bioRxiv preprint doi: https://doi.org/10.1101/471375; this version posted November 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/471375; this version posted November 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

bioRxiv preprint doi: https://doi.org/10.1101/471375; this version posted November 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Main

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

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

bioRxiv preprint doi: https://doi.org/10.1101/471375; this version posted November 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

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

bioRxiv preprint doi: https://doi.org/10.1101/471375; this version posted November 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Results

Generation of a genome-wide dataset of RT variants

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

	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*

Table 1: RT variant discovery in the DDD

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

bioRxiv preprint doi: https://doi.org/10.1101/471375; this version posted November 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

As our study is the first to discover MEIs directly from WES on a large scale, we first utilized matched sample WGS data to determine if MELT could ascertain MEI variants reliably from WES data. We compared MEI variants identified by MELT within the DDD WES data to both WGS data generated on the same individuals and population MEI data previously generated from the 1000 Genomes Project Phase 3 (1KPG)[4-6] WGS data. The latter comparison was to ensure that the number of exonic MEIs identified within DDD WES data was concordant with expectations at the individual and population level. When comparing our WES genotypes to WGS in identical individuals, we had a genotype concordance rate of 94.46% (93.93% *Alu*, 97.29% L1, 98.25% SVA) among calls with at least 10X coverage in our WES data. In total, we were able to re-identify 1,355 (1,289 *Alu*, 160 L1, 1 SVA) MEI genotypes, or 84.5% of all heterozygous or homozygous genotypes identifiable with WGS in WES bait regions (Supplemental Table 1). Based on these findings we were confident that MELT was appropriately calibrated to ascertain MEIs in WES data.

We identified 1,129 MEI variants and 576 polymorphic PPGs, with each individual's exome containing on average 33.5±7.3 variants. All MEIs were genotyped across all individuals to form a comprehensive catalogue of RT-derived variation within and adjacent to (±50 bp) sequences targeted in the WES assay (Methods), including coding exons and targeted non-coding elements (Table 1; Fig. 1). The average time to assess a single family for RT-derived events was approximately 15 minutes and the rate of false findings was low (1 incorrect *de novo* variant per every 320 patients; either a false positive variant or false negative genotype in at least one parent). As expected, the total number of variants per individual for each RT class (Fig. 1a-d) as well as combined number of RT events (Fig. 1e) approximated a Poisson distribution. The vast majority of variants are rare (AF < $1\text{x}10^{-4}$; Fig. 1f), with >65% of *Alu* and L1 variants identified in fewer than 4 unrelated individuals. SVA and PPGs appear to be moderately under ascertained compared to *Alu* and L1 at lower AFs, with >50% of variants identified in the lowest AF bin. The length estimates for the three MEI classes largely fit the findings of previous studies (Fig. 1g)[4,5], except in the case of full-length L1 elements (i.e. L1s >6kbp in length). In our study, we identified a total of 26 full-length L1

bioRxiv preprint doi: https://doi.org/10.1101/471375; this version posted November 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

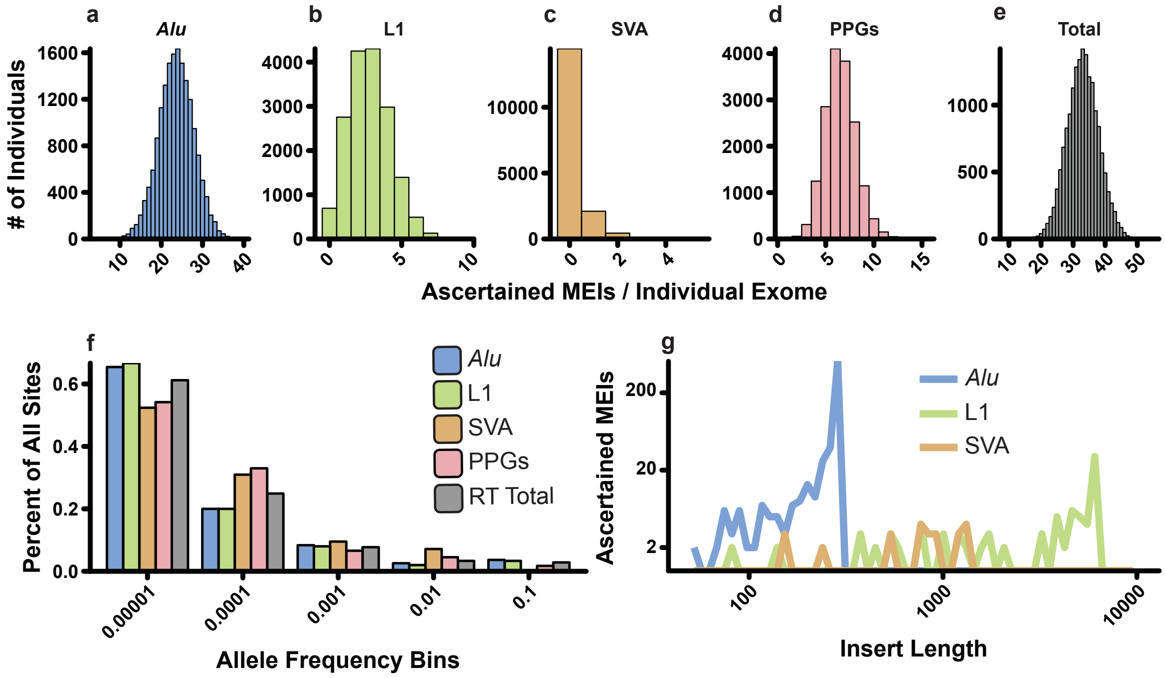

Figure 1: The DDD RT call set

(**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.

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

bioRxiv preprint doi: https://doi.org/10.1101/471375; this version posted November 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

Coding RT burden and constraint

As expected for WES, the vast majority (84.9%) of MEIs impacted the coding or intronic sequence of a protein-coding gene or a regulatory element targeted in the augmented WES assay described in Short et. al.[15] (Fig. 2a). While the number of MEIs identified in this study, based on the proportion of the genome assayed, represent only 2.2% of genome-wide MEI variants, we have ascertained over five-fold more variants that directly impact exons than the next largest study (Supplemental Fig. 5)[4,5].

Our large collection of coding variants allowed us to examine the evolutionary forces acting on coding MEI variation (Fig. 2b). To examine selective constraint, we utilized two

bioRxiv preprint doi: https://doi.org/10.1101/471375; this version posted November 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/471375; this version posted November 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

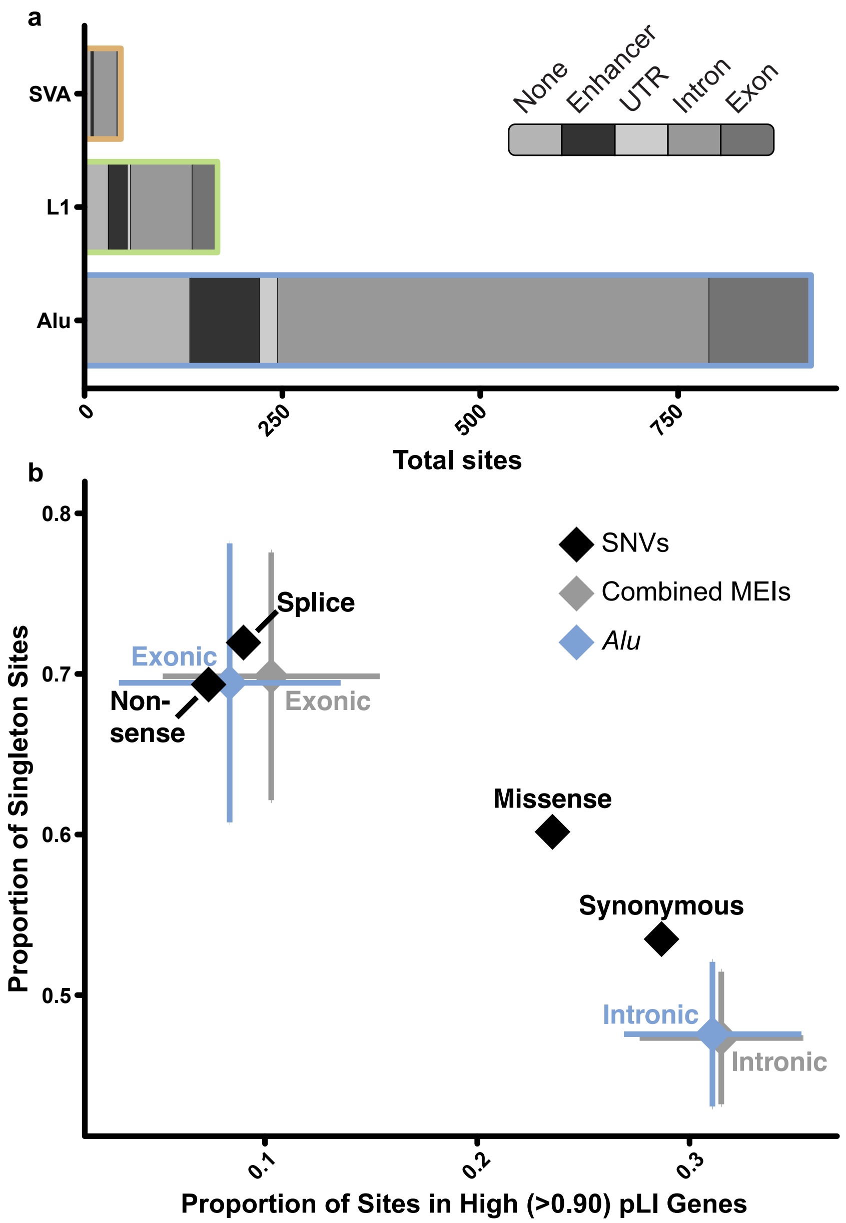

Figure 2: Coding constraint on MEIs

(**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.

bioRxiv preprint doi: https://doi.org/10.1101/471375; this version posted November 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/471375; this version posted November 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

Nine of our validated *de novo* mutations were MEIs (7 *Alu*, 2 L1), or a rate of approximately one *de novo* event per every 1,000 patient exomes sequenced (9/9,738). As expected, based on both the total number of polymorphisms[3-5] and mutation rate (Table 1; Supplemental Table 4), we identified more *Alu de novo* variants than the other RT classes. We also identified 2 PPG germ-line *de novo* variants, or approximately one new PPG per every 5,000 patient whole genomes sequenced (2/9,738). As a further quality control for PPGs, we capillary sequenced all resulting PCR products to confirm the gene of origin (Supplemental Data 1) and performed WGS to identify the PPG insertion site. We were able to localize the *SERINC5* PPG to an ~50Kbp intron of the gene *CLIC4* and the *SLC35F2* event to an intergenic region between the genes *MAK* and *GCM2* (Fig. 3e). Neither of the events directly impacted coding sequence and *CLIC4* is neither under strong selective 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 d*e 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).

bioRxiv preprint doi: https://doi.org/10.1101/471375; this version posted November 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/471375; this version posted November 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

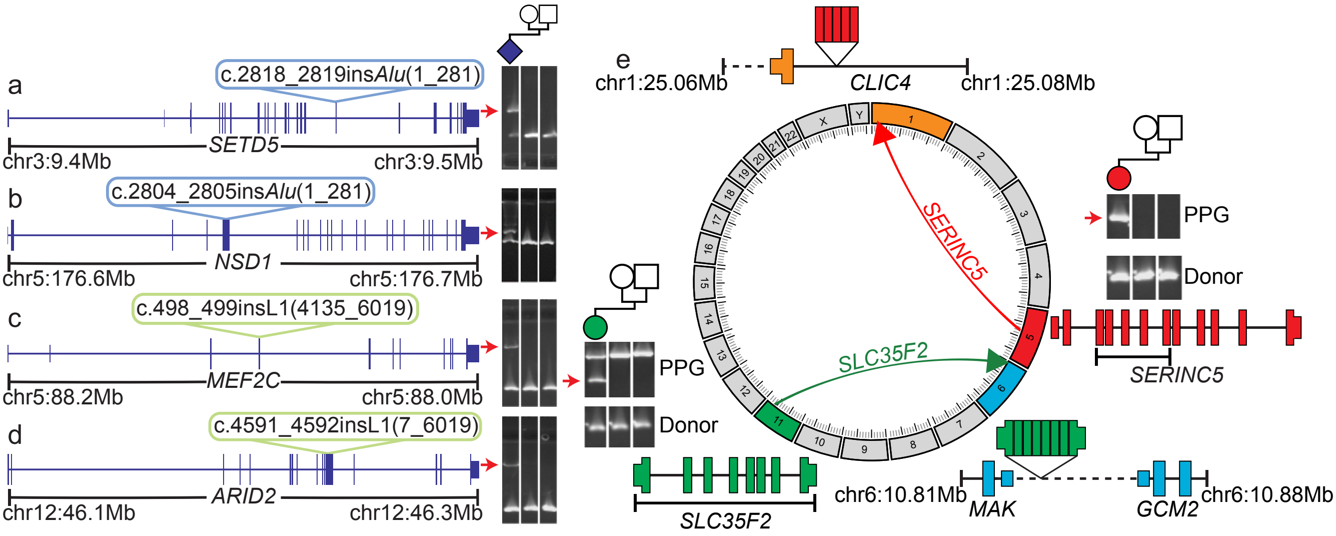

Figure 3: RT-derived de novos in the DDD

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.

bioRxiv preprint doi: https://doi.org/10.1101/471375; this version posted November 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

MEI mutation rate and enrichment of deleterious RT events in DDD

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

bioRxiv preprint doi: https://doi.org/10.1101/471375; this version posted November 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

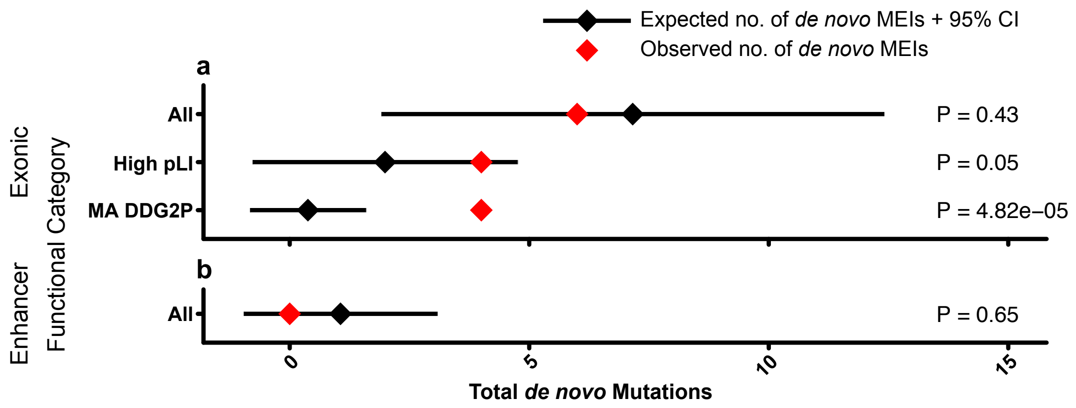

Figure 4: Estimating enrichment of deleterious MEIs

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 (1.2x10^{-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 = 4.82x10^{-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 = 4.3x10^{-5}).

bioRxiv preprint doi: https://doi.org/10.1101/471375; this version posted November 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Discussion

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/471375; this version posted November 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

4,500 neurofibromatosis type I patients (0.3% of patients). This rate is seven times higher than that observed in DDD or Torene et. al.[36] and was attributed to a potential RT mutation "hotspot" associated with the canonical L1 endonuclease cleavage site of 3'-AA/TTTT-5'[39] within the neurofibromatosis-associated gene, *NF1*. Further work is needed to investigate the role of sequence context in determining the overall genomic landscape of RT-mediated disease. Analogously, inclusion of sequence context into the SNV mutation model noticeably improved the ability to determine enrichment/depletion of deleterious SNVs within genes[22,23].

Our study is clearly limited in that we only identified ~2% of the RT variants in each individual human genome[4,5]. Despite a number of known disease-associated intronic MEIs in the literature, we did not identify a pathogenic intronic MEI. As such, it remains an open question as to what contribution RT mutations in the noncoding genome plays in the etiology of DD. While it appears that the contribution of regulatory elements to DD is relatively small, as defined by this (Fig. 4b) and other studies[15], previous work has identified a significant signature of purifying selection against MEI events within 100bp of exons[25] – variants which our study could potentially identify. As our data suggests that the majority of DD cases with pathogenic coding MEIs are due to *de novo* insertions (Table 2; Fig. 3), we conjecture that most additional DD-associated MEIs may be located in the introns of known DD-causing genes and disrupt splicing – a known disease mechanism attributable to RT-derived mutagenesis[7,38,40]. Simulations suggest that under a null genome-wide mutation model we should expect to observe 12.5 (5.5-19.4, 95% CI) *de novo* intronic RT mutations in dominant DD-associated genes in a population sample of 9,738 individuals. As such, a WGS study of a clinical population of similar size to that analyzed here should be well powered to estimate the pathogenic contribution of intronic MEIs.

De novo MEIs are typically readily interpretable with modest informatics expertise, and represent a clinically relevant class of variation to assay in clinical bioinformatics pipelines. While we ultimately find that the overall burden of RT-attributable disease is relatively low in the human population, it is nonetheless an important consideration when elucidating the genetic basis of DD in individual patients.

bioRxiv preprint doi: https://doi.org/10.1101/471375; this version posted November 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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,

bioRxiv preprint doi: https://doi.org/10.1101/471375; this version posted November 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

MEI call set generation and consequence annotation

To identify MEIs in the DDD WES data we utilized the previously published Mobile Element Locator Tool (MELT)[5]. MELT was run with default parameters (except the '-exome' flag during IndivAnalysis) using 'Split' mode to generate a final unified VCF-format file[42] of all 28,132 unfiltered individuals independently for each MEI type (*Alu*, L1, SVA). Following initial data set generation, we found that a subset of variants internal or adjacent to (±50bp) low complexity repeats (defined here as a run of sequence >= 15bp composed of two or fewer nucleotides) were likely false positive. As such, we added an additional filter to the final MELT VCF, lc (low complexity), which removes such false positives from downstream analysis. Variants that could not be genotyped in at least 25% of individuals, had ≤ 2 split reads, had MELT ASSESS score < 3, or had any value in the VCF FILTER column other than PASS or rSD were filtered.

To generate consequences plotted in Fig. 2a, all MEIs were annotated using Variant Effect Predictor v88 (VEP)[43] and intersected with bedtools intersect[44] to enhancers (one of heart[45], VISTA[46], or highly evolutionarily conserved[47]) included on the DDD WES capture[15]. Only a single consequence was retained for each variant, with priority given to enhancer annotation. Primary transcript as determined by VEP was used for all gene-based consequences, pLI score[22] annotation, and DDG2P disease association (Table 2).

bioRxiv preprint doi: https://doi.org/10.1101/471375; this version posted November 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Quality Control of RT data using WGS and 1KGP

To determine if our MEI WES call set was biased compared to WGS data, we performed two independent comparisons: 1.) to high coverage (>30x) WGS data generated for a subset of DDD trios and 2.) to a published collection of MEIs from 1KGP phase III[5].

For WGS quality-control, we used a subset of 30 DDD trios (n = 90 individuals) which were previously whole genome sequenced. MEI discovery using MELT[5] on all 90 individuals was performed and filtered identically to WES data. Genotypes identified in the WGS data but not in WES were then separated based on coverage in the corresponding WES. Genotypes in low coverage areas (<10x) were considered not possible (n.p), while variants where coverage was greater than 10x are considered not detected (n.d). All remaining genotypes were than compared for identity between WGS and WES results (Supplemental Table 1)

To compare the DDD MEI call set to the 1KGP, we first filtered 1KGP calls to variants with >10x coverage in 1,000 randomly sampled WES individuals (leaving 318 *Alu*, 81 L1, 26 SVA). We then randomly selected 2,453 DDD parents 1,000 times, retaining only loci present in downsampled individuals[4,5]. The resulting distribution was then compared to the observed number of variants in the 1KGP-masked data to generate z-scores independently for all three MEI types (Supplemental Fig. 2).

To compare our PPG dataset to Zhang et. al.[6], we downloaded provided supplemental tables. We then summed the total number of unique events per person and determined “allele counts” for each gene reported. Genes were then matched between our call set and Zhang et. al.[6] using ENSEMBL gene identifiers and allele counts between each data set were plotted to create Supplemental Fig. 3.

bioRxiv preprint doi: https://doi.org/10.1101/471375; this version posted November 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

GTEx annotation of processed pseudogenes

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

SNV Variant Calling and Quality Control

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

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

bioRxiv preprint doi: https://doi.org/10.1101/471375; this version posted November 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

SNV and MEI constraint

As sensitivity of variant discovery can bias our results, we generated an "accessibility mask" of the DDD WES data where we expect our variant ascertainment sensitivity to be >95% (Supplemental Fig. 8)[5]. Our mask thus includes only regions of the genome that contain at least 10X average coverage in a mean cohort of 1,000 randomly selected individuals for a total of 74.2Mbp, or ~2.3% of the genome (Supplemental Table 4). Using this mask, we filtered our original 1,129 variants down to 828 (660 *Alu*, 109 L1, 31 SVA) variants in unaffected parents (n = 17,032 individuals). Parents were determined to be affected either by the referring clinician or, where ambiguous, through manual curation of HPO terms for a matching parent-offspring phenotype.

Using this mask subset of variants, we determined genomic constraint as shown in Figure 2b. Allele frequency values were recalculated for all variants, and a pLI score[22] for each MEI was added as described above. MEIs which did not insert into a gene or inserted into a gene without a calculated pLI score[22] were excluded from subsequent analysis. We then calculated proportion of singleton variants and proportion of variants in high pLI genes independently for *Alu* and, due to low overall numbers of the other MEI subtypes, for a combined set of *Alu*, L1, and SVA. SNVs annotated as nonsense, missense, synonymous, or splice acceptor/donor (splice in Fig. 2b) as determined by VEP v88[43] were extracted from the SNV VCF files described above and used to calculate singleton and pLI proportion identically to MEIs.

Mobile element insertion validation by PCR

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

bioRxiv preprint doi: https://doi.org/10.1101/471375; this version posted November 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

Processed pseudogene validation by PCR and capillary sequencing

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

bioRxiv preprint doi: https://doi.org/10.1101/471375; this version posted November 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

WGS of probands with *de novo* processed pseudogenes

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

MEI mutation rate and burden

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

bioRxiv preprint doi: https://doi.org/10.1101/471375; this version posted November 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

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

Author Contributions

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

bioRxiv preprint doi: https://doi.org/10.1101/471375; this version posted November 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

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.

bioRxiv preprint doi: https://doi.org/10.1101/471375; this version posted November 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Mills, R.E., Bennett, E.A., Iskow, R.C. & Devine, S.E. Which transposable elements are active in the human genome? *Trends Genet* **23**, 183-91 (2007).

2. Zhang, Z., Harrison, P.M., Liu, Y. & Gerstein, M. Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. *Genome Res* **13**, 2541-58 (2003).

3. Stewart, C. *et al.* A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genet* **7**, e1002236 (2011).

4. Sudmant, P.H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75-81 (2015).

5. Gardner, E.J. *et al.* The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome Res* **27**, 1916-1929 (2017).

6. Zhang, Y., Li, S., Abyzov, A. & Gerstein, M.B. Landscape and variation of novel retroduplications in 26 human populations. *PLoS Comput Biol* **13**, e1005567 (2017).

7. Hancks, D.C. & Kazazian, H.H., Jr. Roles for retrotransposon insertions in human disease. *Mob DNA* **7**, 9 (2016).

8. Brandler, W.M. *et al.* Frequency and Complexity of De Novo Structural Mutation in Autism. *Am J Hum Genet* **98**, 667-79 (2016).

9. Brandler, W.M. *et al.* Paternally inherited cis-regulatory structural variants are associated with autism. *Science* **360**, 327-331 (2018).

10. Werling, D.M. *et al.* An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nat Genet* **50**, 727-736 (2018).

11. Hehir-Kwa, J.Y. *et al.* A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. *Nat Commun* **7**, 12989 (2016).

12. Deciphering Developmental Disorders Study. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**, 433-438 (2017).

bioRxiv preprint doi: https://doi.org/10.1101/471375; this version posted November 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

13. Martin, H.C. *et al.* Quantifying the contribution of recessive coding variation to developmental disorders. *Science* (2018).

14. King, D.A. *et al.* Detection of structural mosaicism from targeted and whole-genome sequencing data. *Genome Res* **27**, 1704-1714 (2017).

15. Short, P.J. *et al.* De novo mutations in regulatory elements in neurodevelopmental disorders. *Nature* **555**, 611-616 (2018).

16. Lord, J. *et al.* The contribution of non-canonical splicing mutations to severe dominant developmental disorders. *bioRxiv* (2018).

17. Kaplanis, J. *et al.* Mutational origins and pathogenic consequences of multinucleotide mutations in 6,688 trios with developmental disorders. *bioRxiv* (2018).

18. Niemi, M.E.K. *et al.* Common genetic variants contribute to risk of rare severe neurodevelopmental disorders. *Nature* **562**, 268-271 (2018).

19. Goncalves, I., Duret, L. & Mouchiroud, D. Nature and structure of human genes that generate retropseudogenes. *Genome Res* **10**, 672-8 (2000).

20. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648-60 (2015).

21. Huang da, W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57 (2009).

22. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285 (2016).

23. Samocha, K.E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat Genet* **46**, 944-50 (2014).

24. Hormozdiari, F. *et al.* Rates and patterns of great ape retrotransposition. *Proc Natl Acad Sci U S A* **110**, 13457-62 (2013).

25. Zhang, Y., Romanish, M.T. & Mager, D.L. Distributions of transposable elements reveal hazardous zones in mammalian introns. *PLoS Comput Biol* **7**, e1002046 (2011).

bioRxiv preprint doi: https://doi.org/10.1101/471375; this version posted November 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

26. MacArthur, D.G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823-8 (2012).

27. Kubiak, M.R. & Makalowska, I. Protein-Coding Genes' Retrocopies and Their Functions. *Viruses* **9**(2017).

28. Gilbert, N., Lutz, S., Morrish, T.A. & Moran, J.V. Multiple fates of L1 retrotransposition intermediates in cultured human cells. *Mol Cell Biol* **25**, 7780-95 (2005).

29. Jonsson, H. *et al.* Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549**, 519-522 (2017).

30. Firth, H.V. *et al.* DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am J Hum Genet* **84**, 524-33 (2009).

31. Wright, C.F. *et al.* Making new genetic diagnoses with old data: iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders. *Genet Med* (2018).

32. Kersey, P.J. *et al.* Ensembl Genomes 2016: more genomes, more complexity. *Nucleic Acids Res* **44**, D574-80 (2016).

33. Riecken, L.B. *et al.* Inhibition of RAS activation due to a homozygous ezrin variant in patients with profound intellectual disability. *Hum Mutat* **36**, 270-8 (2015).

34. Watterson, G.A. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol* **7**, 256-76 (1975).

35. Ewing, A.D. & Kazazian, H.H., Jr. High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. *Genome Res* **20**, 1262-70 (2010).

36. Torene, R.I. *et al.* Mobile element insertions in 28,00 clinical exomes (Pgmr 187). *Presented at the Annual Meeting of The American Society of Human Genetics* (2018).

37. Witherspoon, D.J. *et al.* Mobile element scanning (ME-Scan) identifies thousands of novel Alu insertions in diverse human populations. *Genome Res* **23**, 1170-81 (2013).

bioRxiv preprint doi: https://doi.org/10.1101/471375; this version posted November 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

38. Wimmer, K., Callens, T., Wernstedt, A. & Messiaen, L. The NF1 gene contains hotspots for L1 endonuclease-dependent de novo insertion. *PLoS Genet* **7**, e1002371 (2011).

39. Jurka, J. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc Natl Acad Sci U S A* **94**, 1872-7 (1997).

40. Aneichyk, T. *et al.* Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly. *Cell* **172**, 897-909.e21 (2018).

41. Wright, C.F. *et al.* Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet* **385**, 1305-14 (2015).

42. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-8 (2011).

43. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016).

44. Quinlan, A.R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr Protoc Bioinformatics* **47**, 11.12.1-11.12.34 (2014).

45. May, D. *et al.* Large-scale discovery of enhancers from human heart tissue. *Nat Genet* **44**, 89-93 (2011).

46. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L.A. VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic Acids Res* **35**, D88-92 (2007).

47. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034-50 (2005).

48. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-303 (2010).

49. Tyner, C. *et al.* The UCSC Genome Browser database: 2017 update. *Nucleic Acids Res* (2016).

50. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-95 (2010).

bioRxiv preprint doi: https://doi.org/10.1101/471375; this version posted November 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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).