InDelible: Detection and Evaluation of Clinically-relevant Structural Variation from Exome Sequencing

Eugene J. Gardner PhD[1,*], Alejandro Sifrim PhD[2,*], Sarah J. Lindsay PhD[1], Elena Prigmore PhD[1], Diana Rajan MSc[1], Petr Danecek PhD[1], Giuseppe Gallone PhD[1], Ruth Y. Eberhardt PhD[1], Hilary C. Martin PhD[1], Caroline F. Wright PhD[3], David R. FitzPatrick DM[3], Helen V. Firth DM[1,4], Matthew E. Hurles PhD[1,a]

[1]Wellcome Trust Sanger Institute, Wellcome Genome Campus, Cambridge, Hinxton, CB10 1SA, United Kingdom
[2]Department of Human Genetics, KU Leuven, Herestraat 49, Box 602, Leuven, Belgium, B-3000
[3]University of Exeter Medical School, Institute of Biomedical and Clinical Science, Royal Devon & Exeter Hospital, Exeter, EX2 5DW, United Kingdom
[4]MRC Human Genetics Unit, University of Edinburgh, WGH, Edinburgh, EH4 2SP, United Kingdom
[5]East Anglian Medical Genetics Service, Box 134, Cambridge University Hospitals NHS Foundation Trust, Cambridge Biomedical Campus, Cambridge CB2 0QQ, United Kingdom
[*]Contributed equally
[a]To whom correspondence should be addressed: meh@sanger.ac.uk, +44 (0)1223 495377

Abstract

Purpose
Identifying structural variations (SVs) associated with developmental disorder (DD) patient phenotype missed by conventional approaches.

Methods
We have developed a novel SV discovery approach that mines split-read information, 'InDelible', and applied it to exome sequencing (ES) of 13,438 probands with severe DD recruited as part of the Deciphering Developmental Disorders (DDD) study.

Results
Using InDelible we were able to find 59 previously undetected variants in genes previously associated with DD, of which 49.2% (29) had phenotypic features that accord with those of the patient in which they were found, and were deemed plausibly pathogenic. InDelible was particularly effective at ascertaining variants between 21-500 bps in size, and increased the total number of potentially pathogenic variants identified by DDD in this size range by 42.0% (n = 29 variants). Of particular interest were seven confirmed *de novo* SVs in the gene *MECP2*; these variants represent 31.8% of all *de novo* protein truncating variants in *MECP2* among DDD patients.

Conclusion
InDelible provides a rapid framework for the discovery of likely pathogenic SVs that are likely to be missed by standard analytical workflows and has the potential to improve the diagnostic yield of ES.

Introduction

Structural Variation (SV) is a broad class of genetic variation greater than 50bps in size[1]. SVs include a diverse collection of genomic rearrangements such as copy number variation (CNV), mobile element insertions (MEIs), inversions, translocations, and others[2]. Depending on population ancestry and technology used, the typical human genome harbours between 7,000-25,000 polymorphic SVs, with the majority constituting biallelic CNVs and MEIs[3]. While most SVs have minimal, if any, functional impact, SVs have been recognized as causative variants in congenital disorders[4–6].

In diagnostic testing of suspected genetic disorders, SVs are typically identified using chromosomal microarrays (CMAs) which offer a low-cost albeit low-resolution method for the identification of large CNVs (typically >20kbp in length for genic regions). CMAs are still widely used by diagnostic laboratories despite the increasing maturity of genome sequencing-based tools for SV discovery[7] and the wealth of clinically-ascertained exome sequencing (ES) data already generated for the ascertainment of single nucleotide variants (SNVs) and small insertions/deletions (InDels)[8]. There are several reasons for this. First, the cost, computational power, and informatics complexity necessary for genome sequencing-based diagnostics is still largely prohibitive to many public and private healthcare providers[9]. Second, current ES-based SV-discovery approaches focus on methods that interrogate sequencing coverage to identify regions of copy number variation within one genome compared to others[10]. As such, ascertainment is typically limited to CNVs of size >10kb, with resolution largely a factor of the sequencing depth and the density and number of baits in the ES assay, analogous to probes in CMAs. Thus, despite potentially offering improvements in CNV ascertainment over CMAs, ES as a tool for the assessment of diagnostic SVs has been slow to enter the clinic[11].

Consequently, patients with genetic abnormalities smaller than the discovery resolution of CMA or standard SV-ES approaches (>10kb) but larger than variants able to be accurately called using typical SNV/InDel based approaches (<50bp)[12] often remain undetected. To address this unmet need, we have developed the tool InDelible, which examines ES data for split read pairs indicative of SV breakpoints. We decided to focus on split reads because the formation of unique junction sequences is a shared characteristic of a broad range of different classes of SVs. We applied InDelible to ES data generated from 13,438 probands with severe developmental disorders (DD) recruited as part of the Deciphering Developmental Disorders (DDD) study. Approximately 29% of DDD probands harbour a pathogenic *de novo* mutation in a gene known to be associated with DD[8] and have been previously assessed for a wide range of potential genetic causes of DDs[13–17], which includes both CMA and ES-ascertained CNVs (unpublished). As such, the DDD study represents an ideal opportunity to demonstrate the additive diagnostic potential of identification of SVs at scale using split-read information.

Materials and Methods

InDelible Algorithm Development

InDelible processes a single ES sample provided in BAM or CRAM format via six primary steps (Figure 1). First, reads which have at least one split end (i.e. split reads; SRs) are identified (Fetch). Next, reads are aggregated to distinct clusters (Aggregate) and scored via random forest active learning (Score). SRs are then provided to BLAST[18] for breakpoint assessment (Blast) and gene(s) of intersect and cluster frequency are determined (Annotate). Finally, where possible, parental samples are assessed for split read evidence at identical loci (denovo). All of these commands can be run on one sample via the "Complete" command (blue box, Figure 1). We also provide the "Train" and "Database" commands to train a new random forest and combine identified breakpoints across multiple samples to generate an allele count database, respectively.

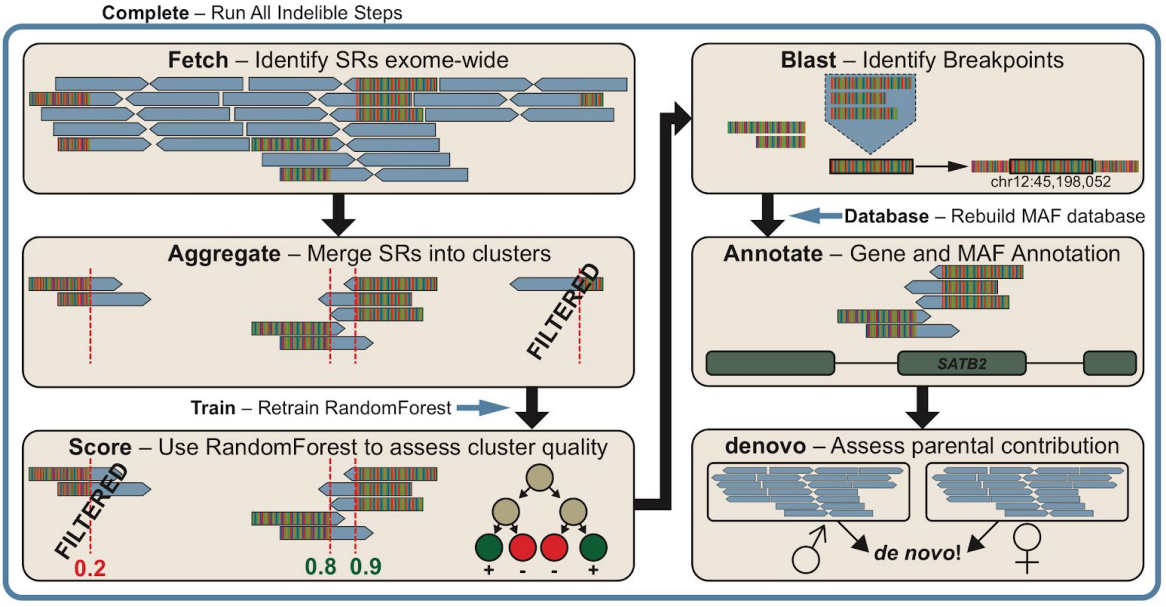

Figure 1: InDelible SV discovery in ES data. *InDelible processes one ES sample provided in BAM or CRAM format via six primary steps (tan boxes). First, alignment files are queried for all reads where part of the aligned sequence matches the reference genome and the other does not (i.e. split reads; Fetch). Next, reads are clustered (Aggregate) and scored using a random forest model[38] trained using a variant truth set (Score; see Methods for more detail). Split reads are then merged within clusters to determine the longest quality junction sequence and mapped back to the genome with blastn[18] to determine likely breakpoints (Blast). Split read clusters are subsequently annotated with population allele frequency and intersection with genomic functional annotations, such as protein-coding genes (Annotate). Finally, clusters are assessed for presence or absence in parental samples, where available, to determine inheritance status and identify likely de novo variants (denovo). All of these commands can be run on one sample via the "Complete" command (blue box). InDelible also includes the "Train" and "Database" commands to train a new random forest model and rebuild the allele frequency database, respectively.*

For variant quality control, InDelible uses a random forest paired with active learning to filter false positive variants using features calculated for each split read cluster (Supplementary Figure 1). The goal of active learning is to focus model training on the categorization of variants that are most difficult to classify. In short, active learning is an iterative approach in which test and training sets are selected from a pool of truth variants. Sites that are hard to

classify are then moved from the test set into the training set and the random forest model retrained iteratively until the improvement in prediction accuracy between rounds of retraining asymptotes. InDelible includes training data comprising variants ascertained from the DDD study and the resulting random forest model, but can be retrained with alternative data. InDelible is developed in version 3.7 of the Python programming language and is available free for academic use under v3.0 of the GNU GPL (see code availability).

A more detailed description of the InDelible algorithm and the random forest training methodology for variant filtering is given in the Supplementary Methods.

Patient Recruitment and InDelible Callset Generation

A total of 13,451 patients were recruited from 24 regional genetics services throughout the United Kingdom and Republic of Ireland as previously described[19]. Sequencing of families and alignment to the human reference genome (GRCh37) with bwa[20] was performed as previously described[19]. We applied InDelible to all 13,451 recruited probands, but excluded 13 probands from further analysis due to excessive runtime, leaving 13,438 probands. This includes probands sequenced with both parents (trios, n = 9,848) or with one or both parents absent (non-trios, n = 3,590). As this was the first dataset analyzed with InDelible, we ran all steps while also training the random forest and constructing the allele count database (Figure 1). Please see the supplementary methods for additional detail.

Quality Control of InDelible Ascertained Variants

Via visual inspection of read alignments[21], we determined whether variants were likely to be real in the proband and inherited from a parent where possible (Figure 2A). Variants with gnomAD non-Finnish European allele frequency $\geq 1 \times 10^{-4}$ based on Karczewski et al.[22] (variants ≤50bp) or Collins et al.[1] (variants >50bp) were then filtered out from further analysis (Figure 2A, Supplementary Data 1). This approach is imperfect – as the variant size range that InDelible detects is under-represented, some variants may be common in the population but may not be represented in these reference datasets. To determine the sensitivity of InDelible for DDD variants previously reported as potentially pathogenic, InDelible variants were intersected with previously reported DDD variants[23] and with CNVs called from read-depth analysis of DDD ES data (Supplementary Data 1). Previously known variants, false negative/positive variants, variants with high DDD/gnomAD allele frequency, and variants located in regions difficult to interpret clinically (e.g. intron or 5'/3' UTR) are annotated as such in Supplementary Data 1.

Data Availability

Sequencing, phenotype data, and variant calls for all data in this paper are accessible via the European Genome-phenome Archive (EGA) under study EGAS00001000775.

Code Availability

InDelible and R code used to generate data presented in this manuscript is available at the InDelible GitHub repository: https://github.com/eugenegardner/Indelible. A Docker image of InDelible is available at the following GitHub repository: https://github.com/wtsi-hgi/indelible-docker/tree/master.

Results

Overview of InDelible

To ascertain SVs and large InDels with breakpoints nearby or directly impacting ES bait regions, we designed the InDelible algorithm (summarised in Methods, detailed description in Supplementary Methods). InDelible variant discovery and analysis proceeds in several steps (Figure 1). In summary, Indelible identifies split reads, aggregates them into clusters at the same genomic location, filters these clusters to remove technical artefacts and retain likely genetic variants, and then combines unaligned portions of split reads and maps them to the genome to characterise the nature of the variant. InDelible also calculates the frequency of each split-read cluster across a population of individuals to facilitate the filtering of variants on the basis of minor allele frequency.

InDelible is coded in Python, uses the pysam[24] library for sequence alignment file manipulation (Supplementary Table 1), and works on bwa-aligned BAM or CRAM format files[25]. We have designed InDelible to be scalable for datasets comprising individual probands to multi-thousand sample cohorts and our estimates suggest that, to analyse a dataset of 1,000 trios, InDelible would require approximately 1610 CPU hours, or 16.1 hours of real time on a 100-core compute cluster (Supplementary Figure 2). Additionally, for easy implementation on cloud compute platforms, we have made InDelible available as a Docker image (see Supplementary Methods and Code Availability).

A key objective for the design of InDelible was to identify *de novo* variants potentially causative of a proband's disorder. As such, variants are primarily filtered on: (i) the population frequency of the split read cluster to remove variants too common to be plausibly causative of a rare disorder, (ii) absence in unaffected parents (when available), and (iii) intersection of variant breakpoints with the coding sequences of known-disease associated genes. Defining the precise molecular structure of SVs from short read sequencing data can be challenging, and even minor errors in breakpoint precision can have large consequences on interpretation (e.g. in versus out of frame InDels). Hence, we opted to identify all variants which intersect relevant DD-associated genes for further manual curation rather than relying on generic variant interpretation tools.

Identification of Putatively Causal SVs in DDD

To evaluate the utility of InDelible for diagnostic analyses, we applied InDelible to identify putatively diagnostic variants in 13,438 probands recruited to the DDD study. Probands were exome sequenced either with both parents (trios, n = 9,848) or with one or both parents absent (non-trios, n = 3,590). We first identified split reads and split read clusters (Figure 1) to ascertain 353,313,108 redundant split read clusters across all probands. Random forest filtering resulted in retention of 30,407,679 high-quality, redundant split read clusters across all probands, or 8.6% of originally ascertained loci (Supplementary Methods, Supplementary Figure 3). After cluster filtering we merged all retained clusters into a set of 1,947,186 non-redundant split read clusters across all 13,438 probands, with 1,346,856 (69.2%) clusters found only in one proband (Supplementary Figure 4). Retained clusters were then annotated with putative breakpoints, intersecting gene(s), and population frequency.

We next restricted our variant set to rare (call frequency < 0.04%) clusters found only in or near (here defined as within ±10bp of any exon) the coding sequence of 399 dominant or

X-linked DD-associated genes from the Developmental Disorders Genotype-to-Phenotype database (DDG2P)[26]. Variants identified within individuals sequenced as a parent-offspring trio were then also assessed for *de novo* status. Filtering on allele frequency, inheritance, and gene intersect resulted in a preliminary set of 244 candidate InDels and SVs across all 13,438 probands (Figure 2A; Supplementary Data 1; Supplementary Methods). Based on manual variant inspection[21], we determined that 16/244 (6.6%) candidate *de novo* events were likely to be present in a parent (i.e. parental false negatives) and 24/244 (9.84%) were unlikely to be real variants (i.e. offspring false positives). Four probands contributed 50% of false positive variants, indicating that sample selection and/or additional sample-level QC could further lower the false positive rate of InDelible (Figure 2A; Supplementary Data 1).

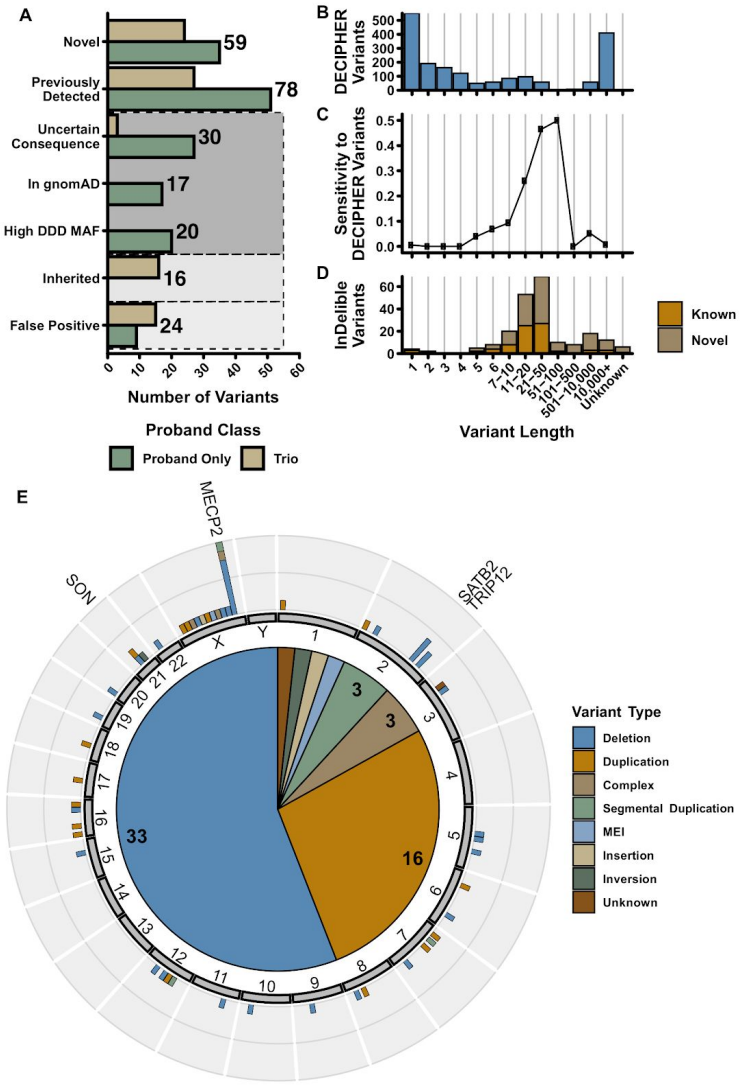

Figure 2: SV ascertainment in the DDD study with InDelible. *(**A**) Breakdown of putative variant consequences for all 244 variants identified in this study delineated by whether or not the proband was sequenced with both parents (Trio – Tan) or not (Proband Only – Dark Green). Grey boxes represent erroneous variants and variants unlikely to be associated with a proband phenotype. (**B**) Total number of previously reported variants among DDD probands with a net size change ≥1bp. (**C**) Sensitivity of InDelible to DDD reported variants among various variant size bins. (**D**) Categorization of InDelible ascertained variants into previously known (Tan) versus those novel (Orange) to this study based on size. (**E**) Distribution of variants unique to InDelible throughout the genome. Shown in the outer plot are the total number of InDelible variants per gene, with genes which have multiple previously undetected variants labeled. In the inner plot, we have displayed the total number of variants for each SV type we identified.*

Following variant quality control, we sought to further curate variants for those likely to be associated with a proband phenotype (Figure 2A). We considered variants with a minor allele frequency in the genome aggregation database (gnomAD)[1,22] of non-Finnish European exomes of ≥1x10^{-4} (17/244; 7.0%) or presence in other unrelated individuals within DDD (20/244; 8.2%) as unlikely to be the cause of the child's disorder. Additionally, variants confined to introns or 5'/3' UTRs were also defined as variants of uncertain significance and were not considered further (30/244; 12.3%). This final round of filtering left 137 SVs and large InDels which could plausibly explain a proband phenotype (51 from probands sequenced as trios, 86 from non-trio probands).

We next sought to determine the sensitivity of InDelible to clinically relevant variants ascertained using alternative methods. DDD has already identified (across both trio and non-trio probands) 1852 rare, plausibly pathogenic variants with a net size difference ≥1bp (i.e. non-SNVs) in the same DDG2P gene set defined above[4,8] – variants potentially detectable with a split read-based method such as that employed by InDelible. The majority of these variants are private or low allele frequency small InDels between 1-10bp in size (1218/1852; 65.8%) or large CMA or ES-ascertained CNVs ≥10kb in length (410/1852; 22.1%; Figure 2B). As anticipated due to the low number of split reads at variant breakpoints as variant size decreases, InDelible performed poorly in identification of very short variants ≤10bp with an overall sensitivity of 1.4% (Figure 2C). Sensitivity improved as a function of variant size, peaking at 50.0% sensitivity for variants between 51-100bp, but dropped again for variants ≥100bp; as variant size increases, it becomes more likely that the breakpoints of SVs which impact coding sequence lie outside of ES target regions (i.e. within intronic and intergenic sequences). Ergo, such variants are refractory to identification with split reads and likely to be missed by InDelible. Overall, InDelible reidentified a total of 78 variants which had been previously ascertained by at least one other variant discovery algorithm in the DDD study, and identified an additional 59 putatively causal variants that had been missed by other methods.

Clinical Interpretation of Variants Ascertained with InDelible

These 59 previously undetected variants (four of which have been described previously[27]) that impact known DD-associated genes (Supplementary Data 1) is composed primarily of deletions and duplications (49/59; 83%) but also includes variants with diverse mutational mechanisms such as MEIs, complex rearrangements, and dispersed duplications/translocations (Figure 2E). Twenty four of these variants were observed in trio probands, with parental data supporting a *de novo* origin for all of these variants. InDelible was particularly effective at identifying variants between 21-500bps in size (Figure 2D) – 29 previously undetected variants (49% of InDelible-specific variants) lie within this size range and represent a 42% increase in putatively pathogenic variants 21-500bps in length among DDD probands (Figure 2D). We also identified four genes with multiple previously undetected SVs among unrelated individuals, of which the most recurrently affected was *MECP2,* the causal gene of Rett Syndrome (Figure 2E)[28].

From an initial round of clinical review, based on intersecting gene(s) and associated phenotypes, we concluded that six (10.2%) of these 59 previously undetected variants were unlikely to explain the referred proband's phenotype, and were thus excluded from future analysis (Supplementary Data 1). We next attempted PCR validation of all 53 putatively pathogenic variants (Supplementary Methods). Of the variants for which conclusive validation results could be obtained, 25/25 (100%) were confirmed as true positives, either by the obvious presence of a mutant band of expected size with gel electrophoresis or by

follow-up capillary sequencing where the gel result was uncertain (Supplementary Data 1). For variants for which PCR was possible, we also confirmed that 13/13 (100%) putative *de novo* variants identified in trio probands were indeed absent from both parents.

All 53 plausibly pathogenic variants were then clinically interpreted by two senior clinical geneticists; 29/53 (54.7%) were classified as pathogenic or likely pathogenic by both clinical geneticists (Supplementary Data 1). *De novo* variants identified by Indelible thus represent 0.8% (20/2592) of all confirmed diagnoses among trio probands in the DDD study.

Novel variants identified in non-trio probands (n = 29 variants), for which inheritance status is unavailable, were less likely to be considered pathogenic (Fisher's $p = 2.0 \times 10^{-3}$). This finding is corroborated by the difference in the proportion of in-frame versus out-of-frame deletions and duplications ≤50bp between trio and non-trio probands; 78.2% of deletions and duplications are in-frame for non-trios versus 10.5% for trios (Fisher's $p = 2.5 \times 10^{-7}$; Supplementary Figure 5). This is consistent with population-level observations – out-of-frame deletions and duplications are typically under stronger negative selection than in-frame variants[29] and an increased proportion of in-frame variants in non-trio probands is suggestive of a greater proportion being benign. The difference is likely attributable to the absence of parental data leading to the inclusion of rare benign inherited variants that are unlikely to be filtered out using population variation data (e.g. gnomAD[1,22]).

InDelible enables assessment of a *de novo* SV Mutational Hotspot in *MECP2*

InDelible identified a total of seven confirmed *de novo* variants ≥20bp in length affecting *MECP2* (Figure 2E; Figure 3A), all predicted to be protein truncating, and all ascertained from female probands. Out of these seven probands, two have phenotypes that could be described as consistent with typical Rett Syndrome presentation[30]. Through in-depth clinical curation of HPO terms (see Supplementary Methods), we grouped probands with putative loss of function mutations caused by SVs in *MECP2* into four categories (Figure 3B). Cases identified by InDelible thus represent the wide variety of diverse clinical presentations that can result from disruption of the C-terminus of *MECP2*[31] and include previously observed *MECP2*-associated phenotypes such as early onset seizures and Angelman-like symptoms (Supplementary Data 2; Figure 3B)[32].

Interestingly, all five of our *MECP2* variants in probands without typical Rett syndrome presentation overlapped the same 634bp region located within the final coding exon and, aside from a previously ascertained whole gene deletion (proband 279220), do not overlap with putatively pathogenic SNVs identified within the DDD study (Figure 3A). The SV-specific region corresponds to an area of low sequence complexity and has been previously ascertained as hyper-mutable by several studies[31,33]. The molecular function of this region of *MECP2* is poorly understood and it is uncertain as to the consequences that our described variants may have on protein structure beyond decreasing transcript abundance and/or overall protein stability[31].

The seven *de novo MECP2* variants constitute 29% (7/24) of all novel *de novo* variants identified by InDelible, and 35% (7/20) of all confirmed *de novo* protein-truncating or gene-deleting variants of *MECP2* in the DDD study[8] (Figure 3A).

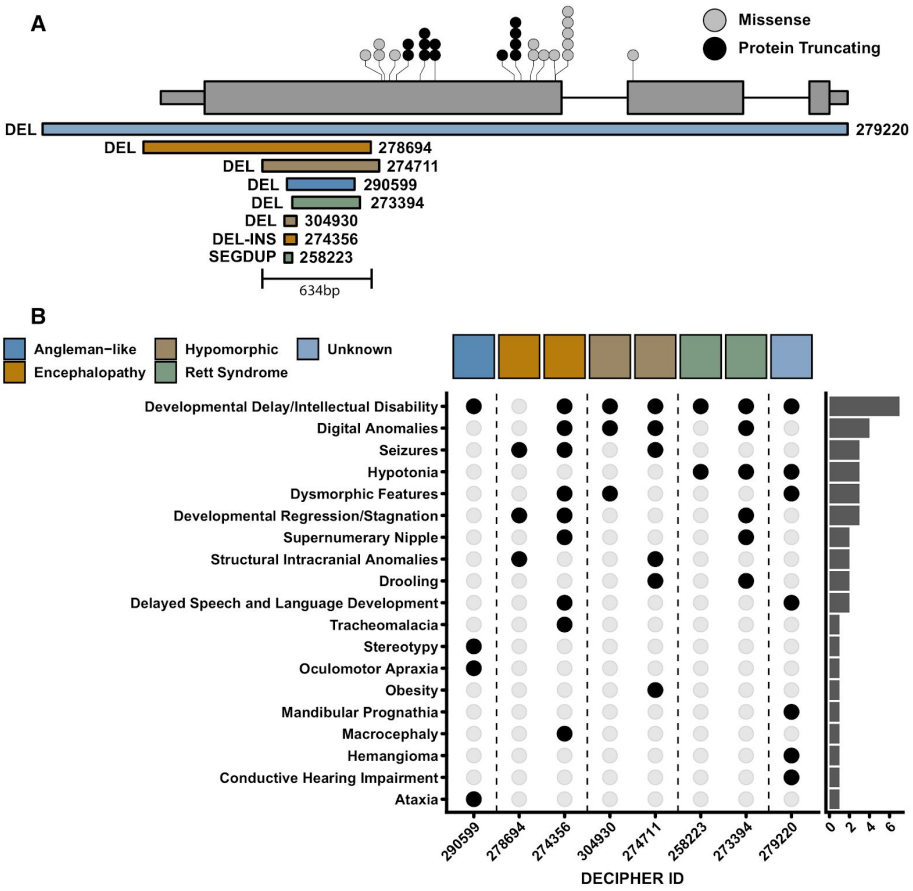

Figure 3: Clustered SVs in MECP2 cause diverse phenotypes. *(**A**) Shown is a cartoon representation of the gene MECP2, with stop-gained (black circles) and missense (grey circles) de novo SNVs identified in DDD trios. Each circle represents one proband, with recurrent variants represented by stacks of circles. Below the MECP2 gene model, we have shown the seven variants identified by InDelible as well as the single whole gene deletion previously identified via CMA (proband 279220). All InDelible-ascertained variants overlap the same 634bp region in the last exon of MECP2. (**B**) Diverse proband phenotypes among MECP2 SV carriers. Each proband carrying a MECP2 SV from (A) is shown on the x-axis, with phenotypes annotated by the referring clinician shown on the y-axis. Filled black circles represent when a corresponding proband displays the corresponding phenotype. Colored boxes on the top of the plot represent the diverse phenotypes we identified following clinical review. The y-axis marginal histogram represents the number of times the corresponding phenotype was observed among our SV probands.*

Discussion

Here we present the development and application of InDelible, a novel tool designed for the rapid assessment of ES data for breakpoints of rare, pathogenic SVs involved in single gene disorders (Figure 1). We applied InDelible to 13,438 proband genomes sequenced as part of the DDD study and identified a total of 137 candidate pathogenic variants impacting genes associated with dominant or X-linked DD (Figure 2A, Supplementary Data 1). Of these 137 variants, 59 were not previously identified in DDD probands, despite the wide range of SV and InDel detection algorithms that have previously been deployed on this cohort[8,27]. Notably, we increased the number of putatively diagnostic variants among DDD probands 21-500bp in length by 42.0% (Figure 2D). Through conservative clinical assessment of these 59 variants, we determined that 29 (49.2%) of our previously undetected variants were considered likely causative of proband phenotype – of particular interest was the large number of protein truncating SVs we identified in *MECP2* (Figure 3).

The variant size range which InDelible interrogates is complementary to other approaches commonly used for variant discovery from ES data[10,12]. While other previously described algorithms have also attempted to mine split read information for structural variant detection[12,34,35], they have different properties that preclude meaningful comparison with InDelible[12]. Some have been trained primarily on genome sequencing data rather than ES data[34,35], others do not explicitly assess *de novo* status, and many are not readily scalable to a dataset of ~10,000 trios. As such, we have built InDelible to be scalable to many thousands of samples (Supplementary Figure 2).

Other studies have previously noted that ~10% of all *MECP2* variants in probands ascertained based on presentation of Rett-associated phenotypes were deletions[33,36] and a large number of pathogenic or likely pathogenic variants in ClinVar fall within the same region of *MECP2* that we report in this manuscript. These observations, combined with the diverse phenotypes that this study has identified (Figure 3B), further complicate the clinical interpretation of variants disrupting *MECP2*. In particular, the work of Guy et al.[31] found that slight differences between the size and sequence context of deletions in the C-terminal domain of *MECP2* can have significant ramifications in RNA/protein expression. Additionally, Huppke et al.[37] found that skewed X-inactivation could play a role in the severity of *MECP2* presentation. Further work is needed to understand how different classes of mutation lead to diverse phenotypes in patients with *MECP2* loss of function variants. However, most importantly and exemplifying the additive power of InDelible, if not applied to the DDD study, 20.6% of DDD probands with clinically relevant *MECP2* variants would not have received a diagnosis for their disorder.

InDelible was designed to detect variant breakpoints missed by other approaches in ES data from DD patients. This has three major ramifications for the design of InDelible and the variants reported as part of this study. Firstly, as the primary cause of DD are highly penetrant dominant *de novo* variants[8], InDelible variant reporting was focused on identifying such variants from a defined list of genes known to be associated with DD[26]. However, this does not preclude the use of InDelible to identify variants acting through other modes of inheritance – InDelible will identify variants across the entire allele frequency spectrum and outside of the provided gene list as part of the primary output.

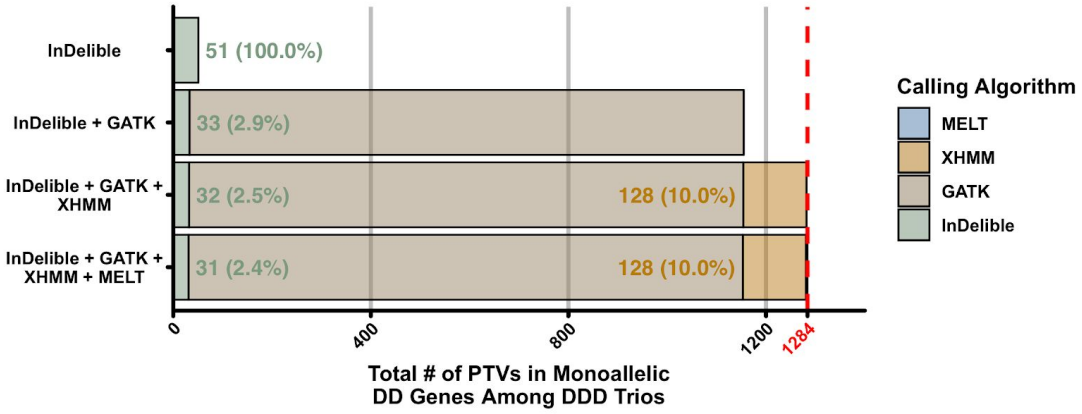

Figure 4: Added diagnostic PTV Yield of InDelible. *Total number of de novo PTVs (y-axis) ascertained in DD-associated genes when using InDelible alone, or in combination with a subset of three additional algorithms (GATK[12], XHMM[10], or MELT[4]). Percentages represent the proportion of all PTVs specific to InDelible (green text) or XHMM (orange text) for each bar. The red line and axis label indicates the maximum number of de novo PTVs identified in DD-associated genes among 9,848 DDD trio probands if combining data from all four algorithms (n = 1,284 variants).*

Secondly, the DDD cohort has been previously investigated for a broader range of variant classes (using both different assays and algorithms) than most ES studies. For ES-based CNV discovery from read-depth, DDD applied four separate algorithms to build a joint call set (unpublished). Thus, the added diagnostic value of running InDelible is probably under-estimated in the DDD study compared to other ES studies. To quantify the added diagnostic value of running InDelible across different settings, we estimated the proportion of unique PTVs InDelible would find if used alone or jointly with other algorithms targeting a breadth of variant types (SNVs, InDels, large deletions, and MEIs; Supplementary Methods)[4,10,12]. Overall, and when using other approaches, InDelible-specific variants will likely represent between 2-3% of all PTVs in a given cohort (Figure 4).

Finally, InDelible is unlikely to be more effective than currently available tools when applied to genome sequencing data. In ES, discordant read pairs are typically much less informative for detecting SVs than in genome sequencing due to the inherent properties of the data. In genome sequencing, data combining split and discordant read-pair information is a better means to identify most SV types.

InDelible provides a rapid framework for the assessment of ES data for intermediate length pathogenic SVs of diverse mutational origins. Our results show that through a combination of improved algorithm design, variant annotation, and clinical interpretation, ongoing interrogation of well-studied datasets will continue to yield novel diagnoses.

Acknowledgements

We would like to thank the DDD participants and their families – without their trust and confidence this work would not be possible. We also wish to acknowledge Jeffrey Barrett for his leadership role in the DDD Study. 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 [decipher.sanger.ac.uk]]), which is funded by the Wellcome.

Author Information

A.S. performed initial design and implementation of the InDelible algorithm. E.J.G. provided additional code, prepared InDelible for publication, assessed the DDD cohort for variation with InDelible, and together with A.S. and M.E.H. designed experiments, oversaw the study, and wrote the manuscript. S.J.L, E.P., and D.R. designed and performed PCR validation experiments of InDelible-ascertained variants. G.G. and R.Y.E curated and maintained DDD sequencing data. P.D. curated ES CNV calls. H.C.M. helped curate DECIPHER variants and DDD patient diagnoses. C.F.W, D.R.F, and H.V.F performed clinical interpretation and helped write the manuscript.

Ethics Statement

M.E.H. is a co-founder of, consultant to, and holds shares in, Congenica Ltd, a genetics diagnostic company. Informed and written 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 REC, and GEN/284/12 granted by the Republic of Ireland REC). 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.

References

1. Collins RL, Brand H, Karczewski KJ, et al. A structural variation reference for medical and population genetics. *Nature*. 2020;581(7809):444-451.

2. Sudmant PH, Rausch T, Gardner EJ, et al. An integrated map of structural variation in 2,504 human genomes. *Nature*. 2015;526(7571):75-81.

3. Zhao X, Collins RL, Lee W-P, et al. Expectations and blind spots for structural variation detection from short-read alignment and long-read assembly. *Genomics*. Published online July 4, 2020.

4. Gardner EJ, Prigmore E, Gallone G, et al. Contribution of retrotransposition to developmental disorders. *Nat Commun*. 2019;10(1):4630.

5. Sanchis-Juan A, Stephens J, French CE, et al. Complex structural variants in Mendelian disorders: identification and breakpoint resolution using short- and long-read genome sequencing. *Genome Medicine*. 2018;10(1). doi:10.1186/s13073-018-0606-6

6. Torene RI, Galens K, Liu S, et al. Mobile element insertion detection in 89,874 clinical exomes. *Genet Med*. 2020;22(5):974-978.

7. Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol*. 2019;20(1):117.

8. Kaplanis J, Samocha KE, Wiel L, et al. Integrating healthcare and research genetic data empowers the discovery of 28 novel developmental disorders. *Genomics*. Published online October 16, 2019:584.

9. Schwarze K, Buchanan J, Taylor JC, Wordsworth S. Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. *Genet Med*. 2018;20(10):1122-1130.

10. Fromer M, Moran JL, Chambert K, et al. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am J Hum Genet*. 2012;91(4):597-607.

11. Srivastava S, Love-Nichols JA, Dies KA, et al. Meta-analysis and multidisciplinary consensus statement: exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental disorders. *Genet Med*. 2019;21(11):2413-2421.

12. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297-1303.

13. Kaplanis J, Akawi N, Gallone G, et al. Exome-wide assessment of the functional impact

and pathogenicity of multinucleotide mutations. *Genome Res*. 2019;29(7):1047-1056.

14. Short PJ, McRae JF, Gallone G, et al. De novo mutations in regulatory elements in neurodevelopmental disorders. *Nature*. 2018;555(7698):611-616.

15. Lord J, Gallone G, Short PJ, et al. Pathogenicity and selective constraint on variation near splice sites. *Genome Res*. 2019;29(2):159-170.

16. Niemi MEK, Martin HC, Rice DL, et al. Common genetic variants contribute to risk of rare severe neurodevelopmental disorders. *Nature*. 2018;562(7726):268-271.

17. Martin HC, Jones WD, McIntyre R, et al. Quantifying the contribution of recessive coding variation to developmental disorders. *Science*. 2018;362(6419):1161-1164.

18. Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. *BMC Bioinformatics*. 2009;10:421.

19. Deciphering Developmental Disorders Study. Prevalence and architecture of de novo mutations in developmental disorders. *Nature*. 2017;542(7642):433-438.

20. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2010;26(5):589-595.

21. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform*. 2013;14(2):178-192.

22. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020;581(7809):434-443.

23. Firth HV, Richards SM, Bevan AP, et al. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am J Hum Genet*. 2009;84(4):524-533.

24. Heger A, Jacobs K. *Pysam: Htslib Interface for Python*.; 2009. https://github.com/pysam-developers/pysam

25. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079.

26. Thormann A, Halachev M, McLaren W, et al. Flexible and scalable diagnostic filtering of genomic variants using G2P with Ensembl VEP. *Nat Commun*. 2019;10(1):2373.

27. Wright CF, McRae JF, Clayton S, 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;20(10):1216-1223.

28. Amir RE, Van den Veyver IB, Wan M, Tran CQ, Francke U, Zoghbi HY. Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nature Genetics*. 1999;23(2):185-188. doi:10.1038/13810

29. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in

60,706 humans. *Nature*. 2016;536(7616):285-291.

30. Neul JL, Kaufmann WE, Glaze DG, et al. Rett syndrome: revised diagnostic criteria and nomenclature. *Ann Neurol*. 2010;68(6):944-950.

31. Guy J, Alexander-Howden B, FitzPatrick L, et al. A mutation-led search for novel functional domains in MeCP2. *Hum Mol Genet*. 2018;27(14):2531-2545.

32. Watson P, Black G, Ramsden S, et al. Angelman syndrome phenotype associated with mutations in MECP2, a gene encoding a methyl CpG binding protein. *J Med Genet*. 2001;38(4):224-228.

33. Bebbington A, Percy A, Christodoulou J, et al. Updating the profile of C-terminal MECP2 deletions in Rett syndrome. *J Med Genet*. 2010;47(4):242-248.

34. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*. 2009;25(21):2865-2871.

35. Chen X, Schulz-Trieglaff O, Shaw R, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*. 2016;32(8):1220-1222.

36. Krishnaraj R, Ho G, Christodoulou J. RettBASE: Rett syndrome database update. *Hum Mutat*. 2017;38(8):922-931.

37. Huppke P, Maier EM, Warnke A, Brendel C, Laccone F, Gärtner J. Very mild cases of Rett syndrome with skewed X inactivation. *J Med Genet*. 2006;43(10):814-816.

38. Liaw A, Wiener M. Classification and Regression by randomForest. *R news*. 2002;2:18-22.