome[33]. The four mutations in DD-associated genes
255 were reported to the referring clinician for clinical interpretation based on both initially
256 reported and updated phenotypes (Supplemental Table 3). Three out of four reported
257 mutations (*NSD1*, *MEF2C*, *ARID2*) were subsequently deemed to be likely causative of the
258 patient's phenotype (Supplemental Table 3) by the referring clinician. The fourth patient, with
259 an *Alu* insertion in *SETD5* (Fig. 3a), has clinical features (polydactyly and truncal obesity;
260 Supplemental Table 3) more suggestive of a ciliopathy. As such, the identified MEI is
261 unlikely to be the sole cause for the patient's DD but may contribute to a composite
262 phenotype.



263

264 *Figure 3: RT-derived de novos in the DDD*

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

278 We also examined our dataset for inherited rare pathogenic RT variants. We
279 evaluated variants inherited from an affected parent, bi-allelic inheritance (either a
280 homozygous MEI or a heterozygous MEI paired with another variant class), and X-linked
281 variants maternally inherited by affected males. We did not identify any rare MEI variants
282 inherited from an affected parent nor any compound heterozygous individuals with a rare
283 MEI and a non-MEI PTV (e.g. SNV/InDel) impacting the same gene. We did identify a single
284 proband-specific homozygous MEI inserted into an exon of *PAN2* which was unique to a
285 single family. This gene was recently identified as nominally significant (genome-wide p =
286 4.2×10^{-4}) in a study investigating the role of recessive variants in DD[13], although more data
287 are required to be confident of its association to DD. We also identified a total of 22 (14 *Alu*,
288 7 L1, 1 SVA) polymorphic MEIs on the X chromosome, of which 4 (3 *Alu*, 1 L1) directly
289 impacted protein-coding sequence. Of these variants, none were at a low enough allele
290 frequency to be reasonably DD-associated, were located within a gene associated with DD,
291 nor fit an inheritance pattern consistent with X-linked disease.
292

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

294 Based on our findings, in the coding and peri-coding portion of the genome, one out
295 of every 2,434 DD cases (0.04%±0.04; 95% CI) is directly attributable to RT-derived
296 mutagenesis. To determine both if our observed number of *de novo* variants meets
297 expectation and if our patient cohort is enriched for causal *de novo* RT events, we estimated
298 the population mutation parameter, Θ[34], from the unaffected parents in the DDD study and
299 from the 1KGP[4,5] (Supplemental Table 4). The resulting calculation gives very similar
300 estimates of MEI mutation rate (combined across *Alu*, L1, SVA) of between 1.4×10^{-11} (1KGP)
301 and 1.2×10^{-11} (DDD) variants per bp per generation (μ), or ~1 new MEI genome-wide per
302 every 12 to 14 births – largely concordant with prior estimates from smaller WGS
303 datasets[3,35].
304

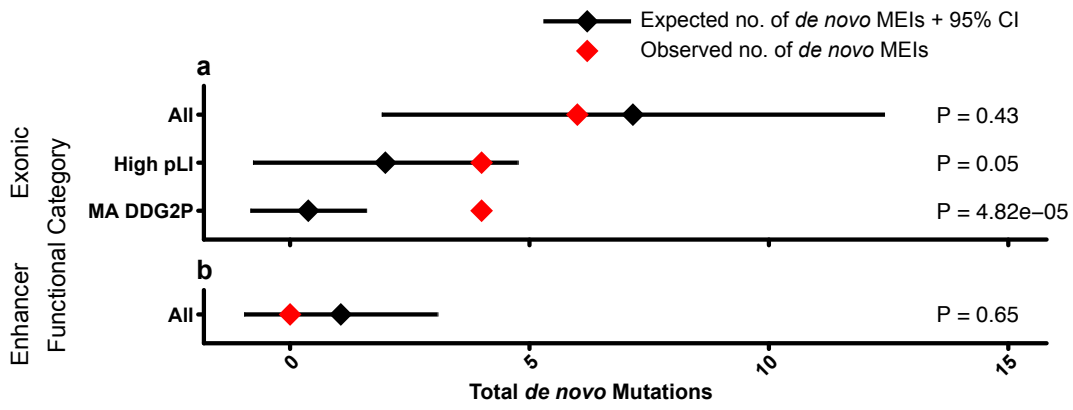


Figure 4: Estimating enrichment of deleterious MEIs

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 ($p = 4.3 \times 10^{-5}$).

323 # Discussion

324 Here we have described the development, validation and exemplification at scale of
325 an analytical pipeline for the rapid assessment of patient genomes for RT variants. We have
326 used these approaches to present the largest study examining the coding genome for RT-
327 derived variation to date (Table 1; Fig. 1). With this dataset, we first demonstrated that
328 exonic MEIs (regardless of insertion length) are under selective constraint on par with
329 protein-truncating SNVs (Fig. 2, Supplemental Fig. 5). We identified four likely pathogenic
330 RT mutations, two *Alu* and two L1 insertions (Fig. 3), all of which arose *de novo* in known
331 haploinsufficient DD-associated genes (Fig. 3a-d), implying that dominant loss-of-function is
332 the major mode of pathogenic exonic RT variation. Finally, we estimated the genome-wide
333 MEI mutation rate and used it to determine that DDD probands are enriched for damaging
334 RT variation within exons of dominant DD-associated genes (Fig. 4a).

335 The total number of polymorphic, exonic RT variants identified in DDD is concordant
336 with previous studies characterizing MEI variation[3,5,37]. Pathogenic MEIs make up 0.04% of
337 diagnoses in the DDD study (4/9,738 probands), a small yet individually significant collection
338 of diagnostic variants. Reassuringly, our proportion of diagnostic variants in DDD is
339 statistically identical to the 7/11,011 (0.06%) diagnostic rate for neurodevelopmental disorder
340 patients as determined by Torene et. al.[36] (Fisher's exact test $p = 0.56$). Unlike Torene et.
341 al.[36], we did not identify a causative inherited MEI, although this difference is not statistically
342 significant (Fisher's exact test $p = 0.51$). We infer that despite making up a significant
343 proportion of reported MEI variants in the clinical literature[7], bi-allelic or X-linked MEI events
344 are a less frequent class of pathogenic variant in developmental disorders. This is in keeping
345 with recent estimates[13] that in a largely outbred clinical population, such as in the UK,
346 recessive disorders caused by coding variants account for a much smaller fraction of
347 patients than dominant disorders.

348 Interestingly, it appears that the contribution of diagnostic RT variants may vary
349 among diseases. Wimmer et. al.[38] reported a total of 13 diagnostic, exonic MEI variants in

350 4,500 neurofibromatosis type I patients (0.3% of patients). This rate is seven times higher
351 than that observed in DDD or Torene et. al.[36] and was attributed to a potential RT mutation
352 "hotspot" associated with the canonical L1 endonuclease cleavage site of 3'-AA/TTTT-5'[39]
353 within the neurofibromatosis-associated gene, *NF1*. Further work is needed to investigate
354 the role of sequence context in determining the overall genomic landscape of RT-mediated
355 disease. Analogously, inclusion of sequence context into the SNV mutation model noticeably
356 improved the ability to determine enrichment/depletion of deleterious SNVs within genes[22,23].
357 Our study is clearly limited in that we only identified ~2% of the RT variants in each
358 individual human genome[4,5]. Despite a number of known disease-associated intronic MEIs in
359 the literature, we did not identify a pathogenic intronic MEI. As such, it remains an open
360 question as to what contribution RT mutations in the noncoding genome plays in the etiology
361 of DD. While it appears that the contribution of regulatory elements to DD is relatively small,
362 as defined by this (Fig. 4b) and other studies[15], previous work has identified a significant
363 signature of purifying selection against MEI events within 100bp of exons[25] – variants which
364 our study could potentially identify. As our data suggests that the majority of DD cases with
365 pathogenic coding MEIs are due to *de novo* insertions (Table 2; Fig. 3), we conjecture that
366 most additional DD-associated MEIs may be located in the introns of known DD-causing
367 genes and disrupt splicing – a known disease mechanism attributable to RT-derived
368 mutagenesis[7,38,40]. Simulations suggest that under a null genome-wide mutation model we
369 should expect to observe 12.5 (5.5-19.4, 95% CI) *de novo* intronic RT mutations in dominant
370 DD-associated genes in a population sample of 9,738 individuals. As such, a WGS study of
371 a clinical population of similar size to that analyzed here should be well powered to estimate
372 the pathogenic contribution of intronic MEIs.
373 *De novo* MEIs are typically readily interpretable with modest informatics expertise,
374 and represent a clinically relevant class of variation to assay in clinical bioinformatics
375 pipelines. While we ultimately find that the overall burden of RT-attributable disease is
376 relatively low in the human population, it is nonetheless an important consideration when
377 elucidating the genetic basis of DD in individual patients.

Online Methods

Patient recruitment and sequencing

A total of 13,462 patients were recruited from 24 clinical genetics centers from throughout the United Kingdom and the Republic of Ireland as previously described[41]. Informed consent was obtained for all families and the study was approved by the UK Research Ethics Committee (10/H0305/83, granted by the Cambridge South Research Ethics Committee and GEN/284/12, granted by the Republic of Ireland Research Ethics Committee). For the purposes of this study, individuals that were not recruited as part of a trio (e.g. individual patients or patients with just one parent), were included on the DDD sample blacklist, or failed to meet MELT QC requirements[5] were excluded from downstream analysis (leaving n = 9,738 probands; 28,132 individuals). Sequencing and SNV/InDel calling of families were performed as previously described[12].

Processed pseudogene pipeline development

PPGs, particularly young polymorphic events, share highly homologous sequence with the source gene from which they are derived. Consequently, the WES bait capture method will capture both DNA from the original “donor” gene and the new “daughter” copy. This allows, compared with our approach for MEI discovery, for ascertainment of PPGs genome-wide. While this approach does come with limitations, such as difficulty in identifying insertion variants, we can still determine events per individual.

Our discovery pipeline functions in two steps: first we collect read evidence on an individual level to determine which genes have been retroduplicated in that individual (Supplemental Fig. 1). Second, we determine presence/absence of each PPG in every individual in the DDD cohort based on the gene models built in the first step. In step one, we iterate over all genes in the ENSEMBL gene database which have a determined pLI score[22] and collect discordant read pairs (DRPs) which map between exons and have an insert size >99.5% of all other reads in the sample. If more than four reads linking two exons are found,

405 the gene is considered to be retroduplicated elsewhere in the genome. In step two, for each
406 gene identified in step one, all evidence across all PPG positive individuals are pooled to
407 make a model of the PPG. This model is then used to check for DRP and split read pair
408 (SRP) evidence in all genomes. If an individual has at least 5 total read pairs of supporting
409 evidence with at least one SRP and one DRP, an individual is considered positive for the
410 given PPG. All genes and individuals were combined into a flat file listing presence or
411 absence of a given PPG in each individual. Source code and more information is available
412 online at github: https://github.com/eugenegardner/Retrogene.git
413

414 MEI call set generation and consequence annotation

415 To identify MEIs in the DDD WES data we utilized the previously published Mobile
416 Element Locator Tool (MELT)[5]. MELT was run with default parameters (except the '-exome'
417 flag during IndivAnalysis) using 'Split' mode to generate a final unified VCF-format file[42] of all
418 28,132 unfiltered individuals independently for each MEI type (*Alu*, L1, SVA). Following initial
419 data set generation, we found that a subset of variants internal or adjacent to (±50bp) low
420 complexity repeats (defined here as a run of sequence >= 15bp composed of two or fewer
421 nucleotides) were likely false positive. As such, we added an additional filter to the final
422 MELT VCF, lc (low complexity), which removes such false positives from downstream
423 analysis. Variants that could not be genotyped in at least 25% of individuals, had ≤ 2 split
424 reads, had MELT ASSESS score < 3, or had any value in the VCF FILTER column other
425 than PASS or rSD were filtered.
426 To generate consequences plotted in Fig. 2a, all MEIs were annotated using Variant
427 Effect Predictor v88 (VEP)[43] and intersected with bedtools intersect[44] to enhancers (one of
428 heart[45], VISTA[46], or highly evolutionarily conserved[47]) included on the DDD WES capture[15].
429 Only a single consequence was retained for each variant, with priority given to enhancer
430 annotation. Primary transcript as determined by VEP was used for all gene-based
431 consequences, pLI score[22] annotation, and DDG2P disease association (Table 2).

432

433 Quality Control of RT data using WGS and 1KGP

434 To determine if our MEI WES call set was biased compared to WGS data, we
435 performed two independent comparisons: 1.) to high coverage (>30x) WGS data generated
436 for a subset of DDD trios and 2.) to a published collection of MEIs from 1KGP phase III[5].
437 For WGS quality-control, we used a subset of 30 DDD trios (n = 90 individuals) which
438 were previously whole genome sequenced. MEI discovery using MELT[5] on all 90 individuals
439 was performed and filtered identically to WES data. Genotypes identified in the WGS data
440 but not in WES were then separated based on coverage in the corresponding WES.
441 Genotypes in low coverage areas (<10x) were considered not possible (n.p), while variants
442 where coverage was greater than 10x are considered not detected (n.d). All remaining
443 genotypes were than compared for identity between WGS and WES results (Supplemental
444 Table 1)
445 To compare the DDD MEI call set to the 1KGP, we first filtered 1KGP calls to
446 variants with >10x coverage in 1,000 randomly sampled WES individuals (leaving 318 *Alu*,
447 81 L1, 26 SVA). We then randomly selected 2,453 DDD parents 1,000 times, retaining only
448 loci present in downsampled individuals[4,5]. The resulting distribution was then compared to
449 the observed number of variants in the 1KGP-masked data to generate z-scores
450 independently for all three MEI types (Supplemental Fig. 2).
451 To compare our PPG dataset to Zhang et. al.[6], we downloaded provided
452 supplemental tables. We then summed the total number of unique events per person and
453 determined “allele counts” for each gene reported. Genes were then matched between our
454 call set and Zhang et. al.[6] using ENSEMBL gene identifiers and allele counts between each
455 data set were plotted to create Supplemental Fig. 3.
456

457 ## GTEx annotation of processed pseudogenes

458 To determine RNA expression levels of donor genes which gave rise to PPGs
459 identified in this study, we queried transcript per kilobase per megabase of sequencing
460 (TPM) scores for all genes in 30 tissues assessed by the current GTEx v7 release (available
461 at https://gtexportal.org/home/datasets). Only the 18,225 protein-coding genes which were
462 assessed for gene PPGs by our project were retained for subsequent analysis. TPM values
463 were then averaged across all GTEx individuals for a given tissue to generate a mean TPM
464 value as plotted in Supplemental Fig. 4. Nonparametric Wilcoxon rank-sum tests were
465 performed using the wilcox.test function in R with default parameters to generate p values
466 for both within tissue and between tissue comparisons.

467

468 ## SNV Variant Calling and Quality Control

469 To call SNVs from all DDD individuals we utilized GATK v3.5[48] in three steps using
470 default settings. First, we called variants in individual samples using HaplotypeCaller. Next,
471 individual VCF files were processed in 200 individual batches using CombineGVCFs.
472 Finally, all batched VCFs were passed to GenotypeGVCF to generate a final joint-called
473 VCF file. This file was then annotated used VEP v88[43]. Unaffected parents (n = 17,032
474 individuals) were then extracted from this VCF and only variants with an allele count greater
475 than 1 in these individuals were retained.

476 For initial filtering, we removed SNVs with a VQSLOD < -2.7971, depth < 10, and
477 genotype quality < 20. We next performed more extensive QC using a 'missingness' score
478 identical to the method described in Martin et. al.[13]. In short, each genotype at a given
479 variant was assessed for genotype quality (GQ), depth (DP), and a binomial test for allelic
480 depth (i.e. number of alternate versus reference supporting reads; AD). If a given genotype
481 had GQ <20, DP < 7, or AD p-value < 0.001 it was considered 'missing'. If more than 50% of
482 genotypes for a given variant were missing, the variant was subsequently filtered from final

483 analysis. Allele frequencies were recalculated based on included individuals while
484 accounting for missing genotypes.
485

486 SNV and MEI constraint

487 As sensitivity of variant discovery can bias our results, we generated an "accessibility
488 mask" of the DDD WES data where we expect our variant ascertainment sensitivity to be
489 >95% (Supplemental Fig. 8)[5]. Our mask thus includes only regions of the genome that
490 contain at least 10X average coverage in a mean cohort of 1,000 randomly selected
491 individuals for a total of 74.2Mbp, or ~2.3% of the genome (Supplemental Table 4). Using
492 this mask, we filtered our original 1,129 variants down to 828 (660 *Alu*, 109 L1, 31 SVA)
493 variants in unaffected parents (n = 17,032 individuals). Parents were determined to be
494 affected either by the referring clinician or, where ambiguous, through manual curation of
495 HPO terms for a matching parent-offspring phenotype.
496 Using this mask subset of variants, we determined genomic constraint as shown in
497 Figure 2b. Allele frequency values were recalculated for all variants, and a pLI score[22] for
498 each MEI was added as described above. MEIs which did not insert into a gene or inserted
499 into a gene without a calculated pLI score[22] were excluded from subsequent analysis. We
500 then calculated proportion of singleton variants and proportion of variants in high pLI genes
501 independently for *Alu* and, due to low overall numbers of the other MEI subtypes, for a
502 combined set of *Alu*, L1, and SVA. SNVs annotated as nonsense, missense, synonymous,
503 or splice acceptor/donor (splice in Fig. 2b) as determined by VEP v88[43] were extracted from
504 the SNV VCF files described above and used to calculate singleton and pLI proportion
505 identically to MEIs.
506

507 Mobile element insertion validation by PCR

508 To validate all 9 *de novo* MEI variants (Table 2) and the homozygous insertion in
509 *PAN2* we used the following PCR protocol: primers were designed using Primer3 to make

510 products spanning the predicted insertion site (Supplemental Table 2). PCR was carried out
511 using Platinum™ Taq DNA Polymerase High Fidelity (Invitrogen); 20ng of genomic DNA
512 extracted from blood or saliva was amplified in the presence of 0.2 μM of each primer and 1
513 unit of Platinum™ Taq. Amplification was carried out using the following cycling conditions;
514 for Alu insertions: 2 min at 94°C, followed by 36 cycles of (30 sec at 94°C, 30 sec at 60°C
515 and 1 min at 68°C); for LINE1 insertions: 2 min at 94°C, followed by 36 cycles of (30 sec at
516 94°C, 30 sec at 60°C and 7 min at 68°C). PCR products were visualized using a 2% agarose
517 E-Gel® (Invitrogen).
518

519 Processed pseudogene validation by PCR and capillary sequencing

520 To validate the 2 *de novo* PPG variants (Table 2) we used the following PCR
521 protocol: primers were designed using Primer3 to make products within the exons of each
522 gene. Forward and reverse primers were then paired between exons to amplify across the
523 excised intronic regions (Supplemental Table 2). PCR was carried out using either
524 Platinum™ Taq DNA Polymerase High Fidelity (Invitrogen) or Thermo-Start Taq DNA
525 Polymerase (Thermo Scientific). Platinum™ Taq assay: 20ng of genomic DNA extracted
526 from blood or saliva was amplified in the presence of 0.2 μM of each primer and 1 unit of
527 Platinum™ Taq. Amplification was carried out using the following cycling conditions; 2 min at
528 94°C, followed by 36 cycles of (30 sec at 94°C, 30 sec at 60°C and 1 min at 68°C). Thermo-
529 Start Taq DNA Polymerase assay: 40 ng genomic DNA was amplified in the presence of 0.2
530 μM of each primer and 0.42 units of Thermo-Start Taq. Cycling conditions were as follows: 5
531 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
532 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
533 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
534 72°C) with a final elongation of 10 min at 72°C. PCR products were visualized using a 2%
535 agarose E-Gel® (Invitrogen). PCR products were sequenced using either the forward or
536 reverse primer used in the amplification protocol by Eurofins GATC Biotech GmbH.

537 Sequence traces were aligned using SeqMan Pro 15 (Lasergene 15) and reads were
538 aligned to the human genome (hg19) using BLAT (UCSC)[49].
539

540 WGS of probands with *de novo* processed pseudogenes

541 To validate and determine the insertion site of the two identified *de novo* PPGs
542 (Table 2), we performed Illumina WGS on all individuals of each trio in which the *de novo*
543 event was identified (n = 6 individuals). Samples were first quantified with Biotium Accuclear
544 Ultra high sensitivity dsDNA Quantitative kit using Mosquito LV liquid platform, Bravo WS
545 and BMG FLUOstar Omega plate reader and cherrypicked to 500ng / 120ul using Tecan
546 liquid handling platform. Cherrypicked plates are then sheared to 450bp using a Covaris
547 LE220 instrument and subsequently purified using SPRI Select beads on Agilent Bravo WS.
548 Library construction (ER, A-tailing and ligation) was performed using 'NEB Ultra II custom kit'
549 on an Agilent Bravo WS automation system. Samples were then tagged using NextFLEX
550 Unique Dual Indexed adapter 1-96 barcodes at the ligation stage. Libraries were then
551 quantified by qPCR using Kapa Illumina ABI Sanger custom qPCR kits using a Mosquito LV
552 liquid handling platform, Bravo WS, and Roche Lightcycler. Libraries are then pooled in
553 equimolar amounts on a Beckman BioMek NX-8 liquid handling platform and normalised to
554 2.4nM for cluster generation on a c-BOT and then sequenced on the Illumina TenX
555 sequencing platform. Following sequencing, reads were aligned with BWA mem[50] (with
556 settings -t 16 -p -Y -K 100000000) to version hg19 of the human reference genome. Reads
557 were then manually inspected using the Integrative Genomics Viewer (IGV)[51] to confirm
558 presence, *de novo* status, and parent of origin of each PPG.
559

560 MEI mutation rate and burden

561 To determine the mutation rate independently for each MEI type (*Alu*, L1, SVA), we
562 utilized data generated by both DDD and the 1KGP[5]. For DDD data we filtered sites as
563 above based on our >10X coverage accessibility mask. For the 1KGP data[5], we created a

564 combined mask from three different data sources: 1.) the pilot accessibility mask generated
565 by the 1KGP project phase III[52], which removes regions of the genome inaccessible to
566 variant calling, 2.) reference ME sequences as identified by repeatmasker[53], as MELT is
567 unable to accurately ascertain MEIs in these regions, and 3.) All sequence ±10Kbp from the
568 5' and 3' terminus of all protein-coding genes from RefSeq[54]. This mask was generated
569 separately for *Alu* and L1 and did not filter 1,113.0Mbp or 959.9Mbp of the genome,
570 respectively. The *Alu* mask was used for filtering SVA and both masks excluded both
571 allosomes. On masking the 1KGP data, we were left with a total of 10,930 autosomal MEIs
572 (8,554 *Alu*, 2,047 L1, 329 SVA). Following filtering of the DDD and 1KGP sets with their
573 corresponding masks, we used the Watterson estimator with an effective population size of
574 10,000 for all calculations to estimate the population mutation parameter, Θ[34], and mutation
575 rate, μ (Supplemental Table 4).

576 We next used or estimate of μ to determine the expected number of *de novo* events
577 in exons, enhancers, and introns genome-wide. Total number of genome-wide mutations to
578 simulate, 686, was determined by extrapolation of μ for 9,738 individuals. Simulated variants
579 were then annotated identically to actual variants reported in this study. Total number of
580 variants in the three categories depicted in Fig. 4 were then summed to determine the
581 Poisson λ of *de novo* variants under neutral mutation rate and compared to number of
582 observed variants using the *ppois* function in R.

583

584 ## Author Contributions

585 E.J.G performed variant calling and annotation, PPG algorithm design, constraint and
586 burden testing, and initial clinical annotation and together with M.E.H. designed experiments,
587 oversaw the study, and wrote the manuscript. E. P. designed and performed PCR
588 experiments. G.G. curated and prepared DDD sequencing data. P.J.S. assisted in
589 estimating genetic burden of deleterious MEIs in the human population. A.S. assisted with
590 the design of the PPG discovery algorithm. T.S. performed variant calling of SNVs. K.E.C,

591 E.C., K.L.L., K.P., E.R., D.R.F, and H.V.F prepared clinical assessments of patients and
592 confirmation of molecular diagnoses as they relate to patient phenotype.
593

594 Acknowledgements

595 The authors wish to thank the Wellcome Sanger Institute sequencing facility staff for their
596 assistance in preparing samples and performing sequencing experiments, all members of
597 the DDD study for providing valuable comments during data analysis and manuscript
598 preparation, and the DDD families – this work would not be possible without their confidence
599 and support. We also thank Panayiotis Constantinou for helping to curate known MEI-
600 associated cases and for annotation of affected parents as well as Hilary Martin for
601 constructive comments during manuscript preparation. We also wish to acknowledge Jeffrey
602 Barrett and Caroline Wright for their leadership of the DDD. The DDD study presents
603 independent research commissioned by the Health Innovation Challenge Fund [grant
604 number HICF-1009-003], a parallel funding partnership between Wellcome and the
605 Department of Health, and the Wellcome Sanger Institute [grant number WT098051]. The
606 views expressed in this publication are those of the author(s) and not necessarily those of
607 Wellcome or the Department of Health. The study has UK Research Ethics Committee
608 approval (10/H0305/83, granted by the Cambridge South REC, and GEN/284/12 granted by
609 the Republic of Ireland REC). The research team acknowledges the support of the National
610 Institute for Health Research, through the Comprehensive Clinical Research Network. This
611 study makes use of DECIPHER (http://decipher.sanger.ac.uk), which is funded by the
612 Wellcome.
613

614 Competing Interests

615 M.E.H. is a co-founder of, consultant to, and holds shares in, Congenica Ltd, a genetics
616 diagnostic company.
617

618 # References

619 1. Mills, R.E., Bennett, E.A., Iskow, R.C. & Devine, S.E. Which transposable elements
620 are active in the human genome? *Trends Genet* **23**, 183-91 (2007).
621 2. Zhang, Z., Harrison, P.M., Liu, Y. & Gerstein, M. Millions of years of evolution
622 preserved: a comprehensive catalog of the processed pseudogenes in the human
623 genome. *Genome Res* **13**, 2541-58 (2003).
624 3. Stewart, C. *et al.* A comprehensive map of mobile element insertion polymorphisms
625 in humans. *PLoS Genet* **7**, e1002236 (2011).
626 4. Sudmant, P.H. *et al.* An integrated map of structural variation in 2,504 human
627 genomes. *Nature* **526**, 75-81 (2015).
628 5. Gardner, E.J. *et al.* The Mobile Element Locator Tool (MELT): population-scale
629 mobile element discovery and biology. *Genome Res* **27**, 1916-1929 (2017).
630 6. Zhang, Y., Li, S., Abyzov, A. & Gerstein, M.B. Landscape and variation of novel
631 retroduplications in 26 human populations. *PLoS Comput Biol* **13**, e1005567 (2017).
632 7. Hancks, D.C. & Kazazian, H.H., Jr. Roles for retrotransposon insertions in human
633 disease. *Mob DNA* **7**, 9 (2016).
634 8. Brandler, W.M. *et al.* Frequency and Complexity of De Novo Structural Mutation in
635 Autism. *Am J Hum Genet* **98**, 667-79 (2016).
636 9. Brandler, W.M. *et al.* Paternally inherited cis-regulatory structural variants are
637 associated with autism. *Science* **360**, 327-331 (2018).
638 10. Werling, D.M. *et al.* An analytical framework for whole-genome sequence association
639 studies and its implications for autism spectrum disorder. *Nat Genet* **50**, 727-736
640 (2018).
641 11. Hehir-Kwa, J.Y. *et al.* A high-quality human reference panel reveals the complexity
642 and distribution of genomic structural variants. *Nat Commun* **7**, 12989 (2016).
643 12. Deciphering Developmental Disorders Study. Prevalence and architecture of de novo
644 mutations in developmental disorders. *Nature* **542**, 433-438 (2017).

645 13. Martin, H.C. *et al.* Quantifying the contribution of recessive coding variation to
646 developmental disorders. *Science* (2018).
647 14. King, D.A. *et al.* Detection of structural mosaicism from targeted and whole-genome
648 sequencing data. *Genome Res* **27**, 1704-1714 (2017).
649 15. Short, P.J. *et al.* De novo mutations in regulatory elements in neurodevelopmental
650 disorders. *Nature* **555**, 611-616 (2018).
651 16. Lord, J. *et al.* The contribution of non-canonical splicing mutations to severe
652 dominant developmental disorders. *bioRxiv* (2018).
653 17. Kaplanis, J. *et al.* Mutational origins and pathogenic consequences of multinucleotide
654 mutations in 6,688 trios with developmental disorders. *bioRxiv* (2018).
655 18. Niemi, M.E.K. *et al.* Common genetic variants contribute to risk of rare severe
656 neurodevelopmental disorders. *Nature* **562**, 268-271 (2018).
657 19. Goncalves, I., Duret, L. & Mouchiroud, D. Nature and structure of human genes that
658 generate retropseudogenes. *Genome Res* **10**, 672-8 (2000).
659 20. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot
660 analysis: multitissue gene regulation in humans. *Science* **348**, 648-60 (2015).
661 21. Huang da, W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis
662 of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57 (2009).
663 22. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature*
664 **536**, 285 (2016).
665 23. Samocha, K.E. *et al.* A framework for the interpretation of de novo mutation in human
666 disease. *Nat Genet* **46**, 944-50 (2014).
667 24. Hormozdiari, F. *et al.* Rates and patterns of great ape retrotransposition. *Proc Natl*
668 *Acad Sci U S A* **110**, 13457-62 (2013).
669 25. Zhang, Y., Romanish, M.T. & Mager, D.L. Distributions of transposable elements
670 reveal hazardous zones in mammalian introns. *PLoS Comput Biol* **7**, e1002046
671 (2011).

672 26. MacArthur, D.G. *et al.* A systematic survey of loss-of-function variants in human
673 protein-coding genes. *Science* **335**, 823-8 (2012).
674 27. Kubiak, M.R. & Makalowska, I. Protein-Coding Genes' Retrocopies and Their
675 Functions. *Viruses* **9**(2017).
676 28. Gilbert, N., Lutz, S., Morrish, T.A. & Moran, J.V. Multiple fates of L1
677 retrotransposition intermediates in cultured human cells. *Mol Cell Biol* **25**, 7780-95
678 (2005).
679 29. Jonsson, H. *et al.* Parental influence on human germline de novo mutations in 1,548
680 trios from Iceland. *Nature* **549**, 519-522 (2017).
681 30. Firth, H.V. *et al.* DECIPHER: Database of Chromosomal Imbalance and Phenotype
682 in Humans Using Ensembl Resources. *Am J Hum Genet* **84**, 524-33 (2009).
683 31. Wright, C.F. *et al.* Making new genetic diagnoses with old data: iterative reanalysis
684 and reporting from genome-wide data in 1,133 families with developmental disorders.
685 *Genet Med* (2018).
686 32. Kersey, P.J. *et al.* Ensembl Genomes 2016: more genomes, more complexity.
687 *Nucleic Acids Res* **44**, D574-80 (2016).
688 33. Riecken, L.B. *et al.* Inhibition of RAS activation due to a homozygous ezrin variant in
689 patients with profound intellectual disability. *Hum Mutat* **36**, 270-8 (2015).
690 34. Watterson, G.A. On the number of segregating sites in genetical models without
691 recombination. *Theor Popul Biol* **7**, 256-76 (1975).
692 35. Ewing, A.D. & Kazazian, H.H., Jr. High-throughput sequencing reveals extensive
693 variation in human-specific L1 content in individual human genomes. *Genome Res*
694 **20**, 1262-70 (2010).
695 36. Torene, R.I. *et al.* Mobile element insertions in 28,00 clinical exomes (Pgmr 187).
696 *Presented at the Annual Meeting of The American Society of Human Genetics*
697 (2018).
698 37. Witherspoon, D.J. *et al.* Mobile element scanning (ME-Scan) identifies thousands of
699 novel Alu insertions in diverse human populations. *Genome Res* **23**, 1170-81 (2013).

700 38. Wimmer, K., Callens, T., Wernstedt, A. & Messiaen, L. The NF1 gene contains
701 hotspots for L1 endonuclease-dependent de novo insertion. *PLoS Genet* **7**,
702 e1002371 (2011).
703 39. Jurka, J. Sequence patterns indicate an enzymatic involvement in integration of
704 mammalian retroposons. *Proc Natl Acad Sci U S A* **94**, 1872-7 (1997).
705 40. Aneichyk, T. *et al.* Dissecting the Causal Mechanism of X-Linked Dystonia-
706 Parkinsonism by Integrating Genome and Transcriptome Assembly. *Cell* **172**, 897-
707 909.e21 (2018).
708 41. Wright, C.F. *et al.* Genetic diagnosis of developmental disorders in the DDD study: a
709 scalable analysis of genome-wide research data. *Lancet* **385**, 1305-14 (2015).
710 42. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-8
711 (2011).
712 43. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122
713 (2016).
714 44. Quinlan, A.R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr*
715 *Protoc Bioinformatics* **47**, 11.12.1-11.12.34 (2014).
716 45. May, D. *et al.* Large-scale discovery of enhancers from human heart tissue. *Nat*
717 *Genet* **44**, 89-93 (2011).
718 46. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L.A. VISTA Enhancer Browser--a
719 database of tissue-specific human enhancers. *Nucleic Acids Res* **35**, D88-92 (2007).
720 47. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and
721 yeast genomes. *Genome Res* **15**, 1034-50 (2005).
722 48. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for
723 analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-303 (2010).
724 49. Tyner, C. *et al.* The UCSC Genome Browser database: 2017 update. *Nucleic Acids*
725 *Res* (2016).
726 50. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler
727 transform. *Bioinformatics* **26**, 589-95 (2010).

51. Thorvaldsdottir, H., Robinson, J.T. & Mesirov, J.P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14**, 178-92 (2013).

52. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).

53. Smit AFA, H.R., Green P. RepeatMasker Open-3.0. (1996-2010).

54. O'Leary, N.A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733-45 (2016).