863 different causes of Rett syndrome and lessons learned from data integration.

Friederike Ehrhart[1], Annika Jacobsen[2], Maria Rigau[3], Mattia Bosio[4], Rajaram Kaliyaperumal[5], Jeroen Laros[5], Egon Willighagen[1], Alfonso Valencia[6], Marco Roos[2], Salvador Capella-Gutierrez[3], Curfs Leopold[1], and Chris Evelo[1,1]

[1]Maastricht University
[2]Leiden University Medical Centre
[3]Barcelona Supercomputing Center
[4]Centre de Regulacio Genomica
[5]Leiden University Medical Center
[6]Barcelona Supercomputing Centre (BSC-CNS)

April 28, 2020

Abstract

Rett syndrome (RTT) is a rare neurological disorder mostly caused by a genetic variation in MECP2. Various RTT causing and benign variants in MECP2 have been identified and due to the advent of sequencing in clinical diagnosis new variants are identified daily. 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. 19 variants were identified, which were annotated as both, disease-causing and benign suggesting that there are more disease-causing factors than a mutation contributing to the disease development.

Keywords

genotype-phenotype databases, FAIR data, genetic variants, rare diseases, data integration, Rett syndrome

Introduction

Rett syndrome (RTT) is a rare neurological disorder first described in 1956 by Andreas Rett occurring predominantly in females (Rett, 1966). 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) (Amir et al., 1999; Percy et al., 2007). Function affecting variations in several other genes can cause a RTT phenotype whereas several of these are

involved directly in the up or downstream pathway of MECP2 or in the same biological processes (Ehrhart, Sangani, & Curfs, 2018; Lopes et al., 2016; Lucariello et al., 2016; Sajan et al., 2017; Vidal et al., 2017).

The disorder usually undergoes a development in four stages. In the first stage, it is typical that pre- and postnatal development are almost normal. However, in stage two, at the age of about 6 - 18 months, deceleration and stop of motoric and communication learning becomes apparent. In the third stage, patients are usually stable and typical phenotypes include moderate to severe intellectual disability, lack of motoric and (oral) communication skills, abnormal breathing patterns, sleep problems, stereotypic movements (hand wringing). Further, due to dystonia they often develop scoliosis (Neul et al., 2010). During development, the patients often appear to have autistic features due to the lack of communication skills. This is the reason why the disorder is often misclassified within the autism spectrum. In stage four, the motoric abilities continue on slowly decreasing while social and communication skills improve. The spectrum and development of RTT patients' phenotype was investigated in several large natural history studies (Percy et al., 2010; Weaving et al., 2003). The phenotype severity is thought to vary generally due to X-inactivation, mosaicism, severity of the variation (loss of function vs. impaired function), genetic background ((Pizzo et al., 2018) 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 activity (Nan et al., 1998). Thus, MECP2 represses transcription on the level of chromatin organization. MECP2 has several phosphorylation sites that when phosphorylated, e.g., after an incoming electric signal in a neuron, releases the DNA and allows gene transcription (Ebert et al., 2013; Tao et al., 2009). As MECP2 regulates the expression of many genes, the molecular downstream effects are very broad (Ehrhart et al., 2016; Liyanage & Rastegar, 2014). Several meta studies 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 cytoskeleton (Bedogni et al., 2014; F Ehrhart et al., 2018; Shovlin & Tropea, 2018).

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) (Adams, McBryant, Wade, Woodcock, & Hansen, 2007). Ballestar and coworkers found that *MECP2* variations that slightly decrease the specific recognition of the binding site on DNA are able to cause RTT (Ballestar et al., 2005). The majority of RTT causing missense variations are found in the methyl-DNA binding domain, but RTT causing variations have been found in all parts of the protein (Christodoulou, Grimm, Maher, & Bennetts, 2003). Some studies have found a distinctive correlation of phenotype severity and variation type (Neul et al., 2008), while others found a rather small or insignificant correlation (Amir et al., 2000; Auranen et al., 2001; Huppke, Laccone, Kramer, Engel, & Hanefeld, 2000; Nielsen et al., 2001).

Due to the rareness of RTT (prevalence about 1:10.000 (Laurvick et al., 2006)), 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 RTT (Townend et al., 2018). 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 workflow (Jacobsen et al., 2020) 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 CSV distribution is also shared on Figshare (see DOI in results). To our knowledge, the combined data created and used in this study is the largest collection on disease causing and benign *MECP2* variations available at this moment.

Materials and Methods

Workflow of genetic variant data integration

Data selection and retrieval

In a recent study (Townend et al., 2018), we identified 13 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: ClinVar (Landrum et al., 2016),*https://www.ncbi.nlm.nih.gov/clinvar/*, DECIPHER (Firth et al., 2009),*https://decipher.sanger.ac.uk/*, EVA (*http://www.ebi.ac.uk*), EVS (*http://evs.gs.washington.edu*), ExAC (Lek et al., 2016),*http://exac.broadinstitute.org/*, KMD (*https://kmd.nih.go.kr*), LOVD (Fokkema et al., 2011), MECP2 collection:*https://databases.lovd.nl/shared/genes/MECP2*), and RettBASE (Krishnaraj, Ho, & Christodoulou, 2017),*http://mecp2.chw.edu.au/*. Additionally, an anonymized dataset from local RTT patients were 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 (Townend et al., 2018). Figure 1 shows the data processing (step 1-3) and analysis (step 4) workflow of this study.

Best place for figure 1

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/*] (Wildeman, van Ophuizen, den Dunnen, & Taschner, 2008). 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), 2. benign (verified by finding them in a healthy control subject), and 3. unknown evidence (only pathogenicity prediction scores provided by database). 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 principles (Wilkinson et al., 2016). The data was made machine-readable (in RDF format) using a semantic data model (see below) and a general-purpose FAIRifier tool (Thompson, Burger, Kaliyaperumal, Roos, & Bonino da Silva Santos, 2020) based on the OpenRefine data cleaning and wrangling tool (*http://openrefine.org/*) and an RDF plugin (*https://github.com/stkenny/grefine-rdf-extension*). Similarly, machine-readable metadata (information about the data) was generated using the Metadata Editor (Thompson et al., 2020). The machine-readable metadata was made available on a FAIR Data Point ((Bonino da Silva Santos et al., 2016) 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.*

We applied and extended the semantic data model of a genetic variant described in (Horst; et al., 2015) 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

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 Cytoscape (Shannon et al., 2003) 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://github.com/mbosio85/HGVSparse*. Afterwards, we proceeded to annotate variants with Ensembl Variant Effect Predictor, VEP, (McLaren et al., 2016) v94 using the GRCh38 assembly, selecting all available features, plus optional plugins to estimate variant pathogenicity (i.e., PolyPhen (Adzhubei et al., 2010), SIFT (Sim et al., 2012), MetaLR (Dong et al., 2015), CADD (Kircher et al., 2014), FATHMM-MKL (Shihab et al., 2015) from dbNSFP and dbscSNV scores (Liu, Wu, Li, & Boerwinkle, 2016)) both in coding and splicing regions.

The resulting VEP annotated data was processed with R scripts, available at*https://gitlab.bsc.es/mbosio85/rtt_summary_plot*, 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 (Lombardi, Baker, & Zoghbi, 2015)). This allowed us to make a finer characterization of differential distribution of RTT causing and benign variants across MECP2 domains.

Results

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 included in further analysis. The same problem occurred with the annotation of diagnosis and/or phenotypes. As described before (Townend et al., 2018) only a few databases link original diagnostic information to the genetic information. If this information was given different formats or definitions were used.

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.153282026G>A to NC_000023.11:g.154016575G>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).

Best place for Table 1

Size and content of the FAIR dataset

Number of disease causing and benign *MECP2* genetic variants available

Based on the 13 genotype-phenotype databases identified in (Townend et al., 2018), 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 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 also 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. 19 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. These processed datasets are available as csv on gdrive (*link*). Out of the 10,968 curated *MECP2* variations only 11 occur in more than 1% of all database entries, and these account for 53.7% of all database entries (data not shown).

Best place for Table 2

The 863 unique RTT causing variations are distributed over 4,573 database entries. Also here, only 12 variations are found in more than 1% of all database entries (Table 3) and these 12 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 hotspots (Wan et al., 1999). These eight *MECP2* hotspot variations contribute to 49.7% of all *MECP2* variation entries. The most abundant *MECP2*variation in this dataset is NC_000023.11:g.g.154031355G>A (NM_004992.3:c.473C>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.

Best place for Table 3

Distribution of the variants across databases

Table 2 shows the number of unique MECP2 variations for each investigated database. The relative number of unique MECP2 variations in each of the databases differ. The number of how many variations in one database are unique can also give 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 from where this 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 1 (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.

Best place for figure 2

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 results from VEP (see Methods) from six descriptive features (Figure 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 GnomAD (Lek et al., 2016) (i.e., GnomAD_AF).

We split variants by benign, both and RTT causing, as we identified a subset of 19 variants appearing in both datasets. Overall, we see expected results: the RTT causing variants were found to be in positions significantly more conserved than the benign or both variants (Figure 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 (Figure 3, GnomAD_AF). Analysis of the obtained estimation of pathogenicity from multiple scores (Figure 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 characterization of the both group is located in three of five predictions between the benign and RTT causing, and in two of five closer to the benign group.

Best place for figure 3

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 Figure 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.

Best place for figure 4

Considering protein consequences, in 1350 cases, which is about half of the RTT causing single nucleotide mutations, are protein truncating variants, changing an Arginine into a stop codon. Also frequently, Arginine is changed into a Cysteine (533) or Tryptophan (179) which are major changes considering 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.

Discussion

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 within six months from 4738 (March 2018, (Townend et al., 2018)) to 4757 (November 2018) to 4806 (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 (Das, Raha, Sanghavi, Maitra, & Udani, 2013; Inui et al., 2001) or large populations (e.g. (Bienvenu et al., 2002; Percy et al., 2010)) that have published their data in the previous years. (Bienvenu et al., 2002) 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 participants (Percy et al., 2010)) 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 (Krishnaraj et al., 2017) 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. (Percy et al., 2007) 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 before (Wang, Tang, Lai, & Zhang, 2014).

Best place for Table 5

In our integrated dataset most pathogenic mutations in *MECP2*occur in the methyl-DNA or transcription repressor binding domain. This has been found and confirmed before (Ballestar, Yusufzai, & Wolffe, 2000; Ghosh, Horowitz-Scherer, Nikitina, Gierasch, & Woodcock, 2008; Heckman, Chahrour, & Zoghbi, 2014; Krishnaraj et al., 2017). The functionality of the methyl-DNA binding domain is reported to be extremely sensitive for changes (Ballestar et al., 2000). 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 MECP2 (Tillotson et al., 2017). 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 has 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 (Figure 3). Possible explanations why a variant can be disease causing in one individual and 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 database. Also, X inactivation patterns (Weaving et al., 2003) and genetic background related to other participating genes in *MECP2* related pathways (Pizzo et al., 2018) influence the severity of a rare monogenic (X-linked) disease and can possibly even save individuals with a documented pathogenic variation from disease development (Chen et al., 2016). In principle, patients could also have an unreported second mutation that could cause the effect either alone or through epistatic interaction.

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.1160C>T (P387L, NC_000023.11:g.154030668G>A) variation from a healthy (!) father (Bhanushali, Mandsaurwala, & Das, 2016). Exactly this variant we found 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. To unravel the different influences of*MECP2* variations in the context of an individual patient, we need to evaluate how genetic background 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 also a significant number of patients (54 in our integrated dataset) whose *MECP2* gene carries more than one variation. In these cases, we presume that the disease is caused by one (pathogenic)*MECP2* variation while the other variation *can* be benign if it occurs alone. Other possibilities are positive or negative epistatic effects if these variants occur on the same allele. All of these possibilities may lead to wrong classification of variants.

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.

Next step: automatization

The first step in the integration of genetic variation data across multiple resources was a time consuming study, which included a lot of manual data acquisition and curation. Additionally, analysing *MECP2* variations as the causative entities in RTT was the leading example in this study but this method should be available and applicable for any other gene, too. The next step therefore would be to automate this process. This can only be efficient and robust when the data resources provide an interface by which machines can predict how to find, access, and use their data. This interface is complementary to the specific features that each source provides for its users. FAIR principles provide useful guidance here: they do not prescribe any specific implementation, but do enforce a higher level of transparency for machines. In other words, the feasibility and quality of automation depends on the resources being FAIR. Consequently, a workflow can be developed and used as a tool to retrieve the known disease causing, benign, or other variants of yet unknown significance for any gene. The FAIRification of the databases is a process, which has already started and will hopefully continue to support efficient data science. Interesting for the implementation of FAIR principles are activities towards new standards for processing variant data, such as by the genetic variant workstream of the Global Alliance for Genomics and Health as well as its GA4GH Beacon project, which allows cross database search for variants. Generic FAIR services and service specifications, such as produced in FAIRtrain, FAIRsFAIR, and EOSC-Life (e.g. FAIRsharing.org and the FAIR Data Point specification (Bonino da Silva Santos et al., 2016) https://github.com/FAIRDataTeam/FAIRDataPoint-Spec, enable the general task of identifying and visiting interoperable RDF) from restricted access databases.

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, Eric Smeets for collection of the Maastricht Rett dataset, and Mark Wilkinson for information about automatization.

Conflict of Interest Statement

The authors report no conflict of interest.

References

Adams, V. H., McBryant, S. J., Wade, P. A., Woodcock, C. L., & Hansen, J. C. (2007). Intrinsic disorder and autonomous domain function in the multifunctional nuclear protein, MeCP2. *J Biol Chem, 282* (20), 15057-15064. doi:10.1074/jbc.M700855200

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., . . . Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nat Methods, 7* (4), 248-249. doi:10.1038/nmeth0410-248

Amir, R. E., Van den Veyver, I. B., Schultz, R., Malicki, D. M., Tran, C. Q., Dahle, E. J., . . . Zoghbi, H. Y. (2000). Influence of mutation type and X chromosome inactivation on Rett syndrome phenotypes. *Ann Neurol, 47* (5), 670-679.

Amir, R. E., Van den Veyver, I. B., Wan, M., Tran, C. Q., Francke, U., & Zoghbi, H. Y. (1999). Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nat Genet, 23* (2), 185-188. doi:10.1038/13810

Auranen, M., Vanhala, R., Vosman, M., Levander, M., Varilo, T., Hietala, M., . . . Jarvela, I. (2001). MECP2 gene analysis in classical Rett syndrome and in patients with Rett-like features. *Neurology, 56* (5), 611-617. doi:10.1212/wnl.56.5.611

Ballestar, E., Ropero, S., Alaminos, M., Armstrong, J., Setien, F., Agrelo, R., . . . Esteller, M. (2005). The impact of MECP2 mutations in the expression patterns of Rett syndrome patients. *Hum Genet, 116* (1-2), 91-104. doi:10.1007/s00439-004-1200-0

Ballestar, E., Yusufzai, T. M., & Wolffe, A. P. (2000). 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* (24), 7100-7106.

Bedogni, F., Rossi, R. L., Galli, F., Cobolli Gigli, C., Gandaglia, A., Kilstrup-Nielsen, C., & Landsberger, N. (2014). Rett syndrome and the urge of novel approaches to study MeCP2 functions and mechanisms of action. *Neurosci Biobehav Rev, 46 Pt 2* , 187-201. doi:10.1016/j.neubiorev.2014.01.011

Bhanushali, A. A., Mandsaurwala, A., & Das, B. R. (2016). Homozygous c.1160C>T (P38L) in the MECP2 gene in a female Rett syndrome patient. *J Clin Neurosci, 25* , 127-129. doi:10.1016/j.jocn.2015.08.040

Bienvenu, T., Villard, L., De Roux, N., Bourdon, V., Fontes, M., Beldjord, C., . . . French Consortium for, M. G. A. (2002). Spectrum of MECP2 mutations in Rett syndrome. *Genet Test, 6* (1), 1-6. doi:10.1089/109065702760093843

Bonino da Silva Santos, L. O., Wilkinson, M., Kuzniar, A., Kaliyaperumal, R., Thompson, M., Dumontier, M., & Burger, K. (2016).*Enterprise Interoperability in the Digitized and Networked Factory of the Future - FAIR Data Points Supporting Big Data Interoperability.*(M. Zelm;, G. Doumeingts;, & J. P. Mendonça. Eds.): ISTE Press.

Cheadle, J. P., Gill, H., Fleming, N., Maynard, J., Kerr, A., Leonard, H., . . . Clarke, A. (2000). 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* (7), 1119-1129. doi:10.1093/hmg/9.7.1119

Chen, R., Shi, L., Hakenberg, J., Naughton, B., Sklar, P., Zhang, J., . . . Friend, S. H. (2016). Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. *Nat Biotechnol, 34* (5), 531-538. doi:10.1038/nbt.3514

Christodoulou, J., Grimm, A., Maher, T., & Bennetts, B. (2003). RettBASE: The IRSA MECP2 variation database-a new mutation database in evolution. *Hum Mutat, 21* (5), 466-472. doi:10.1002/humu.10194

Das, D. K., Raha, S., Sanghavi, D., Maitra, A., & Udani, V. (2013). Spectrum of MECP2 gene mutations in a cohort of Indian patients with Rett syndrome: report of two novel mutations. *Gene, 515* (1), 78-83. doi:10.1016/j.gene.2012.11.024

Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., & Liu, X. (2015). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies.*Hum Mol Genet, 24* (8), 2125-2137. doi:10.1093/hmg/ddu733

Ebert, D. H., Gabel, H. W., Robinson, N. D., Kastan, N. R., Hu, L. S., Cohen, S., . . . Greenberg, M. E. (2013). Activity-dependent phosphorylation of MeCP2 threonine 308 regulates interaction with NCoR.*Nature, 499* (7458), 341-345. doi:10.1038/nature12348

Ehrhart, F., Coort, S. L., Cirillo, E., Smeets, E., Evelo, C. T., & Curfs, L. M. (2016). Rett syndrome - biological pathways leading from MECP2 to disorder phenotypes. *Orphanet J Rare Dis, 11* (1), 158. doi:10.1186/s13023-016-0545-5

Ehrhart, F., Coort, S. L., Eijssen, L., Cirillo, E., Smeets, E. E., Bahram Sangani, N., . . . Curfs, L. M. G. (2018). Integrated analysis of human transcriptome data for Rett syndrome finds a network of involved genes. *bioRxiv* . doi:10.1101/274258

Ehrhart, F., Sangani, N. B., & Curfs, L. M. G. (2018). Current developments in the genetics of Rett and Rett-like syndrome. *Curr Opin Psychiatry, 31* (2), 103-108. doi:10.1097/YCO.0000000000000389

Firth, H. V., Richards, S. M., Bevan, A. P., Clayton, S., Corpas, M., Rajan, D., . . . Carter, N. P. (2009). DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am J Hum Genet, 84* (4), 524-533. doi:10.1016/j.ajhg.2009.03.010

Fokkema, I. F., Taschner, P. E., Schaafsma, G. C., Celli, J., Laros, J. F., & den Dunnen, J. T. (2011). LOVD v.2.0: the next generation in gene variant databases. *Hum Mutat, 32* (5), 557-563. doi:10.1002/humu.21438

Ghosh, R. P., Horowitz-Scherer, R. A., Nikitina, T., Gierasch, L. M., & Woodcock, C. L. (2008). Rett syndrome-causing mutations in human MeCP2 result in diverse structural changes that impact folding and DNA interactions. *J Biol Chem, 283* (29), 20523-20534. doi:10.1074/jbc.M803021200

Heckman, L. D., Chahrour, M. H., & Zoghbi, H. Y. (2014). Rett-causing mutations reveal two domains critical for MeCP2 function and for toxicity in MECP2 duplication syndrome mice. *Elife, 3* . doi:10.7554/eLife.02676

Horst;, E. v. d., Kaliyaperumal;, R., Tatum;, Z., Thompson;, M., Schultes;, E., Mina;, E., . . . Hoen., P. A. C. t. (2015). 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* .

Huppke, P., Laccone, F., Kramer, N., Engel, W., & Hanefeld, F. (2000). Rett syndrome: analysis of MECP2 and clinical characterization of 31 patients. *Hum Mol Genet, 9* (9), 1369-1375. doi:10.1093/hmg/9.9.1369

Inui, K., Akagi, M., Ono, J., Tsukamoto, H., Shimono, K., Mano, T., . . . Okada, S. (2001). Mutational analysis of MECP2 in Japanese patients with atypical Rett syndrome. *Brain Dev, 23* (4), 212-215. doi:10.1016/s0387-7604(01)00197-8

Jacobsen, A., Kaliyaperumal, R., Bonino da Silva Santos, L., Mons, B., Schultes, E., Roos, M., & Thompson, M. (2020). A Generic Workflow for the Data FAIRification Process *Data Intelligence, 2* (1-2), 56-65. doi:10.1162/dint_a_00028

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet, 46* (3), 310-315. doi:10.1038/ng.2892

Krishnaraj, R., Ho, G., & Christodoulou, J. (2017). RettBASE: Rett syndrome database update. *Hum Mutat, 38* (8), 922-931. doi:10.1002/humu.23263

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., . . . Maglott, D. R. (2016). ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res, 44* (D1), D862-868. doi:10.1093/nar/gkv1222

Laurvick, C. L., de Klerk, N., Bower, C., Christodoulou, J., Ravine, D., Ellaway, C., . . . Leonard, H. (2006). Rett syndrome in Australia: a review of the epidemiology. *J Pediatr, 148* (3), 347-352. doi:10.1016/j.jpeds.2005.10.037

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., . . . Exome Aggregation, C. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature, 536* (7616), 285-291. doi:10.1038/nature19057

Liu, X., Wu, C., Li, C., & Boerwinkle, E. (2016). dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Hum Mutat, 37* (3), 235-241. doi:10.1002/humu.22932

Liyanage, V. R., & Rastegar, M. (2014). Rett syndrome and MeCP2.*Neuromolecular Med, 16* (2), 231-264. doi:10.1007/s12017-014-8295-9

Lombardi, L. M., Baker, S. A., & Zoghbi, H. Y. (2015). MECP2 disorders: from the clinic to mice and back. *J Clin Invest, 125* (8), 2914-2923. doi:10.1172/JCI78167

Lopes, F., Barbosa, M., Ameur, A., Soares, G., de Sa, J., Dias, A. I., . . . Maciel, P. (2016). Identification of novel genetic causes of Rett syndrome-like phenotypes. *J Med Genet, 53* (3), 190-199. doi:10.1136/jmedgenet-2015-103568

Lucariello, M., Vidal, E., Vidal, S., Saez, M., Roa, L., Huertas, D., . . . Esteller, M. (2016). Whole exome sequencing of Rett syndrome-like patients reveals the mutational diversity of the clinical phenotype.*Hum Genet, 135* (12), 1343-1354. doi:10.1007/s00439-016-1721-3

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., . . . Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biol, 17* (1), 122. doi:10.1186/s13059-016-0974-4

Nan, X., Ng, H. H., Johnson, C. A., Laherty, C. D., Turner, B. M., Eisenman, R. N., & Bird, A. (1998). Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex.*Nature, 393* (6683), 386-389. doi:10.1038/30764

Neul, J. L., Fang, P., Barrish, J., Lane, J., Caeg, E. B., Smith, E. O., . . . Glaze, D. G. (2008). Specific mutations in methyl-CpG-binding protein 2 confer different severity in Rett syndrome. *Neurology, 70* (16), 1313-1321. doi:10.1212/01.wnl.0000291011.54508.aa

Neul, J. L., Kaufmann, W. E., Glaze, D. G., Christodoulou, J., Clarke, A. J., Bahi-Buisson, N., . . . Rett-Search, C. (2010). Rett syndrome: revised diagnostic criteria and nomenclature. *Ann Neurol, 68* (6), 944-950. doi:10.1002/ana.22124

Nielsen, J. B., Henriksen, K. F., Hansen, C., Silahtaroglu, A., Schwartz, M., & Tommerup, N. (2001). 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* (3), 178-184. doi:10.1038/sj.ejhg.5200600

Percy, A. K., Lane, J. B., Childers, J., Skinner, S., Annese, F., Barrish, J., . . . MacLeod, P. (2007). Rett syndrome: North American database. *J Child Neurol, 22* (12), 1338-1341. doi:10.1177/0883073807308715

Percy, A. K., Neul, J. L., Glaze, D. G., Motil, K. J., Skinner, S. A., Khwaja, O., . . . Barnes, K. (2010). Rett syndrome diagnostic criteria: lessons from the Natural History Study. *Ann Neurol, 68* (6), 951-955. doi:10.1002/ana.22154

Pizzo, L., Jensen, M., Polyak, A., Rosenfeld, J. A., Mannik, K., Krishnan, A., . . . Girirajan, S. (2018). Rare variants in the genetic background modulate cognitive and developmental phenotypes in individuals carrying disease-associated variants. *Genet Med* . doi:10.1038/s41436-018-0266-3

Rett, A. (1966). [On a unusual brain atrophy syndrome in hyperammonemia in childhood]. *Wien Med Wochenschr, 116* (37), 723-726.

Sajan, S. A., Jhangiani, S. N., Muzny, D. M., Gibbs, R. A., Lupski, J. R., Glaze, D. G., . . . Neul, J. L. (2017). Enrichment of mutations in chromatin regulators in people with Rett syndrome lacking mutations in MECP2. *Genet Med, 19* (1), 13-19. doi:10.1038/gim.2016.42

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., . . . Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res, 13* (11), 2498-2504. doi:10.1101/gr.1239303

Shihab, H. A., Rogers, M. F., Gough, J., Mort, M., Cooper, D. N., Day, I. N., . . . Campbell, C. (2015). An integrative approach to predicting the functional effects of non-coding and coding sequence variation.*Bioinformatics, 31* (10), 1536-1543. doi:10.1093/bioinformatics/btv009

Shovlin, S., & Tropea, D. (2018). Transcriptome level analysis in Rett syndrome using human samples from different tissues. *Orphanet J Rare Dis, 13* (1), 113. doi:10.1186/s13023-018-0857-8

Sim, N. L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., & Ng, P. C. (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res, 40* (Web Server issue), W452-457. doi:10.1093/nar/gks539

Tao, J., Hu, K., Chang, Q., Wu, H., Sherman, N. E., Martinowich, K., . . . Sun, Y. E. (2009). Phosphorylation of MeCP2 at Serine 80 regulates its chromatin association and neurological function. *Proc Natl Acad Sci U S A, 106* (12), 4882-4887. doi:10.1073/pnas.0811648106

Thompson, M., Burger, J., Kaliyaperumal, R., Roos, M., & Bonino da Silva Santos, L. (2020). Making FAIR easy with FAIR tools: From creolization to convergence. *Data Intelligence, 2* , 87–95. doi:10.1162/dint_a_00031

Tillotson, R., Selfridge, J., Koerner, M. V., Gadalla, K. K. E., Guy, J., De Sousa, D., . . . Bird, A. (2017). Radically truncated MeCP2 rescues Rett syndrome-like neurological defects. *Nature, 550* (7676), 398-401. doi:10.1038/nature24058

Townend, G. S., Ehrhart, F., van Kranen, H. J., Wilkinson, M., Jacobsen, A., Roos, M., . . . Curfs, L. M. G. (2018). MECP2 variation in Rett syndrome-An overview of current coverage of genetic and phenotype data within existing databases. *Hum Mutat, 39* (7), 914-924. doi:10.1002/humu.23542

Vidal, S., Brandi, N., Pacheco, P., Gerotina, E., Blasco, L., Trotta, J. R., . . . Rett Working, G. (2017). The utility of Next Generation Sequencing for molecular diagnostics in Rett syndrome. *Sci Rep, 7* (1), 12288. doi:10.1038/s41598-017-11620-3

Wan, M., Lee, S. S., Zhang, X., Houwink-Manville, I., Song, H. R., Amir, R. E., . . . Francke, U. (1999). Rett syndrome and beyond: recurrent spontaneous and familial MECP2 mutations at CpG hotspots. *Am J Hum Genet, 65* (6), 1520-1529. doi:10.1086/302690

Wang, J., Tang, J., Lai, M., & Zhang, H. (2014). 5-Hydroxymethylcytosine and disease. *Mutat Res Rev Mutat Res, 762* , 167-175. doi:10.1016/j.mrrev.2014.09.003

Weaving, L. S., Williamson, S. L., Bennetts, B., Davis, M., Ellaway, C. J., Leonard, H., . . . Christodoulou, J. (2003). Effects of MECP2 mutation type, location and X-inactivation in modulating Rett syndrome phenotype. *Am J Med Genet A, 118A* (2), 103-114. doi:10.1002/ajmg.a.10053

Wildeman, M., van Ophuizen, E., den Dunnen, J. T., & Taschner, P. E. (2008). Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. *Hum Mutat, 29* (1), 6-13. doi:10.1002/humu.20654

Zappella, M., Meloni, I., Longo, I., Hayek, G., & Renieri, A. (2001). Preserved speech variants of the Rett syndrome: molecular and clinical analysis. *Am J Med Genet, 104* (1), 14-22. doi:10.1002/ajmg.10005

Tables

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.

Phenotype annotation format	Database	Number of *MECP2* or RTT variations	Number of variations m
Phenotype	DECIPHER	34	25
Diagnosis	Maastricht Rett dataset	429	428
	ClinVar	1,134	743
	RettBase	4,705	3,986

	KMD	35	35
Pathogencity scores	EVS	190	190
	LOVD	808	808
	EVA	4,226	4,226
	ExAC	599	599
TOTAL	TOTAL	12,158	11,040

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

Database	Number of total *MECP2* variation entries	Number of unique *MECP2* variations #	Number of u % of total *M*
EVA	4,226	4,192	99.2
LOVD	808	802	99.3
RettBase	4,705	740	15.7
ExAC	599	599	100.0
ClinVar	1,134	716	63.1
EVS	190	95	50.0
Maastricht Rett dataset	429	68	15.9
KMD	35	35	100.0
DECIPHER	34	23	67.6

Table 3: Most abundant RTT causing variants in this study.

Genomic position +	count	%	cDNA ++ and protein change §	Effect and previous reports
g.154031355G>A	463	10.1	c.473C>T, p.(Thr158Met) ¶	Missense variation (Bienvenu et al., 2002; Krishnaraj et al., 2017; Percy et al., 2010)
g.154031326G>A	409	8.9	c.502C>T, p.(Arg168*) ¶	Nonsense variation, leading to truncation (Bienvenu et al., 2002; Krishnaraj et al., 2017; Percy et al., 2010)
g.154031065G>A	345	7.5	c.763C>T, p.(Arg255*) ¶	Nonsense variation, leading to truncation (Bienvenu et al., 2002; Krishnaraj et al., 2017; Percy et al., 2010)

g.154031020G>A	309	6.8	c.808C>T, p.(Arg270*) ¶	Nonsense variation, leading to truncation (Bienvenu et al., 2002; Krishnaraj et al., 2017; Percy et al., 2010)
g.154030948G>A	281	6.1	c.880C>T, p.(Arg294*) ¶	Nonsense variation, leading to truncation (Krishnaraj et al., 2017; Percy et al., 2010)
g.154030912G>A	279	6.1	c.916C>T, p.(Arg306Cys) ¶	Missense variation (Bienvenu et al., 2002; Krishnaraj et al., 2017; Percy et al., 2010)
g.154031431G>A	249	5.4	c.397C>T, p.(Arg133Cys) ¶	Missense variation (Krishnaraj et al., 2017; Percy et al., 2010; Zappella, Meloni, Longo, Hayek, & Renieri, 2001)
g.154032268G>A	161	3.5	c.316C>T, p.(Arg106Trp) ¶	Missense variation (Krishnaraj et al., 2017; Percy et al., 2010)
g.154031373G>C	80	1.7	c.455C>G, p.(Pro152Arg)	Missense variation (Cheadle et al., 2000)
g.154031022delC	67	1.5	c.806delG, p.(Gly269fs)	Frameshift deletion leading to missense (Das et al., 2013)
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

RefSeq + NC_000023.11, ++ NM_004992.3, § NP_004983.1

¶ one of the eight hotspot variations (Wan et al., 1999)

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

	Domain length (% of total)	**RTT causing** % of missense variations per region	**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	Conserved regions		93.6	16.3

Table 5: Comparison of most abundant *MECP2* variations in different studies.

Sample size and citation	Variations mentioned in studies (abundance in % if known)	V
4573 variations annotated with RTT causing (this study)	Thr158Met (10.1)	A
301 RTT patients (Bienvenu et al., 2002)	Thr158Met (7.8)	A
819 RTT patients (Percy et al., 2010)	Thr158Met (11.0)	A
RettBase 4668 total entries (Krishnaraj et al., 2017)	Thr158Met	A

RefSeq: NP_004983.1:p.

Figure legends

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

Figure 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).

Figure 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.

Figure 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).

Appendices

Supplementary table 1 and 2