











OPEN

DATA DESCRIPTOR

A catalogue of 863 Rett-syndrome-causing *MECP2* mutations and lessons learned from data integration

Friederike Ehrhart ^{1,2}✉, Annika Jacobsen ³, Maria Rigau ⁴, Mattia Bosio⁴, Rajaram Kaliyaperumal³, Jeroen F. J. Laros³, Egon L. Willighagen ¹, Alfonso Valencia ^{4,5}, Marco Roos ³, Salvador Capella-Gutierrez ⁴, Leopold M. G. Curfs ² & Chris T. Evelo ^{1,2}

Rett syndrome (RTT) is a rare neurological disorder mostly caused by a genetic variation in *MECP2*. Making new *MECP2* variants and the related phenotypes available provides data for better understanding of disease mechanisms and faster identification of variants for diagnosis. This is, however, currently hampered by the lack of interoperability between genotype-phenotype databases. Here, we demonstrate on the example of *MECP2* in RTT that by making the genotype-phenotype data more Findable, Accessible, Interoperable, and Reusable (FAIR), we can facilitate prioritization and analysis of variants. In total, 10,968 *MECP2* variants were successfully integrated. Among these variants 863 unique confirmed RTT causing and 209 unique confirmed benign variants were found. This dataset was used for comparison of pathogenicity predicting tools, protein consequences, and identification of ambiguous variants. Prediction tools generally recognised the RTT causing and benign variants, however, there was a broad range of overlap. Nineteen variants were identified that were annotated as both disease-causing and benign, suggesting that there are additional factors in these cases contributing to disease development.

Background and Summary

Rett syndrome (RTT) is a rare neurological disorder first described in 1956 by Andreas Rett occurring predominantly in females¹. In most cases, the disorder is caused by a loss-of-function variation on the X-bound gene for *MECP2* (methyl-CpG-binding protein 2)^{2,3}. Phenotypic severity is thought to vary due to X-inactivation, mosaicism, severity of the variation (loss of function vs. impaired function), genetic background (ref. ⁴ and literature cited therein) and environmental factors.

On the molecular level, the *MECP2* protein recognizes and binds to specific methylated and hydroxymethylated DNA regions, and attracts several other proteins to form a transcription repression block. This block makes the DNA sequence accessible for histone deacetylases, which increases the packing density of these regions, reducing their transcriptional activity⁵. Several metastudies on omics data revealed that the influence of *MECP2* affects dominantly dendritic connectivity, synapse function, glial cell differentiation, mitochondrial function, mRNA processing and translation, inflammation, and cytoskeleton⁶⁻⁸.

The *MECP2* protein has five different domains: N-terminal domain (NTD), methyl-DNA binding domain (MDB), transcription repressor binding domain (TRD), intermediate domain between methyl-DNA binding and transcription repressor binding domain also called interdomain (ID), C-terminal domain (CTD)⁹. Ballestar and coworkers found that *MECP2* variations that slightly decrease the specific recognition of the binding site on DNA are able to cause RTT¹⁰. The majority of RTT causing missense variations are found in the methyl-DNA binding

¹Department of Bioinformatics - BiGCaT, NUTRIM School of Nutrition and Translational Research in Metabolism, MHeNS School of Mental Health and Neuroscience, Maastricht University, Maastricht, The Netherlands. ²GKC - Rett Expertise Centre, Maastricht University Medical Center, Maastricht, The Netherlands. ³Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands. ⁴Barcelona Supercomputing Centre (BSC), Barcelona, Spain. ⁵Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain. ✉e-mail: friederike.ehrhart@maastrichtuniversity.nl

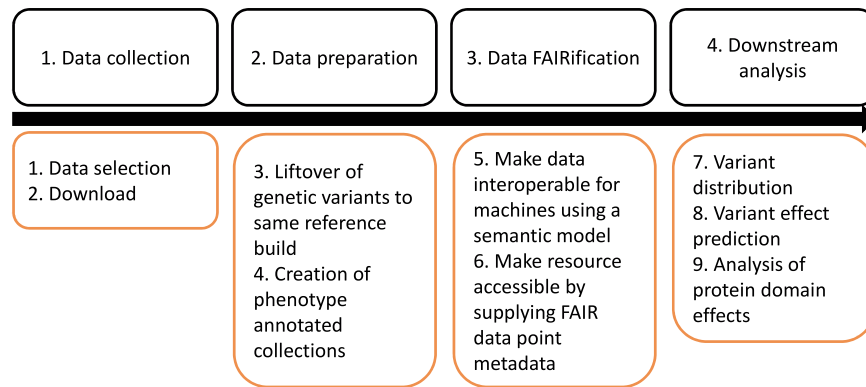


Fig. 1 Schematic drawing of the workflow of this study: data collection, preparation, FAIRification and downstream analysis.

domain, but RTT causing variations have been found in all parts of the protein¹¹. Some studies have found a distinctive correlation of phenotype severity and variation type¹², while others found a rather small or insignificant correlation^{13–16}.

Due to the rareness of RTT (prevalence about 1:10.000, ref. ¹⁷), it is important to share and communicate information about disease causing variations to increase the success of identifying genetic causes. In a previous study, we investigated the status of RTT genotype-phenotype databases and the methods that different resources use to share newly identified genetic variants on the example of RTT¹⁸. Thirteen different genotype-phenotype databases were identified that are used to collect and share genetic variants annotated with observed or predicted effects. Our main conclusion was that databases store and provide information in very different ways, such that now it is technically infeasible to query multiple databases and combine the results in an efficient and automated way. In line with the IRDiRC aims for rare diseases (<http://www.irdirc.org/about-us/vision-goals/>), the bioinformatics infrastructure should contribute to store, curate and make data about known disease causing and benign variations available. Therefore, the interoperability of these databases needs to improve to be able to efficiently use their contents in combination.

In this study, we show how to integrate the available RTT genetic and phenotypic data across multiple databases and use the integrated data for further analysis about RTT, in order to investigate variant abundance and distribution and to test variant effect prediction algorithms. We followed the FAIRification workflow¹⁹ to make the data more findable, accessible, interoperable, and reusable for computer processing. In line with the FAIR data point specification, a combination of DCAT and Re3Data vocabularies were used to describe the data set [<https://github.com/FAIRDataTeam/FAIRDataPoint-Spec/blob/v0.1.0/spec.md>]. The resulting ‘FAIR data point’ refers to two distribution formats: one in RDF and one in CSV. RDF was used to create a self-describing, machine interpretable version of the data using existing global ontologies. The resulting datasets (CSV) are also shared on Figshare (see DOI in results). To our knowledge, the dataset created and used in this study is the largest collection of annotated disease-causing and benign *MECP2* variants available at this moment, and may help researchers investigate and test disease models.

Methods

Workflow of genetic variant data integration. *Data selection and retrieval.* In a recent study¹⁸, we identified thirteen genotype-phenotype databases containing RTT-specific *MECP2* variation data. We evaluated each of these for specific requirements for data integration. Data should be 1) available and permitted to be re-used and redistributed, 2) the given description of genetic variants should be for an unambiguous variation. The latter means that the exact position (chromosome build and location) as well as the variation of the genetic variants are available or retrievable by conversion, thus, they can be described using the HGVS nomenclature. For this study, we selected eight databases and downloaded all *MECP2* genetic variants with available linked phenotype information from each of these databases: ClinVar²⁰, <https://www.ncbi.nlm.nih.gov/clinvar/>, DECIPHER²¹, <https://decipher.sanger.ac.uk/>, EVA (<http://www.ebi.ac.uk>), EVS (<http://evs.gs.washington.edu>), ExAC²², <http://exac.broadinstitute.org/>, KMD (<https://kmd.nih.gov.kr>), LOVD²³, *MECP2* collection: <https://databases.lovd.nl/shared/genes/MECP2>), and RettBASE²⁴, <http://mecp2.chw.edu.au/>. Additionally, an anonymized dataset from local RTT patients was included (Maastricht Rett dataset, permission granted by Niet-WMO verklaring 2018-0597, Maastricht University METC approval). Either the integrated download function was used to get the data or data was extracted from HTML (see the availability of download functions in ref. ¹⁸). Figure 1 shows the data processing (step 1–3) and analysis (step 4) workflow of this study.

Liftover to enable compatible genetic variant description formats. The *MECP2* genetic variant descriptions from the different sources were made compatible and therefore comparable by application of the HGVS nomenclature and the same reference sequence. This is the first step to make the data interoperable. For this, we used the reference sequence for chromosome 23 (X) NC_000023.11, which is part of the current human genome reference assembly (GRCh38). Genomic descriptions were used to ensure that variations in and outside the gene region (exonic, intronic, up- and downstream) were included. The process of re-describing all variants with the

HGVS nomenclature using the same reference build, liftover, was done by using the Mutalyzer position converter webtool [<https://mutalyzer.nl/>]²⁵. Mutalyzer can perform a conversion between different reference sequences and categories (e.g. complete genomic regions NC and mRNA NM) but requires nomenclature-compliant input. Manual correction was performed on genetic variant descriptions that did not have the complete and correct format for conversion but provided enough information to correct the format.

Creation of phenotype annotated collections. Genetic variants were assigned by their linked phenotype information to three different categories: 1) RTT causing (verified by identification as disease causing variant according to the requirements of the databases) (data from DECIPHER, Maastricht Rett dataset, ClinVar, RettBase and KMD), 2) benign (verified by finding them in a healthy control subject) (data from ClinVar and RettBase), and 3) unknown evidence (only pathogenicity prediction scores provided by database) (EVS, EVA, LOVD, ExAC). The exact annotation selection criteria for the RTT causing dataset based on the databases individual annotations are given in Sup. Table 1. These lists are collected and used for further analysis.

Data FAIRification. We made the prepared genetic variant and phenotype data more Findable, Accessible, Interoperable, and Reusable for humans and computers following the FAIR guiding principles²⁶. The data was made machine-readable (in RDF format) using a semantic data model (see below) and a general-purpose FAIRifier tool²⁷ based on the OpenRefine data cleaning and wrangling tool (<http://openrefine.org/>) and an RDF plugin (<https://github.com/stkenny/grfine-rdf-extension>). Similarly, machine-readable metadata (information about the data) was generated using the Metadata Editor²⁷.

We applied and extended the semantic data model of a genetic variant described in ref. ²⁸ to convert the prepared data to RDF. The model is available on GitHub (<https://github.com/LUMC-BioSemantics/rett-variant>) and describes the important data elements of the datasets: 1) the genetic variant: HGVS nomenclature, start/end position of the variation, and genome build, and 2) the phenotype information that describes whether a variant is thought to be RTT causing, benign or unknown.

Downstream analysis examples. *Network analysis of data distribution in RTT databases.* To analyse the distribution of *MECP2* variations in the RTT databases, a network was created where the nodes represent databases and the node size the number of available *MECP2* variations. The thickness of the lines connecting the databases indicate how many *MECP2* variations they share. Network visualization and analysis software Cytoscape²⁹ was used for this purpose.

Variant annotation and characterization by genomic features. To characterize all the collected *MECP2* variants, we developed an automatic analysis pipeline for variant annotation. We used the HGVS corrected variants to integrate custom scripts with HGVS conversion tool from <https://github.com/counsyl/hgvs> and generated VCF files for annotation within an automated pipeline available at <https://gitlab.bsc.es/inb/fair-rett>. Afterwards, we proceeded to annotate variants with Ensembl Variant Effect Predictor, VEP³⁰ v94 using the GRCh38 assembly, selecting all available features, plus optional plugins to estimate variant pathogenicity (i.e., PolyPhen³¹, SIFT³², MetaLR³³, CADD³⁴, FATHMM-MKL³⁵ from dbNSFP and dbSNV scores³⁶) both in coding and splicing regions.

The resulting VEP annotated data was processed with R scripts, available at <https://gitlab.bsc.es/inb/fair-rett>, to compare RTT causing and benign variants as subsets, and to generate summary statistics for these. The scripts allow to compare and visualize the two classes in terms of any of the available VEP annotation features, (e.g. variant frequency in the population, estimated variant consequence, and conservation score of the genomic location). Using this we compared the two datasets of RTT causing and benign variants by pathogenicity scores, impact (i.e. estimation of the consequence of each variant on the protein sequence), variant frequency, and genomic location. Because a few variations appear both as RTT causing and benign, we represented this subset of variants as a third class (“both”) in all visualizations.

Finally, we focused on exonic missense variants and used VEP information about the amino acid change and position within the *MECP2*-e2 transcript to visualize the variation distribution across protein domains and conserved regions (as described in ref. ³⁷). This allowed us to make a finer characterization of differential distribution of RTT causing and benign variants across *MECP2* domains.

Data Records

The machine-readable metadata was made available on a FAIR Data Point³⁸ (<https://github.com/FAIRDataTeam/FAIRDataPoint-Spec>) available via: <http://purl.org/biosemantics-lumc/rettbase/fdp>. The FAIR Data Point metadata provides URIs that resolve to the RDF and CSV files for each of the nine sources on Figshare (<https://doi.org/10.6084/m9.figshare.c.4769153.v2>)³⁹.

Technical Validation

Data integration challenges identified. We encountered several challenges while integrating data from the different RTT databases: 1) different descriptions of genetic variants were used, 2) liftover process and limitations in automated liftover, and 3) findability of terms of use/re-use, detailed below.

1) For the descriptions of genetic variants, the most commonly used nomenclature was HGVS. HGVS still comes in different, correct, flavours, e.g. using genomic or cDNA positions or different (versions of) reference sequences, which still need conversions from one to the other using for instance Mutalyzer. The other most common standard was the RS number (reference SNP identifier, from dbSNP). These are usually linked to loci and can therefore not be used as unambiguous identifiers for a variant. Databases that give only RS identifiers were therefore not

Phenotype annotation format	Database	Number of <i>MECP2</i> or RTT variations	Number of variations meeting annotation criteria	Number of variations with sufficient genetic annotation			
				Using Mutalyzer		After manual curation	
				#	%	#	%
Phenotype	DECIPHER	34	25	25	100	25	100
Diagnosis	Maastricht Rett dataset	429	428	388	90.7	393	91.8
	ClinVar	1,134	743	681	91.7	726	97.7
	RettBase	4,705	3,986	3,798	95.3	3,980	99.8
	KMD	35	35	35	100	35	100
Pathogenicity scores	EVS	190	190	190	100	190	100
	LOVD	808	808	738	91.3	804	99.5
	EVA	4,226	4,226	4,193	99.2	4,216	99.8
	ExAC	599	599	559	93.3	599	100
TOTAL		12,158	11,040	10,607	average 95.7%	10,968	average 98.7%

Table 1. Overview for the different databases, their phenotype annotation format, number of available *MECP2* variants, and data liftover success rates using automated (Mutalyzer) and manual curation.

included in further analysis. The same problem occurred with the annotation of diagnosis and/or phenotypes. As described before in ref. ¹⁸, only a few databases link original diagnostic information to the genetic information, and whenever it was provided, this information was presented using different formats or definitions.

2) For the liftover to one common, comparable variant description (GRCh38 (hg19)), genomic position) Mutalyzer was used. It can be used programmatically via API (Application programming interface) or via Graphical User Interface (GUI). After liftover to HGVS nomenclature it was possible for the majority of variants (90.7–100% per dataset) to use Mutalyzer without further curation (Table 1). Nevertheless, for up to 9.3% of the variations in a dataset (Maastricht Rett dataset, the average was 4.3%, Table 1) the data needed curation due to typos, incorrect nomenclature (e.g., symbols which are not in the official nomenclature), or outdated/historic position description (e.g., Genbank variation description nomenclature). Mutalyzer itself cannot deal with insertions of a number on unknown base pairs (e.g., ins3 instead of insATT), round brackets () to indicate uncertainty (they are gone after translation while square brackets [] to indicate different alleles or group alleles are fine), asterisk * to indicate stop (protein) according to the official HGVS nomenclature. These variations required manual curation, e.g. changing round brackets to square brackets, use Mutalyzer to do the liftover, changing square brackets back to round brackets. Furthermore, it is currently not possible to do a direct liftover from one genomic reference sequence to another (e.g., NC_000023.10:g.153282026 G > A to NC_000023.11:g.154016575 G > A) due to the size of the reference sequence. At the moment, this must be done in two steps via transcript (NC -> NM -> NC).

3) The permission to reuse and redistribute was difficult to find for some databases (RettBase, KMD).

Size and content of the FAIR dataset. *Number of disease causing and benign *MECP2* genetic variants available.* Based on the thirteen genotype-phenotype databases identified in ref. ¹⁸, the inclusion criteria for this study were not met by DisGeNET, dbSNP, dbVAR, Café Variome, and HGMD. DisGeNET, dbSNP and dbVAR did not provide unambiguous descriptions of variations as the RS identifier only indicates a location of polymorphism and needs evaluation of the, sometimes ambiguous, additional information about the nucleotide change. Café Variome provided only protein change information, which, although very relevant itself, cannot be translated back to an unambiguous genetic change. HGMD, the only commercial database, did not allow re-use and re-distribution of the content. The eight databases that did fulfil our inclusion criteria and data previously anonymized from local RTT patients were used in this study (see Table 2). At the time of research, in total 12,158 *MECP2* variation entries were found in these databases. The databases contained between 34 (DECIPHER) and 4,706 (RettBASE) *MECP2* variations (Table 2). Between 15% and 100% of these variations were unique database entries (occur only once in one single database). Multiple entries of one variation were found frequently in disease specific databases, giving an indication of the abundance of this variant and confirming its pathogenicity. In total we identified 4,573 RTT causing *MECP2* variants (of which 863 were unique) that annotate genetic information with diagnosis (RettBase, ClinVar, Maastricht Rett dataset, KMD) and/or clear phenotype descriptions (DECIPHER) clearly stating that they cause RTT (or similar e.g., X-linked mental retardation) (intake criteria Sup. Table 1). We identified 617 benign *MECP2* variants, of which 209 were unique, from two of the databases that annotate with diagnosis information (RettBase and ClinVar). These were clearly stated to be benign. Nineteen variants were found annotated both as RTT causing and benign (Sup. Table 2).

In total, we collected 12,158 *MECP2* variants, which resulted in a collection of 10,968 (5,038 unique) curated and integrated variants. Out of the 10,968 curated *MECP2* variations only eleven occur in more than 1% of all database entries, and these account for 53.7% of all database entries (data not shown).

The 863 unique RTT causing variations are distributed over 4,573 database entries. Also here, only twelve variations are found in more than 1% of all database entries (Table 3) and these twelve make in total 60% of the database entries. The most abundantly found *MECP2* variations were found in seven of nine databases (Table 3). The majority (eight) of these are C > T transitions at CpG hotspots⁴⁰. These eight *MECP2* hotspot variations contribute to 49.7% of all *MECP2* variation entries. The most abundant *MECP2* variation in this dataset is NC_0000

Database	Number of total <i>MECP2</i> variation entries	Number of unique <i>MECP2</i> variations		Number of unique variations which occur only in this database
		#	% of total <i>MECP2</i> variation entries	
EVA	4,226	4,192	99.2	3,329
LOVD	808	802	99.3	144
RettBase	4,705	740	15.7	209
ExAC	599	599	100.0	40
ClinVar	1,134	716	63.1	126
EVS	190	95	50.0	1
Maastricht Rett dataset	429	68	15.9	34
KMD	35	35	100.0	9
DECIPHER	34	23	67.6	2

Table 2. Numbers of total and unique *MECP2* variations in each database.

Genomic position [†]	count	%	cDNA [‡] and protein change [§]	Effect and previous reports
g.154031355 G > A	463	10.1	c.473 C > T, p.(Thr158Met) [‡]	Missense variation ^{24,43,44}
g.154031326 G > A	409	8.9	c.502 C > T, p.(Arg168*) [‡]	Nonsense variation, leading to truncation ^{24,43,44}
g.154031065 G > A	345	7.5	c.763 C > T, p.(Arg255*) [‡]	Nonsense variation, leading to truncation ^{24,43,44}
g.154031020 G > A	309	6.8	c.808 C > T, p.(Arg270*) [‡]	Nonsense variation, leading to truncation ^{24,43,44}
g.154030948 G > A	281	6.1	c.880 C > T, p.(Arg294*) [‡]	Nonsense variation, leading to truncation ^{24,44}
g.154030912 G > A	279	6.1	c.916 C > T, p.(Arg306Cys) [‡]	Missense variation ^{24,43,44}
g.154031431 G > A	249	5.4	c.397 C > T, p.(Arg133Cys) [‡]	Missense variation ^{24,44,53}
g.154032268 G > A	161	3.5	c.316 C > T, p.(Arg106Trp) [‡]	Missense variation ^{24,44}
g.154031373 G > C	80	1.7	c.455 C > G, p.(Pro152Arg)	Missense variation ⁵⁴
g.154031022delC	67	1.5	c.806delG, p.(Gly269fs)	Frameshift deletion leading to missense ⁴¹
g.154030621_154030664del44	50	1.1	c.1164_1207del44, p.(Pro389*)	Frameshift deletion leading to truncation
g.154030631_154030671del41	49	1.1	c.1157_1197del41, p.(Leu386fs)	Deletion leading to frameshift

Table 3. Most abundant RTT causing variants in this study. RefSeq [†]NC_000023.11, [‡]NM_004992.3, [§]NP_004983.1. [¶]one of the eight hotspot variations⁴⁰.

23.11:g.g.154031355 G > A (NM_004992.3:c.473 C > T, NP_004983.1:p.(Thr158Met)) with 463 counts (Table 3). In total 54% of RTT causing variations are a deletion, 9% insertion, 37% substitution, and 9% duplication. Many of the database entries contain multiple variations (e.g., a deletion and insertion) on the same or different chromosomes. 452 RTT causing variations have only one single database entry and of these 269 are a deletion, 43 insertion, and/or 153 substitution.

Distribution of the variants across databases. Table 2 shows the number of unique *MECP2* variations for each investigated database. The different databases contain very different numbers of unique *MECP2* variations. The number of unique *MECP2* variations in a database gives an indication whether it is a database focusing on collecting pathogenic variations (RettBase, ClinVar, Maastricht Rett dataset, DECIPHER) (exception KMD) or general population sequencing results (no disease annotation) (EVA, LOVD, ExAC) (exception EVS). LOVD, for example, lists all different variations and provides background information about the abundance of one variation in the variations' information sheet. RettBase also gives the reference where a specific entry is from. From Table 2 it also becomes clear that every database has unique *MECP2* variations, which are found in no other database. The number of such unique variants differ between 3,329 (EVA) and one (EVS).

Figure 2 shows the size of *MECP2* variation collections in the different databases, their shared and their unique variations. There are databases that focus on collections of genome and/or exome sequencing data of mostly healthy individuals (EVA, EVS, ExAC), curated collections of disease causing variants (LOVD, RettBase, ClinVar, Decipher), and hospital derived collections (KMD, Maastricht Rett dataset). The overlap or shared *MECP2* variations between databases can be explained by the occurrence of this variation in multiple patients, data exchange between databases, or by recruitment from the same resources. For instance, ExAC and LOVD share 559 unique variants, LOVD and ClinVar 546, LOVD and RettBase 512, RettBase and ClinVar 504.

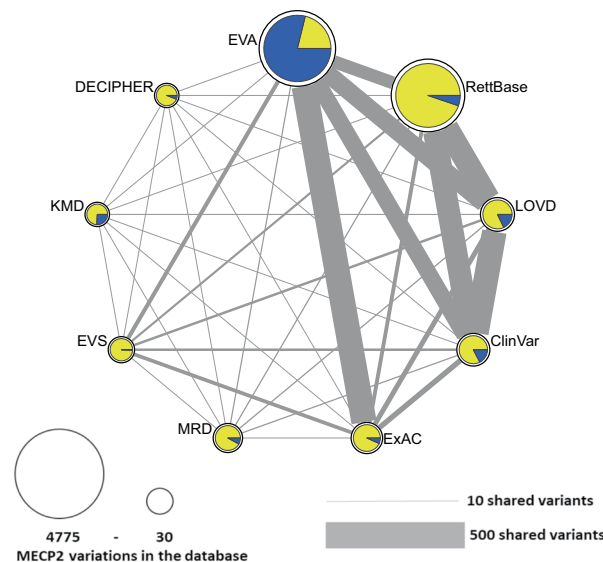


Fig. 2 Network illustrating the number of unique and overlapping *MECP2* variations within and between nine Rett syndrome databases: DECIPHER, Maastricht Rett dataset (MRD), ClinVar, Rettbase, KMD, EVS, LOVD, EVA, and ExAC. Each node (circle) represents a database. The node size correlates with the number of variants (between 30 and 4775), the edge thickness correlates with the number of overlapping/shared variants between the two databases (between 0 and 500). The colour of the charts in the nodes represent the proportion of unique variants (blue) versus variants shared with other databases (yellow).

Advantages of making the *MECP2* genetic variant data FAIR. The FAIR guiding principles have emerged from analysing the general, and often repeated, process that data scientists go through when preparing data from multiple sources for data integration and analysis. The *MECP2* genotype-phenotype data from this study were retrieved from nine heterogeneous resources, which we prepared for analysis by making them more FAIR. This was first and foremost done to enable integration of the data for analysis as correctly as possible, which also facilitates integration with other interoperable data such as protein functionality data from, for instance, UniProt, NextProt or Phyre databases. Another reason was to ensure reusability of the integrated data for other research studies. Note, all the FAIRified resources allow redistribution.

The FAIRified data was described with machine-readable metadata and distributed at a new location, which prospectively allows other researchers to reuse this data. Thus, as data users, we made the data FAIR after retrieving them from their respective distributions. This was necessary, because the way that the data were provided by the different sources was not sufficiently uniform for machines to integrate multiple sources. The disadvantage of leaving the implementation of FAIR principles to data consumers is that they are more likely to make mistakes in the interpretation of the meaning of the data, which may not be the same as the sources. Ideally, data are made FAIR at the source to minimize that risk and optimize transparency. This would have allowed us to directly use the data in automated workflows that can be run regularly to update our findings.

Usage Notes

General use of the data. *The CSV formatted files.* The CSV formatted files contain in the first column genetic variants generated and used in this effort, which were described with the HGVS nomenclature, a common format to describe genetic variants. The second column indicates the source (name of the database) and the third column classifies the variations into RTT causing, benign or NA (not applicable). Any analysis or visualization tool that can deal with HGVS formatted files like VCF can work with this format.

The RDF files. The RDF output of this dataset is serialized in the Turtle format. To use these RDF files, we need to upload it to a triple store, which is a special database designed to store and query RDF files. We can use the SPARQL query language to query the RDF files. We linked our RDF to external RDF data sources such as Ensembl RDF and Orphanet Rare Disease ontology. Within a SPARQL query we can also exploit these links to do further integration queries.

Visualization in genome browsers. Broadly used genome browsers, e.g. UCSC accepts HGVS variant descriptions in their interactive interface for visualization purposes, e.g., see the result [Link] of NC_000023.11:g.1000000 C > T. In addition, conversion from the CSV formatted file to a “BED detail” formatted file [Link] should be relatively straightforward prior to its ingestion by the genome browsers. This alternative format can also be used in most modern genome browsers.

Biological questions answered using this data. *Variant pathogenicity prediction vs. curated datasets.* To explore differences between RTT causing and benign *MECP2* genetic variants we analyzed the annotated

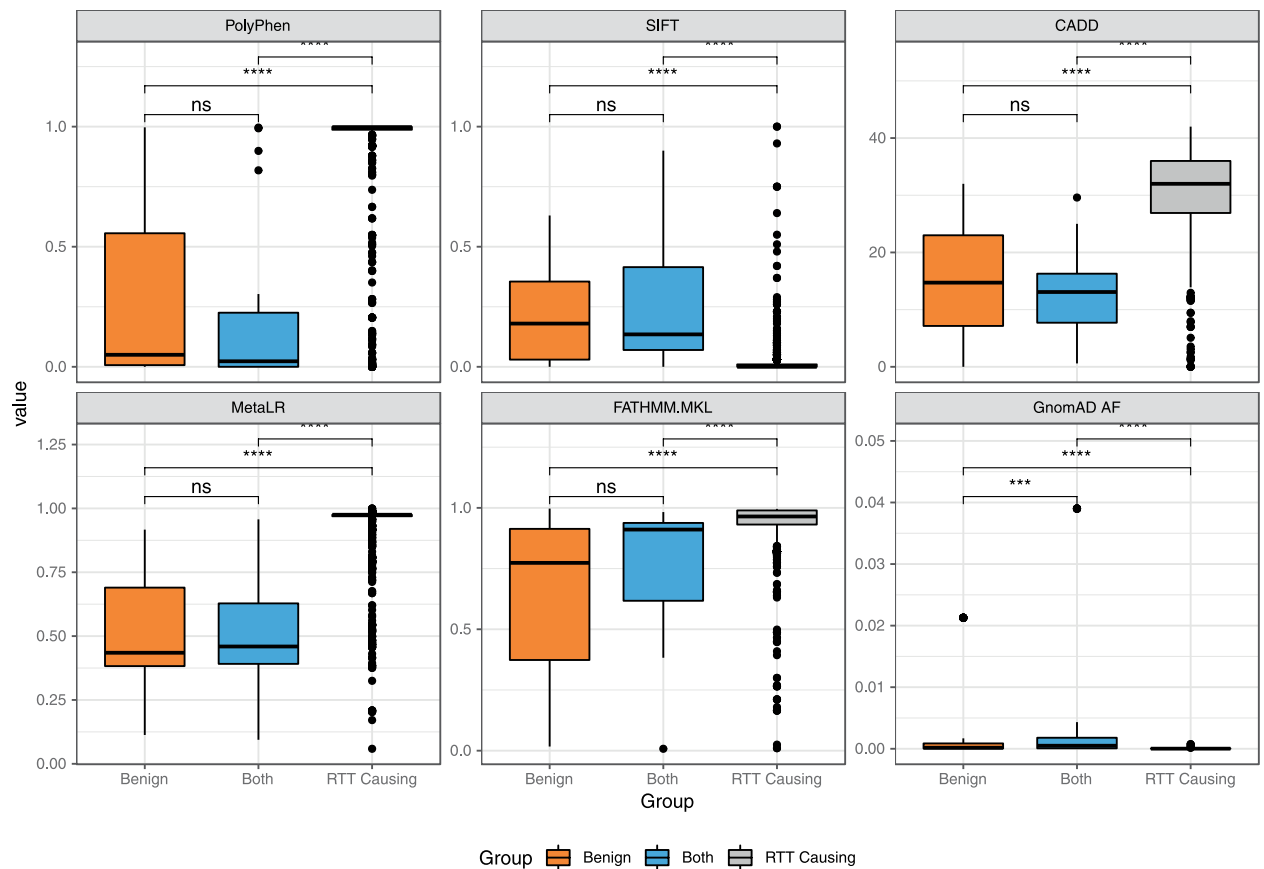


Fig. 3 Boxplots comparing prediction score value distribution calculated by different tools from the benign, both and RTT causing *MECP2* genetic variants. The effect prediction was done based on conservation score (PolyPhen), four pathogenicity scores (SIFT, CADD, MetaLR, and FATHMM.MKL), and the variant allele frequency in the GnomAD dataset.

results from VEP (see Methods) from six descriptive features (Fig. 3). We chose to visualize the obtained scores about conservation (i.e., PolyPhen), pathogenicity estimation scores (i.e., SIFT, CADD, MetaLR, FATHMM-MKL), and the variant frequency in normal population from GnomAD²² (i.e., GnomAD_AF).

We classified variants by benign, RTT causing and “both”, as we identified a subset of 19 variants appearing in both datasets. The detailed list of the *MECP2* variations, which occur in both, can be found in Sup. Table 2. Overall, we see expected results: the RTT causing variants were found to be in positions significantly more conserved than the benign or both variants (Fig. 3, PolyPhen (Wilcoxon test)), as well as less frequent than benign variations even though, all variants presented here are not abundant in the normal population (Fig. 3, GnomAD_AF). Analysis of the obtained estimation of pathogenicity from multiple scores (Fig. 3 panels SIFT, CADD, MetaLR and FATHMM-MKL), shows that RTT causing variants are on average predicted as more damaging than the benign and both variants ($p < 0.0001$ in all cases after applying Wilcoxon test). Note that SIFT associates more pathogenic variants to lower scores, whereas CADD, MetaLR and FATHMM-MKL associates more pathogenic variants to higher scores. MetaLR is better than the other three pathogenicity scores in distinguishing benign and RTT causing variant types. This may be because this novel meta-score integrates more features than the other three prediction tools, amongst other pathogenicity scores and frequency information.

The pathogenicity estimates of the both group place the variants between the benign and RTT-causing in three out of five predictions, while in the other two give a prediction more similar to the benign group.

Distribution of pathogenic and benign missense variations to protein domains. In this experiment the position of RTT causing and benign missense variants in different domains and conserved regions of *MECP2* are compared (Table 4 and Fig. 4). Most RTT causing missense variations are found in the methyl-DNA binding domain (MDB) (68.3%) and in the transcription repressor binding domain (TRD). However, at lower frequencies, RTT causing missense variations can also be found in the other domains. The benign variants are most frequent in the C-terminal domain (55.1%) and the interdomain (28.1%), but can likewise also be found in the other domains at lower frequencies. The distribution across the conserved regions of *MECP2* shows that 93.6% of the missense RTT causing variants are found in conserved regions while only 16.3% of the benign variants are found in conserved regions.

Regarding mutations affecting the protein sequence, in 1350 cases, which is about half of the RTT-causing single nucleotide mutations, the variations are truncating, changing an Arginine into a stop codon. Also frequently,

		Domain length (% of total)	RTT causing	Benign
			% of missense variations per region	
Domains	N-terminal domain	78 (16.0)	0.2	1
	Methyl-DNA binding domain	84 (17.3)	68.3	1.5
	Interdomain	45 (9.3)	1.9	28.1
	Transcription repressor domain	103 (21.2)	24.1	14.3
	C-terminal domain	176 (36.2)	5.5	55.1
Conserved regions			93.6	16.3

Table 4. Location of RTT causing and benign missense variants in different domains and conserved regions of MECP2.

Arginine is changed into a Cysteine (533) or Tryptophan (179) which are major changes in protein 3D structure. The average BLOSUM62 value of all amino acid changes for the RTT causing dataset is -1.8 . For the benign MECP2 variations, the most abundant variations are silent (= not amino acid changing), coding for Serine (65), Threonine (44) and Proline (40). The most abundant amino acid change is Glutamic acid to Lysine (33) and the average BLOSUM62 value of all amino acid changes indicates with -0.3 less severe consequences for the protein structure than the RTT causing group.

Added-value of integration of data across different sources. This is to our knowledge the first study that integrates genetic variation data from multiple databases on MECP2. Despite best efforts of individual sources to reach the largest possible coverage, our results demonstrate that the number of usefully annotated variants increases when databases are combined. The greatest advantage of the integrated approach is therefore that more variants become available for further research and diagnosis. This is especially interesting for rare diseases which have relatively small study populations. By mapping to a common reference sequence, the information of different sources becomes comparable and we are getting closer to the “true” number of variants known. In this study, we were able to increase the previously estimated numbers of a few hundred RTT causing unique sequence variations to 863. However, databases, at least the active ones, get regular updates and input of data. In the time from the beginning of this study the number of variants in e.g. RettBase increased from 4,738 (March 2018¹⁸) to 4,757 (November 2018) to 4,806 (NM_004992.3, April 2020). Consequently, the number of 863 known RTT causing variants is likely outdated when this study is published. We argue that it is unrealistic to assume that any single database will ever be completely comprehensive, unless it automatically pulls in updates from other databases. A possible contribution to the solution of this problem would be to create the combined list of pathogenic variants by automated workflows that find and summarize data from across databases on demand or continuously. To make that possible we need to standardize how databases provide data for machine processing. The role of FAIR data principles to achieve this is discussed later in more detail.

This integrated dataset gives the possibility to study abundance and prevalence of certain variations in a larger population than any of the study populations published before. There are several studies on relatively small^{41,42} or large populations (e.g.^{43,44}) that have published their data in the previous years. Study ref. ⁴³ analysed 301 different MECP2 alleles in a French population and found 69 different variations, which cause 64% of RTT. They identified NP_004983.1:p.R168*, R255*, R270*, T158M, and R306C (Table 5) as the most abundant variations and 59 variations were found in only one or two patients. In the list from the US national history study (819 participants⁴⁴) the variations R106W, R133C, T158M, R168*, R255*, R270*, R294*, and R306C were responsible for more than 60% of RTT. The MECP2 variation content of RettBase was analyzed recently by ref. ²⁴ and the following eight hotspot variations are responsible for a total of 47% of RTT cases (of total number of MECP2 entries was at that time 4668, disease causing and benign): R106W, R133C, T158M, R168*, R255*, R270*, R294*, and R306C. ref. ³ provides information about eleven more datasets from different countries.

Although our study resulted in a different ranking of the eight hotspots, we could confirm these as the most abundant ones which occur in our dataset in 54.6% of all RTT causing database entries. All eight hotspot mutations are C > T transitions leading in seven of eight cases to a change from Arginine to a stop codon, Cysteine or Tryptophan which are changes with a high probability to change the 3D structure of the protein. The special vulnerability of certain Cytosine positions to errors in base excision repair was described before⁴⁵.

In our integrated dataset most pathogenic mutations in MECP2 occur in the methyl-DNA or transcription repressor binding domain. This finding has been described and confirmed before^{24,46–48}. The functionality of the methyl-DNA binding domain is reported to be extremely sensitive to changes⁴⁶. The importance of the domain also shows from the observation that a construct consisting only of methyl-DNA binding and transcription repressor domain could preserve some basic functions of MECP2⁴⁹. There is also a clear distinction between conserved and non-conserved regions. As expected, disease-causing mutations occur much more often in the conserved regions. However, the data shows clearly that mutations in all domains, both conserved and non-conserved regions, can cause RTT. The open question here remains how much influence does a particular mutation have and how much is contributed by other genetic aspects or environmental influences. This question becomes more important considering the discovery of variants that in one individual can be benign and RTT causing in another.

How can the same variation be benign AND cause RTT in different individuals. The majority of the MECP2 genetic variations, which are described as RTT causing in one, and benign in another database entry, are predicted to be benign (Fig. 3). Possible explanations why a variant can be disease causing in one individual and

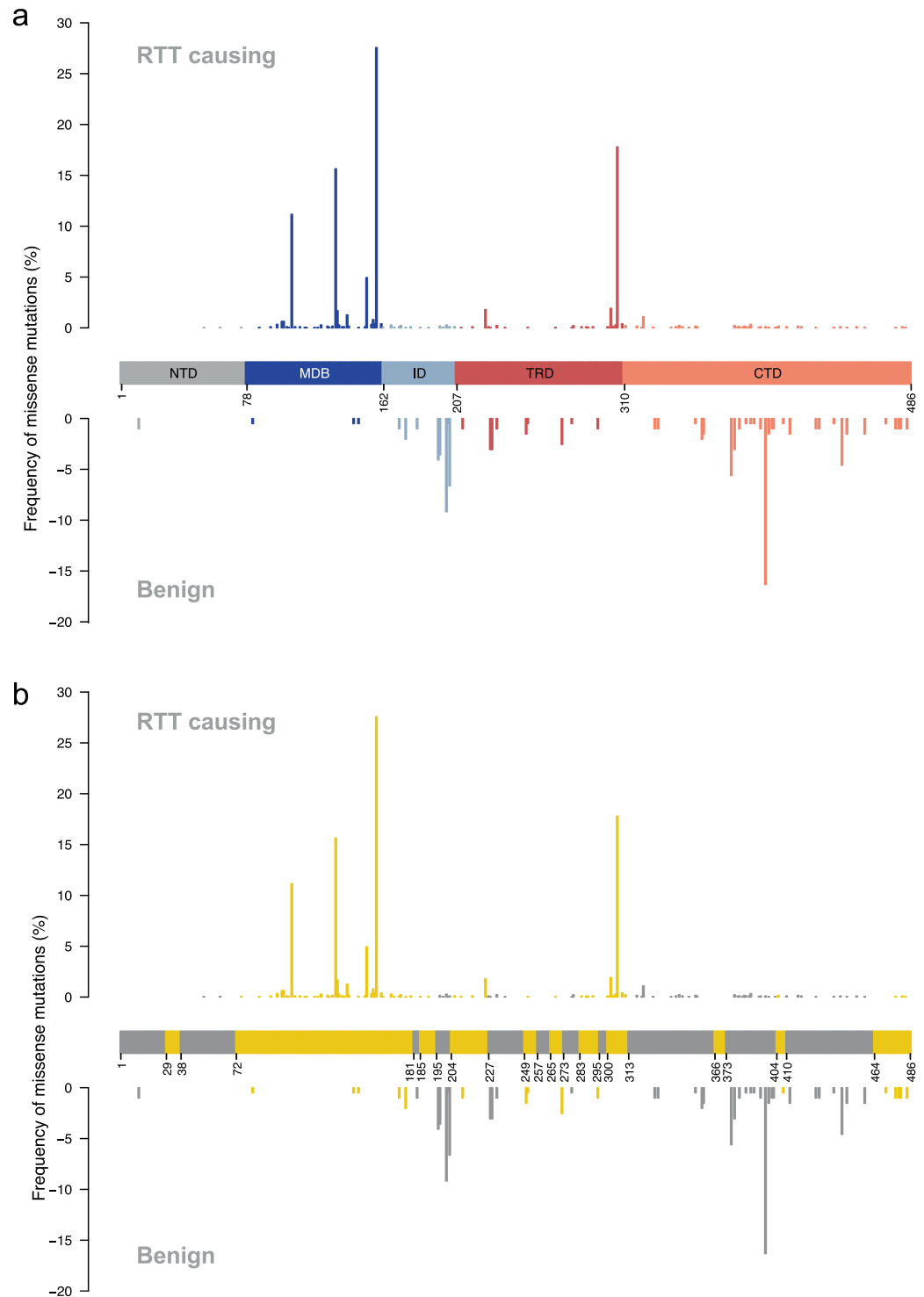


Fig. 4 Distribution of RTT causing and benign *MECP2* missense variations. Amino acid positions correspond to isoform *MECP2*-e2 (the result of translation initiated at exon 2). Frequency is represented as the percentage of missense variations falling in each position, from the total of missense variations in cases or controls. In a) each *MECP2* domain is coloured differently, while in b) conserved deletions are coloured in yellow. Domain abbreviations: N-terminal domain (NTD), methyl-DNA binding domain (MDB), interdomain (ID), transcription repressor binding domain (TRD), C-terminal domain (CTD).

benign in another could be due to the location of the gene on the X chromosome which may result in a subclinical phenotype in females but a fully-fledged RTT in male patients. The sex of patients is usually not given in these genotype-phenotype databases. In addition, X-inactivation patterns⁵⁰ and genetic background related to other participating genes in *MECP2* related pathways⁴ influence the severity of a rare monogenic (X-linked) disease and

Sample size and citation	Variations mentioned in studies (abundance in % if known)							
4573 variations annotated with RTT causing (this study)	Thr158Met (10.1)	Arg168* (8.9)	Arg255* (7.5)	Arg270* (6.8)	Arg294* (6.1)	Arg306Cys (6.1)	Arg133Cys (5.4)	Arg106Trp (3.5)
301 RTT patients ⁴³	Thr158Met (7.8)	Arg168* (11.5)	Arg255* (10.9)	Arg270* (10.5)		Arg306Cys (6.8)		
819 RTT patients ⁴⁴	Thr158Met (11.0)	Arg168* (10.4)	Arg255* (11.0)	Arg270* (5.5)	Arg294* (6.3)	Arg306Cys (6.9)	Arg133Cys (3.8)	Arg106Trp (3.1)
RettBase 4668 total entries ²⁴	Thr158Met	Arg168*	Arg255*	Arg270*	Arg294*	Arg306Cys	Arg133Cys	Arg106Trp

Table 5. Comparison of most abundant *MECP2* variations in different studies. RefSeq: NP_004983.1:p.

can possibly even save individuals with a documented pathogenic variation from disease development⁵¹. In principle, patients could also have an unreported second mutation that could cause the effect either alone or through epistatic interaction. Another reason for misinterpretation of a variant may be due to linkage disequilibrium, where the causal mutation is not the reported one but another unreported.

For several variations, a high pathogenicity score was predicted but they were still documented in healthy individuals. This has been observed before in a girl with RTT who inherited a germline disease causing *MECP2* c.1160 C > T (P387L, NC_000023.11:g.154030668 G > A) variation from a healthy (!) father⁵². We found exactly this variant only in our RTT causing dataset (documented in ClinVar and RettBase), the annotation with the benign outcome was not added to one of these databases yet. These effects may contribute to the limited penetrance of some mutations. To unravel the different influences of *MECP2* variations in the context of an individual patient, we need to evaluate how genetic background (ancestry) can affect other process related genes. For this, genotype-phenotype databases with detailed phenotype capture will be highly important and data integration tools and methods must be developed to investigate this further.

There is a possibility that a gene carries more than one variation. Indeed, in our integrated dataset we found a total 54 individuals with multiple variants. However, multiple variants are difficult to predict, there may be positive or negative epistatic effects if these variants occur on the same allele or two mutations affecting the same codon, one cancelling out to another. All of these possibilities may lead to wrong classification of variants.

Code availability

Any custom code used to generate and analyse this dataset is openly available on Git-based repositories. For data FAIRification see <https://github.com/stkenny/grefine-rdf-extension> and <https://github.com/LUMC-BioSemantics/rett-variant>, for VEP data analysis see <https://github.com/counsyl/hgvs>, and <https://gitlab.bsc.es/inb/fair-rett> for _summary_plots and HGVS pipelines.

Received: 1 July 2020; Accepted: 30 November 2020;

Published online: 15 January 2021

References

1. Rett, A. On a unusual brain atrophy syndrome in hyperammonemia in childhood. *Wien Med Wochenschr* **116**, 723–726 (1966).
2. Amir, R. E. *et al.* Rett syndrome is caused by mutations in X-linked *MECP2*, encoding methyl-CpG-binding protein 2. *Nat Genet* **23**, 185–188, <https://doi.org/10.1038/13810> (1999).
3. Percy, A. K. *et al.* Rett syndrome: North American database. *J Child Neurol* **22**, 1338–1341, <https://doi.org/10.1177/0883073807308715> (2007).
4. Pizzo, L. *et al.* Rare variants in the genetic background modulate cognitive and developmental phenotypes in individuals carrying disease-associated variants. *Genet Med*, <https://doi.org/10.1038/s41436-018-0266-3> (2018).
5. Nan, X. *et al.* Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature* **393**, 386–389, <https://doi.org/10.1038/30764> (1998).
6. Ehrhart, F. *et al.* Integrated analysis of human transcriptome data for Rett syndrome finds a network of involved genes. *World J Biol Psychiatry accepted*, 1–14, <https://doi.org/10.1080/15622975.2019.1593501> (2019).
7. Shovlin, S. & Tropea, D. Transcriptome level analysis in Rett syndrome using human samples from different tissues. *Orphanet J Rare Dis* **13**, 113, <https://doi.org/10.1186/s13023-018-0857-8> (2018).
8. Bedogni, F. *et al.* Rett syndrome and the urge of novel approaches to study MeCP2 functions and mechanisms of action. *Neurosci Biobehav Rev* **46 Pt 2**, 187–201, <https://doi.org/10.1016/j.neubiorev.2014.01.011> (2014).
9. Adams, V. H., McBryant, S. J., Wade, P. A., Woodcock, C. L. & Hansen, J. C. Intrinsic disorder and autonomous domain function in the multifunctional nuclear protein, MeCP2. *J Biol Chem* **282**, 15057–15064, <https://doi.org/10.1074/jbc.M700855200> (2007).
10. Ballestar, E. *et al.* The impact of *MECP2* mutations in the expression patterns of Rett syndrome patients. *Hum Genet* **116**, 91–104, <https://doi.org/10.1007/s00439-004-1200-0> (2005).
11. Christodoulou, J., Grimm, A., Maher, T. & Bennetts, B. RettBASE: The IRSA *MECP2* variation database—a new mutation database in evolution. *Hum Mutat* **21**, 466–472, <https://doi.org/10.1002/humu.10194> (2003).
12. Neul, J. L. *et al.* Specific mutations in methyl-CpG-binding protein 2 confer different severity in Rett syndrome. *Neurology* **70**, 1313–1321, <https://doi.org/10.1212/01.wnl.0000291011.54508.aa> (2008).
13. Amir, R. E. *et al.* Influence of mutation type and X chromosome inactivation on Rett syndrome phenotypes. *Ann Neurol* **47**, 670–679 (2000).
14. Auranen, M. *et al.* *MECP2* gene analysis in classical Rett syndrome and in patients with Rett-like features. *Neurology* **56**, 611–617, <https://doi.org/10.1212/wnl.56.5.611> (2001).
15. Huppke, P., Laccone, F., Kramer, N., Engel, W. & Hanefeld, F. Rett syndrome: analysis of *MECP2* and clinical characterization of 31 patients. *Hum Mol Genet* **9**, 1369–1375, <https://doi.org/10.1093/hmg/9.9.1369> (2000).

16. Nielsen, J. B. *et al.* MECP2 mutations in Danish patients with Rett syndrome: high frequency of mutations but no consistent correlations with clinical severity or with the X chromosome inactivation pattern. *Eur J Hum Genet* **9**, 178–184, <https://doi.org/10.1038/sj.ejhg.5200600> (2001).
17. Laurvick, C. L. *et al.* Rett syndrome in Australia: a review of the epidemiology. *J Pediatr* **148**, 347–352, <https://doi.org/10.1016/j.jpeds.2005.10.037> (2006).
18. Townend, G. S. *et al.* MECP2 variation in Rett syndrome—An overview of current coverage of genetic and phenotype data within existing databases. *Hum Mutat* **39**, 914–924, <https://doi.org/10.1002/humu.23542> (2018).
19. Jacobsen, A. *et al.* A Generic Workflow for the Data FAIRification Process. *Data Intelligence* **2**, 56–65, https://doi.org/10.1162/dint_a_00028 (2020).
20. Landrum, M. J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res* **44**, D862–868, <https://doi.org/10.1093/nar/gkv1222> (2016).
21. Firth, H. V. *et al.* DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am J Hum Genet* **84**, 524–533, <https://doi.org/10.1016/j.ajhg.2009.03.010> (2009).
22. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291, <https://doi.org/10.1038/nature19057> (2016).
23. Fokkema, I. F. *et al.* LOVD v.2.0: the next generation in gene variant databases. *Hum Mutat* **32**, 557–563, <https://doi.org/10.1002/humu.21438> (2011).
24. Krishnaraj, R., Ho, G. & Christodoulou, J. RettBASE: Rett syndrome database update. *Hum Mutat* **38**, 922–931, <https://doi.org/10.1002/humu.23263> (2017).
25. Wildeman, M., van Ophuizen, E., den Dunnen, J. T. & Taschner, P. E. Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. *Hum Mutat* **29**, 6–13, <https://doi.org/10.1002/humu.20654> (2008).
26. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018, <https://doi.org/10.1038/sdata.2016.18> (2016).
27. Thompson, M., Burger, J., Kaliyaperumal, R., Roos, M. & Bonino da Silva Santos, L. Making FAIR easy with FAIR tools: From creolization to convergence. *Data Intelligence* **2**, 87–95, https://doi.org/10.1162/dint_a_00031 (2020).
28. Horst, E. v. d. *et al.* Finding novel associations across domains using linked data: a case study on genetic variants disrupting transcription start sites. *Proceedings of the 8th International Conference on Semantic Web Applications and Tools for Life Sciences* **1546** (2015).
29. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498–2504, <https://doi.org/10.1101/gr.1239303> (2003).
30. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122, <https://doi.org/10.1186/s13059-016-0974-4> (2016).
31. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248–249, <https://doi.org/10.1038/nmeth0410-248> (2010).
32. Sim, N. L. *et al.* SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res* **40**, W452–457, <https://doi.org/10.1093/nar/gks539> (2012).
33. Dong, C. *et al.* Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum Mol Genet* **24**, 2125–2137, <https://doi.org/10.1093/hmg/ddu733> (2015).
34. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310–315, <https://doi.org/10.1038/ng.2892> (2014).
35. Shihab, H. A. *et al.* An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* **31**, 1536–1543, <https://doi.org/10.1093/bioinformatics/btv009> (2015).
36. Liu, X., Wu, C., Li, C. & Boerwinkle, E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Hum Mutat* **37**, 235–241, <https://doi.org/10.1002/humu.22932> (2016).
37. Lombardi, L. M., Baker, S. A. & Zoghbi, H. Y. MECP2 disorders: from the clinic to mice and back. *J Clin Invest* **125**, 2914–2923, <https://doi.org/10.1172/JCI78167> (2015).
38. Bonino da Silva Santos, L. O. *et al.* Enterprise Interoperability in the Digitized and Networked Factory of the Future - FAIR Data Points Supporting Big Data Interoperability. 270–279 (ISTE Press, 2016).
39. Jacobsen, A. & Ehrhart, F. MECP2 genetic variant data. *figshare* <https://doi.org/10.6084/m9.figshare.c.4769153> (2019).
40. Wan, M. *et al.* Rett syndrome and beyond: recurrent spontaneous and familial MECP2 mutations at CpG hotspots. *Am J Hum Genet* **65**, 1520–1529, <https://doi.org/10.1086/302690> (1999).
41. Das, D. K., Raha, S., Sanghavi, D., Maitra, A. & Udani, V. Spectrum of MECP2 gene mutations in a cohort of Indian patients with Rett syndrome: report of two novel mutations. *Gene* **515**, 78–83, <https://doi.org/10.1016/j.gene.2012.11.024> (2013).
42. Inui, K. *et al.* Mutational analysis of MECP2 in Japanese patients with atypical Rett syndrome. *Brain Dev* **23**, 212–215, [https://doi.org/10.1016/s0387-7604\(01\)00197-8](https://doi.org/10.1016/s0387-7604(01)00197-8) (2001).
43. Bienvenu, T. *et al.* Spectrum of MECP2 mutations in Rett syndrome. *Genet Test* **6**, 1–6, <https://doi.org/10.1089/109065702760093843> (2002).
44. Percy, A. K. *et al.* Rett syndrome diagnostic criteria: lessons from the Natural History Study. *Ann Neurol* **68**, 951–955, <https://doi.org/10.1002/ana.22154> (2010).
45. Wang, J., Tang, J., Lai, M. & Zhang, H. 5-Hydroxymethylcytosine and disease. *Mutat Res Rev Mutat Res* **762**, 167–175, <https://doi.org/10.1016/j.mrrev.2014.09.003> (2014).
46. Ballestar, E., Yusufzai, T. M. & Wolffe, A. P. Effects of Rett syndrome mutations of the methyl-CpG binding domain of the transcriptional repressor MeCP2 on selectivity for association with methylated DNA. *Biochemistry* **39**, 7100–7106 (2000).
47. Ghosh, R. P., Horowitz-Scherer, R. A., Nikitina, T., Gierasch, L. M. & Woodcock, C. L. Rett syndrome-causing mutations in human MeCP2 result in diverse structural changes that impact folding and DNA interactions. *J Biol Chem* **283**, 20523–20534, <https://doi.org/10.1074/jbc.M803021200> (2008).
48. Heckman, L. D., Chahrouh, M. H. & Zoghbi, H. Y. Rett-causing mutations reveal two domains critical for MeCP2 function and for toxicity in MECP2 duplication syndrome mice. *Elife* **3**, <https://doi.org/10.7554/eLife.02676> (2014).
49. Tillotson, R. *et al.* Radically truncated MeCP2 rescues Rett syndrome-like neurological defects. *Nature* **550**, 398–401, <https://doi.org/10.1038/nature24058> (2017).
50. Weaving, L. S. *et al.* Effects of MECP2 mutation type, location and X-inactivation in modulating Rett syndrome phenotype. *Am J Med Genet A* **118A**, 103–114, <https://doi.org/10.1002/ajmg.a.10053> (2003).
51. Chen, R. *et al.* Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. *Nat Biotechnol* **34**, 531–538, <https://doi.org/10.1038/nbt.3514> (2016).
52. Bhanushali, A. A., Mandsaurwala, A. & Das, B. R. Homozygous c.1160C>T (P38L) in the MECP2 gene in a female Rett syndrome patient. *J Clin Neurosci* **25**, 127–129, <https://doi.org/10.1016/j.jocn.2015.08.040> (2016).
53. Zappella, M., Meloni, I., Longo, I., Hayek, G. & Renieri, A. Preserved speech variants of the Rett syndrome: molecular and clinical analysis. *Am J Med Genet* **104**, 14–22, <https://doi.org/10.1002/ajmg.10005> (2001).
54. Cheadle, J. P. *et al.* Long-read sequence analysis of the MECP2 gene in Rett syndrome patients: correlation of disease severity with mutation type and location. *Hum Mol Genet* **9**, 1119–1129, <https://doi.org/10.1093/hmg/9.7.1119> (2000).

Acknowledgements

The authors would like to thank the Mutalyzer team for support and feedback, Henk van Kranen for support in liftover of ancient genetic variant descriptions, and Eric Smeets for collection of the Maastricht Rett dataset. This work was funded by ELIXIR (funded by the European Commission within the Research Infrastructures programme of Horizon 2020), the research infrastructure for life-science data (MolData2). FE and LC were also funded by The Dutch Rett Syndrome Foundation (Stichting Terre). CE, AJ, RK, AV, SCG, MB, MRi and MR also received funding from EXCELERATE (H2020, Grant No. 676559). AJ, RK, MR, MB, and SCG also received funding from RD-Connect, European Union Seventh Framework Programme (FP7/2007–2013, Grant No. 305444). FE, CE, AJ, RK, MR, MB, and SCG received funding from the European Union's Horizon 2020 Research and Innovation Program under grant agreement EJP RD N°825575. RK was also funded by NWO in project VWData (grant no. 400.17.605) and BBMRI-NL (NWO, National Roadmap for Large-Scale Research Facilities, grant no. 184.033.111). AV and SCG also received funding from INB Grant (Grant No. PT17/0009/0001 - ISCIII-SGEFI / ERDF).

Author contributions

F.E. original idea, study design, data acquisition, data analysis. A.J. study design, data analysis, data deposition. MRi data analysis (*MECP2* domain analysis). M.B. data analysis (VEP analysis). R.K. data analysis, data deposition, code development for FAIRification. J.F.J.L. critical review, contributions to nomenclature and gene variant standards. E.L.W. critical review, contributions to data structure and metadata standards. A.V. critical review. M.R.o. critical review (FAIRification and metadata). S.C.G. study design, critical review. L.M.G.C. data acquisition (Maastricht dataset), critical review. C.T.E. study design, critical review. All authors contributed in writing and reviewing of the study.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41597-020-00794-7>.

Correspondence and requests for materials should be addressed to F.E.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

The Creative Commons Public Domain Dedication waiver <http://creativecommons.org/publicdomain/zero/1.0/> applies to the metadata files associated with this article.

© The Author(s) 2021