Article

Evolutionary analysis of Rett Syndrome-causing proteins and their pathogenic missense point mutations: structural order–disorder, post-translational modifications, evolutionary rates, and interacting proteins

Muhamad Fahmi[1], Gen Yasui[1], Kaito Seki[1], Syouichi Katayama[2], Takako Kaneko-Kawano[2], Tetsuya Inazu[2], Yukihiko Kubota[1,3] and Masahiro Ito[1,3,*]

[1] Graduate School of Life Sciences, Ritsumeikan University; gr0343rp@ed.ritsumei.ac.jp (M.F.), sj0048hh@ed.ritsumei.ac.jp(G.Y.), sj0036kf@ed.ritsumei.ac.jp(K.S.), yukubota@fc.ritsumei.ac.jp(Y.K.), maito@sk.ritsumei.ac.jp(M.I.)

[2] College of Pharmaceutical Sciences, Ritsumeikan University; s-kata@fc.ritsumei.ac.jp(S.K.), takanek@fc.ritsumei.ac.jp(T.K.-K.), tinazu@fc.ritsumei.ac.jp(T.I.)

[3] College of Life Sciences, Ritsumeikan University

* Correspondence: maito@sk.ritsumei.ac.jp (M.I.)

Abstract: Rett syndrome (RTT) is mainly caused by mutations in methyl CpG-binding protein 2, cyclin-dependent kinase-like 5, or forkhead box protein G1. These RTT-causing proteins harbor an intrinsically disordered region (IDR) whose conformation exhibits spatiotemporal heterogeneity, which not only confer versatility to the protein, but also implicates them in diseases. The IDR generally evolves more rapidly than an ordered structure. In this study, we examined the relationship between pathogenic RTT-associated point mutations in RTT-causing proteins and the evolutionary dynamics of sequence features including structural order–disorder, phosphorylation sites, and evolutionary rates. We also analyzed the molecular properties and evolution of proteins that interact with RTT-causing proteins in terms of phylogenetic profiles, tissue specificity, subcellular localization, expression level, and functions. The results indicate that constrained IDRs may function by forming contacts with other regions in the protein sequence causing pathogenic missense mutations likely to arise in the rapidly evolving IDR and affect molecular networks, leading to disease. The results also provide novel insights into the genetic basis for RTT and the evolution of the neocortex in higher vertebrates.

Keywords: Rett Syndrome; Intrinsically disordered region; phylogenetic profile analysis; post-transcriptional modification; methyl-CpG-binding protein 2; cyclin-dependent kinase-like 5; forkhead box protein G1

1. Introduction

Rett syndrome (RTT; OMIM entry #312750) is a rare disease that was first described by Andreas Rett in 1966 [1]. It mainly affects girls aged 6 to 18 months and is characterized by severe neurodevelopmental impairment such as intellectual disability, movement disorder, and epilepsy [2]. Mutations in *methyl CpG-binding protein* (*MECP*)*2*, an X-linked gene involved in the regulation of RNA

splicing and chromatin remodeling, are the predominant cause of classic RTT and are present in > 90% of patients [3, 4]. Atypical cases of this syndrome are associated with mutations in either *cyclin-dependent kinase-like* (*CDKL*)*5*, *forkhead box protein* (*FOXG*)*1*, or undefined genes [5, 6]. Collectively, MeCP2, CDKL5, and FOXG1 are known as RTT-causing proteins. MeCP2 has been determined shown to have a disordered structure by using various experimental methods approaches, the only predicted structures of are available for FOXG1 have only been investigated by predictions [7-9]. In case of CDKL5, structure of amino terminal kinase domain is already identified, but long carboxy terminal tail has not been clarified [10]. These proteins contain polypeptide segments that are unable to fold spontaneously into three-dimensional structures; the so-called intrinsically disordered regions (IDRs) exist as dynamic ensembles of conformations that rapidly interconvert from molten globule (collapsed) to coiled or premolten globules (extended) as a result of relatively flat energy landscapes [11–14].

The different conformations of IDRs and structured regions are dictated by the amino acid sequence; the former generally lack bulky hydrophobic residues [14, 15]. Proteins are composed of either fully structured or fully disordered regions (with the latter referred to as intrinsically disordered proteins [IDPs]) or a combination of the two, which is the case for most eukaryotic proteins [16]. Although protein function has traditionally been elucidated based on a well-defined structure, it is now widely acknowledged that IDRs contribute to diverse functions, which can be classified into six types: entropic chain activity, display site, chaperone, molecular effector, molecular assembler, and molecular scavenger [17–19]. Excluding entropic chain activity, IDRs adopt specific tertiary conformations—at least locally—in order to perform these functions by binding to other proteins, nucleic acids, membranes, and small molecules or respond to changes in their environment that alter their relative free energy landscape [11, 20, 21]. Hence, IDR conformation varies over time—i.e., it exhibits spatiotemporal heterogeneity [22]. Moreover, long IDRs contain more modification sites than fully ordered regions and their conformational flexibility provides more opportunities for displaying these sites [23, 24]. These features explain how proteins with IDRs or IDPs interact with and are tightly regulated by various factors to ensure that appropriate levels of protein are available at the right time to minimize the possibility of inappropriate protein–protein interactions [19]. Thus, altered conformation and availability of proteins with IDRs or IDPs are more likely to be associated with disease states. However, IDRs generally (but not always) evolve more rapidly than ordered structures owing to different accepted point mutation caused by differences of residue composition, intramolecular contacts, and function [25].

Disease-related alleles are introduced into the human population by mutation, are directed in part by random genetic drift, and disappear through purifying selection [26–28]. Restoring *MeCP2* gene function in an animal model abolished the symptoms of RTT, suggesting that the disorder is treatable [29]. In addition to gene therapy, reactivation of inactivated X chromosome is noted as a new therapeutic method [30, 31]. Moreover, growth factor stimulation (e.g. insulin-like growth factor 1) and activation of neurotransmitter pathways (e.g. β2-adrenergic receptor pathway) can partially rescue phenotypes of MeCP2 knockout mice (RTT model mice) [32, 33]. Elucidating the molecular basis for RTT can lead to the development of effective treatments.

Proteins function as part of interaction networks with other proteins, with human-specific protein–protein interactions established through evolution. RettBASE is a point mutation database for RTT that provides reliable prevalence data on deleterious mutations [34]. In this study, we used RettBASE to examine the relationship between pathogenic RTT point mutations and the evolution of structural order–disorder, phosphorylation sites, and evolutionary rates of *MECP2*, *CDKL5*, and *FOXG1* in order to clarify the role of pathogenic mutations in these genes in the development of RTT, and determine how IDRs are involved. In addition, we also analyzed the molecular features and evolution of proteins that interact with RTT-causing proteins based on phylogenetic profile, tissue specificity, subcellular localization, expression level, and function. Our results suggest that

pathogenic RTT missense mutations tend to occur in domain regions and affect ordered residues for CDKL5 and FOXG1, and an intrinsically disordered domain (IDD) which can form intramolecular contacts with other disordered regions in MeCP2 could promote the development of pathogenic RTT by missense mutations in the rapidly evolved residues beyond the domain region. The phylogenetic cluster analysis revealed that the RTT-causing proteins MeCP2, CDKL5, and FOXG1 and their interaction partners may have been acquired during metazoan evolution and play essential roles in neocortical development.

2. Results

2.1. *Structural order–disorder properties of RTT-causing proteins during chordate evolution*

We retrieved 97, 113, and 108 chordates sequences of *MeCP2*, *CDKL5*, and *FOXG1*, respectively, and constructed a heat map of structural order–disorder propensity for each RTT-causing protein according to aligned sequences and taxonomic position in the phylogenetic tree (Supplementary Table S1 and Figure 1). All RTT-causing proteins harbored both ordered and disordered regions; by comparing their distribution to domain and non-domain regions, we found that the catalytic domain and non-domain regions of CDKL5 were ordered and disordered, respectively (Figure 1B). While most regions of MeCP2 were predicted to be disordered, some ordered structures were observed in the methyl-CpG binding domain (MBD) (Figure 1A). Furthermore, FOXG1 showed a varied distribution of ordered–disordered regions corresponding to domain and non-domain regions, with the former predicted to be fully ordered (Figure 1). Despite insertions and deletions were frequently detected in disordered regions, particularly in MeCP2 and FOXG1 (Figure 1A, C), the order-disorder conformation of RTT-causing proteins showed to be stable in chordates, excluding a few conformational transitions of FOXG1 and CDKL5 in mammals and fishes, respectively.

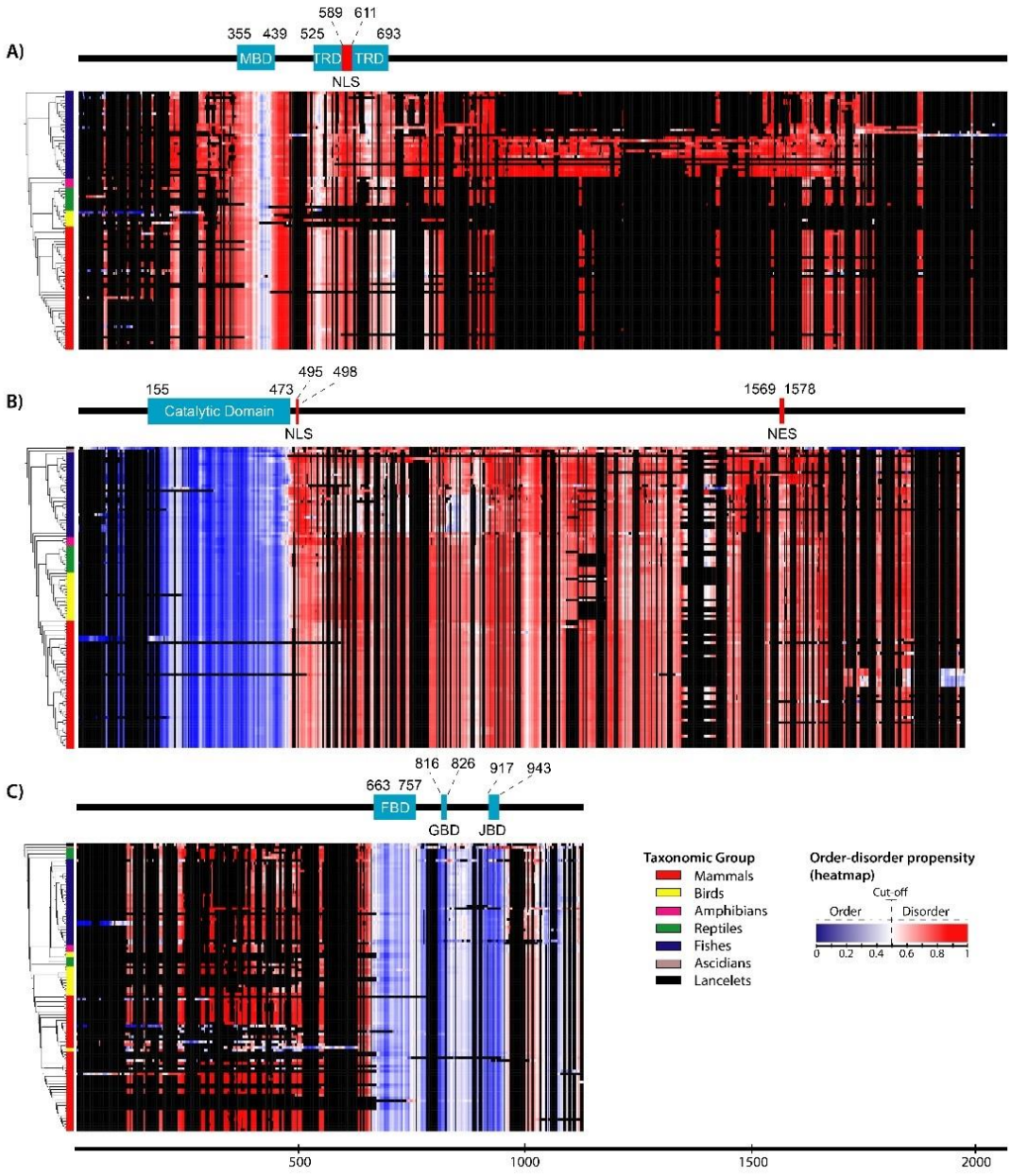

Figure 1. Order–disorder propensity of RTT-causing proteins in chordates. Heat maps of order–disorder propensity were generated according to taxonomic position in the phylogenetic tree (rows) and multiple sequence alignment (columns). The heat maps show a color gradient of blue (ordered) to red (disordered), with white as the boundary between the two and black as gaps. Colored boxes between the trees and heat maps indicate the taxonomic group; and bars above the heat maps indicate domain position in the multiple sequence alignment, with light blue and black areas indicating the domain and absence of a domain, respectively. (**a–c**) Heat maps for MeCP2 (A), CDKL5 (B), and FOXG1 (C) are shown. MBD, TRD, FBD, GBD, JBD, NLS, and NES indicate methyl-CpG binding domain, transcriptional repression domain, forkhead binding domain, Groucho-binding domain, JARID1B binding domain, nuclear localization signal, and nuclear export signal, respectively.

2.2 Rate of evolution per site in RTT-causing proteins

We calculated the evolutionary rate of RTT-causing proteins in chordates to investigate its relationship with structural features and the distribution of pathogenic RTT-causing missense mutations. We used the human sequence as reference and determined standardized evolutionary

rate scores (Z scores), with values greater than or less than zero reflecting evolution at a faster and slower than average rate, respectively (Figure 2 and Supplementary Table S2). The results showed that evolutionary rates per site showed similar patterns in all proteins, with low rates of evolution more commonly observed in domains and ordered regions; some exceptional cases such as the transcriptional repression domain (TRD) of MeCP2 showed a higher rate of amino acid substitution. On the other hand, non-domain regions that were also usually disordered—excluding the ordered region surrounding a domain in FOXG1—typically exhibited a higher evolutionary rate, although some regions with low rates of evolution were nonetheless detected (Fig. 2). This was corroborated by the distribution of evolutionary rates for predicted structural order–disorder residues in the three RTT-causing proteins, with disordered residues showing a wide and overlapping distribution that reflected their conservation ($P < 2.2e{-}16$ for CDKL5 and FOXG1 and $P < 6.409e{-}08$ for MeCP2, Mann-Whitney U test; Fig. 3).

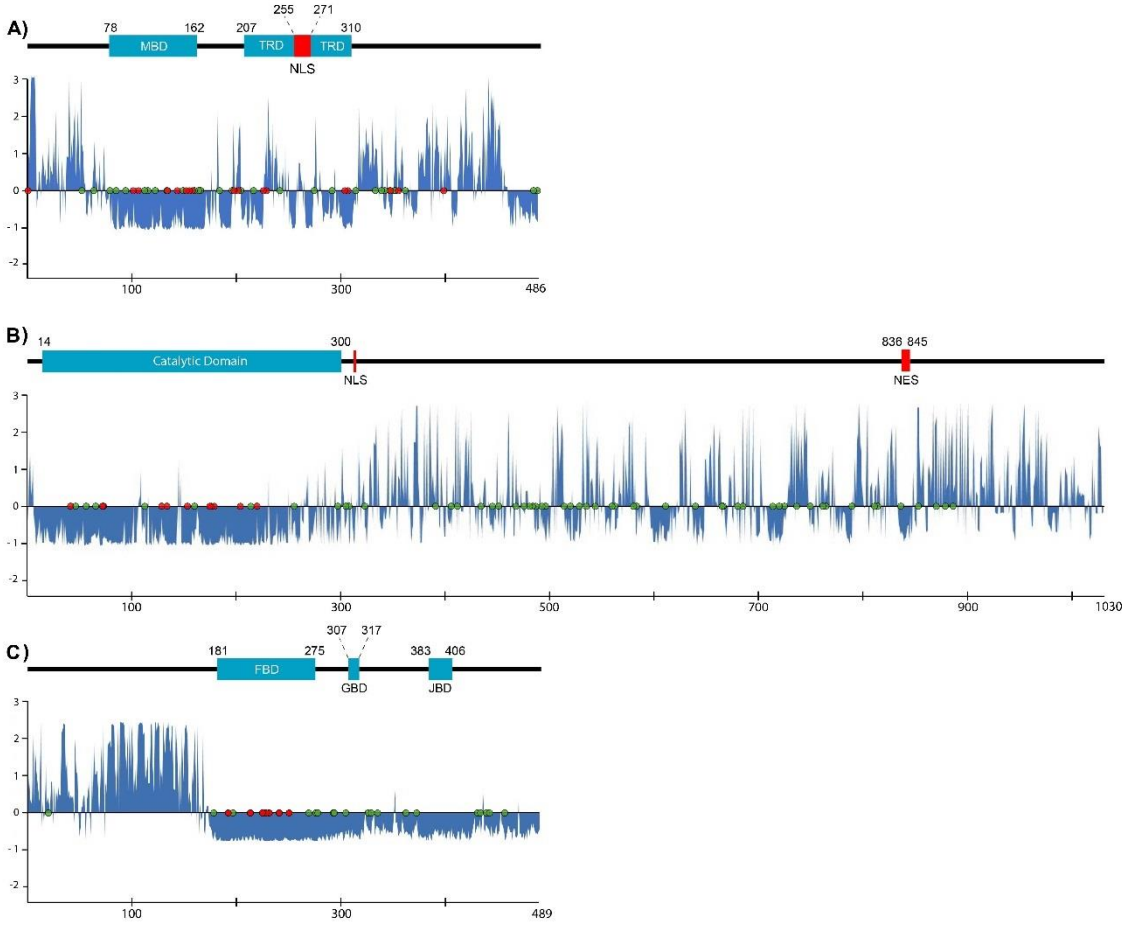

Figure 2. Rate of evolution per site in human RTT-related proteins. (a–c) Rates of amino acid substitution in MECP2 (A), CDKL5 (B), and FOXG1 (C) are shown as blue areas. The bars above charts indicate the position of the domain in the human sequence, with light blue areas indicating the domain and black lines indicating no domain. Phosphorylated amino acids and pathogenic missense mutation sites are indicated by green and red dots, respectively. The x and y axes represent the sequence length and Z score of evolutionary rate, respectively.

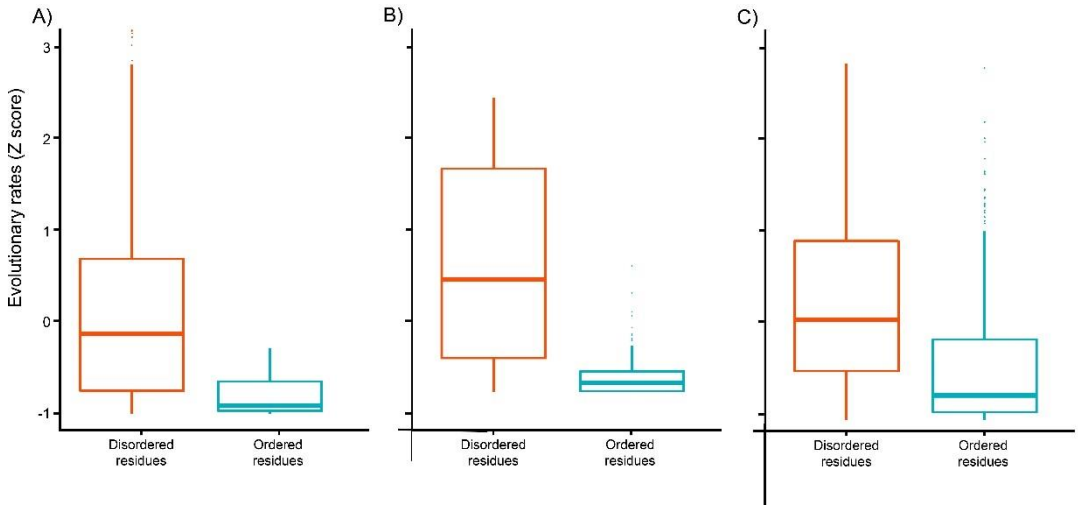

Figure 3. Boxplots of evolutionary rates for predicted structural order–disorder residues of human RTT-causing proteins. (a–c) Boxes representing predicted ordered (blue) and disordered (red) structure residues in MECP2 (A), CDKL5 (B), and FOXG1 (C). The x axis and y axes represent predicted conformation and Z score of evolutionary rate, respectively.

2.3 Post-translational modifications (PTMs)

We predicted PTM (phosphorylation) sites in chordate sequences of RTT-causing proteins and identified conserved PTM sites. We found numerous conserved phosphorylation sites including 60/82 in CDKL5, 30/45 in MeCP2, and all 23 sites in FOXG1 in human RTT-causing proteins (Fig. 2 green dots and Supplementary Table S3). Structural disorder makes such sites accessible for phosphorylation. PTMs affect the stability, turnover, interaction potential, and localization of proteins within the cell; proteins with disordered regions are more likely to have many PTMs [14, 15]. However, half of the phosphorylation sites of FOXG1 were distributed in the ordered region near a domain.

2.4 Disease-associated missense mutation distribution in the sequence of RTT-causing proteins

Sites of pathogenic RTT-associated missense mutation in human MeCP2, CDKL5, and FOXG1 were identified using RettBASE and the features of the corresponding sequence including domain or non-domain regions, predicted ordered or disordered residues, and rapid or slow evolutionary rates were examined. There were 7, 12, and 18 sites in FOXG1, CDKL5, and MeCP2, respectively, that harbored pathogenic missense mutations associated with RTT (Figure 2 and Supplementary Table S4). When the frequencies were combined with that of cases observed for each mutation, MeCP2 had a higher number of cases (1225) than CDKL5 (30) and FOXG1 (eight) (Supplementary Table S4). Pathogenic RTT-associated missense mutations were more frequently detected in domain regions for all proteins, and in ordered and slowly evolving regions for MeCP2 and CDKL5 (Table 1). Based on these results and the tendency for domain regions to have a slower evolutionary rate, we suggest that pathogenic RTT-associated missense mutations in the RTT-causing proteins tend to occur in the domain region irrespective of whether the structure is ordered or disordered. Furthermore, missense mutations in disordered residues have the highest probability of being pathogenic in RTT. On the other hand, many mutation sites in MeCP2 were located close to (or in the case of Ser346Arg and Ser134Cys, overlapped with) phosphorylation sites (Fig. 2), although the frequency of cases harboring these mutation sites was low (only one for each).

Table 1. Summary of the relationship between frequency of observed cases of pathogenic RTT-associated missense mutation and sequence features at each residue of RTT-causing proteins (ordered vs. disordered, domain vs. non-domain, and low vs. high rate of evolution)

Sequence Property	FOXG1			CDKL5			MECP2		
	Proportion of predicted properties	Frequency of observed RTT case	p-value	Proportion of predicted properties	Frequency of observed RTT case	p-value	Proportion of predicted properties	Frequency of observed RTT case	p-value
Structural order-disorder									
Ordered	253	8	> 0.001	378	30	< 0.001	26	243	< 0.001
disordered	236	0		652	0		460	982	
Domain non-domain region									
Domain	127	8	< 0.001	286	30	< 0.001	188	1217	< 0.001
Non-domain	362	0		744	0		298	8	
Rates of evolution									
Low rates	331	8	> 0.001	630	30	< 0.001	283	1217	< 0.001
Fast rates	158	0		400	0		203	8	

2.5 Phylogenetic profiling of RTT-causing proteins and their interaction partners

We retrieved 240 human proteins interacting with RTT-causing proteins from BioGRID and UniProt databases (Supplementary Table S5) [35, 36]. Phylogenetic profiling and cluster analysis of 326 eukaryotes were performed using the retrieved sequences and the sequences of the three RTT-causing proteins as queries (Fig. 4, Supplementary Table S6). The dataset was divided into four clusters, which were defined as class 1 to 4. There were 58 conserved proteins in chordates of class 1, 92 in metazoans of class 2, 17 in plants of class 3, and 73 in eukaryotes of class 4. MeCP2 and CDKL5 belonged to class 1 whereas FOXG1 belonged to class 2 (Fig. 4).

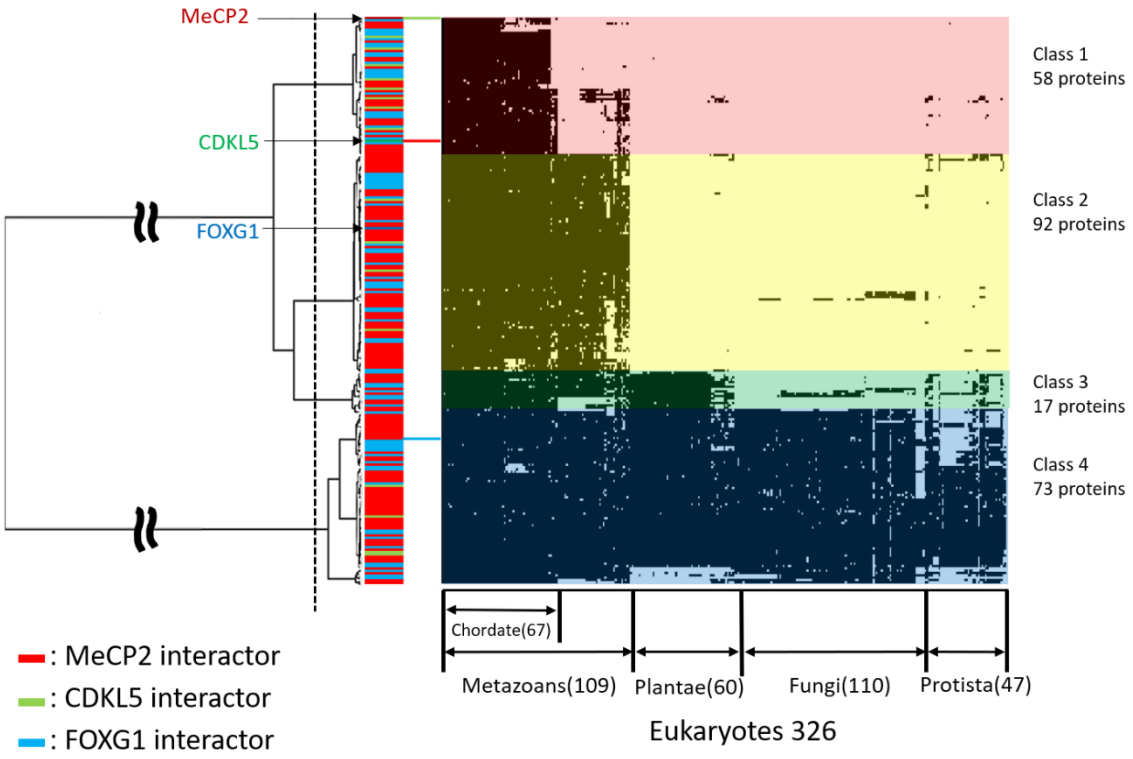

Figure 4. Phylogenetic profiling of MeCP2, CDKL5, and FOXG1 proteins and their interaction partners. The horizontal axis shows 326 eukaryotes for which whole genome sequences are available, and the vertical axis shows 240 human proteins related to RTT. Human proteins in each species are shown in black. The phylogenetic tree was divided into four clusters (class 1–4); those conserved across chordates, metazoan, plants, and eukaryotes are shown.

2.6 Subcellular localization and Gene Ontology (GO) analysis

Based on the evolutionary classification, we determined the subcellular localization of each protein and GO categories in each class (Fig. 5, Supplementary Table S7). Specific GO categories included epigenetic regulation of gene expression, transcriptional regulation, and organ or organ morphogenesis (Fig. 5). We confirmed the evolutionary trends of proteins with specific GO categories and their subcellular localization and found that 129 and 48 proteins in classes 1–4 were expressed in the nucleus only or in the nucleus and cytoplasm, respectively. Proteins in classes 1–4 were represented in the epigenetic regulation of gene expression category, whereas transcriptional regulation was observed only in classes 1 and 2 and organogenesis and organ morphogenesis were mainly observed in class 2 (Fig. 5).

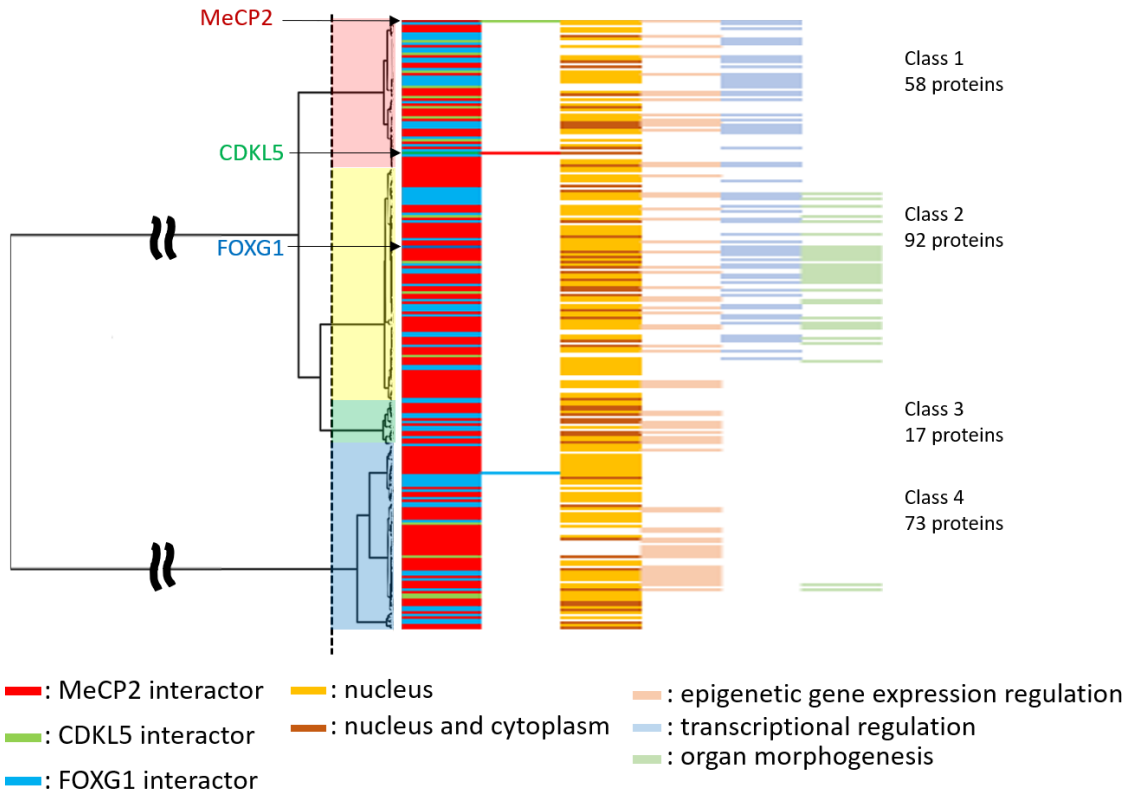

Figure 5. Subcellular localization and specific GO categories of human RTT-related proteins: Phylogenetic trees show interactors, subcellular localization, and specific GO categories for each protein. The vertical axis shows 240 RTT-related proteins, and each bar shows MeCP2, CDKL5, and FOXG1 interactors; nuclear localization; epigenetic regulation of gene expression; transcriptional regulation; and organogenesis from left to right.

2.7 Tissue and organ localization

Tissue and organ expression data for 237 proteins were extracted from The Human Protein Atlas as transcripts per million (TPM) values [37]. In addition, four proteins were not expressed in the cerebral cortex. Tissues and organs with specific expression were identified using 195 RTT-related human proteins as queries (Fig. 6, Supplementary Table S8). There were nine proteins that were specifically expressed in the cerebral cortex including apoliprotein E, CDKL5, special AT-rich sequence-binding protein (SATB)2, spalt-like transcription factor (SALL)1, zinc finger protein (ZNF)483, FOXG1, (sex determining region Y)-box (SOX)2, homeodomain-interacting protein kinase (HIPK)2, and histone cluster 2 H3 family member A.

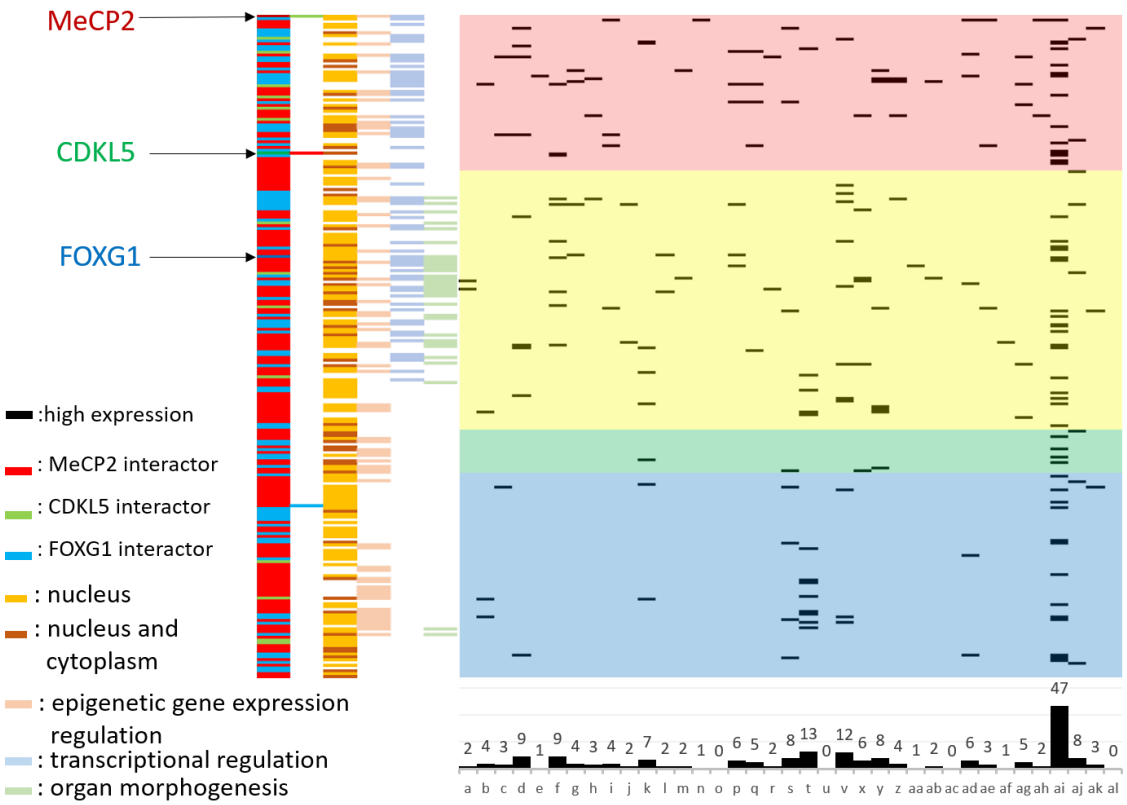

Figure 6. Tissue and organ expression analysis of human RTT-related proteins. The vertical and horizontal axes show RTT related proteins and 37 tissue types classified according to the Human Protein Atlas [37]. The tissue expressing each protein satisfying the range determined with Equation 3 is shown in black. The lower part of the figure shows the number of specifically expressed proteins. a, adipose tissue; b, adrenal gland; c, appendix; d, bone marrow; e, breast; f, cerebral cortex; g, cervix; uterine; h, colon; i, duodenum; j, endometrium; k, epididymis; l, esophagus; m, fallopian tube; n, gallbladder; o, heart muscle; p, kidney; q, liver; r, lung; s, lymph node; t, ovary; u, pancreas; v, parathyroid gland; x, placenta; y, prostate; z, rectum; aa, salivary gland; ab, seminal vesicle; ac, skeletal muscle; ad, skin; ae, small intestine; af, smooth muscle; ag, spleen; ah, stomach; ai, testis; aj, thyroid gland; ak, tonsil; al, urinary bladder.

3. Discussion

The spatiotemporal heterogeneity of IDR conformations allows a protein to bind to multiple partners. Proteins with long disordered region are tightly regulated to ensure their availability at the appropriate level and time [19, 22–24]. While mutations in IDR-containing proteins are often associated with diseases, IDR sequences generally evolve more rapidly than sequences of ordered regions. RTT is a progressive postnatal neurodevelopmental disorder in females characterized by intellectual disability, with an incidence of ~1:10,000 [38, 39]. It is mainly caused by mutations in MeCP2 (which contains an IDR) and CDKL5 or FOXG1 (which are only predicted to have this conformation) [7-9]. To obtain insight into the evolution of proteins with IDRs in association with pathogenic RTT-associated missense mutation, in this study we analyzed the evolution of the RTT-causing proteins MeCP2, FOXG1, and CDKL5, in the context of structural order–disorder and rate of amino acid substitution per site. We found that all three proteins had a conformation that was conserved across chordates, with IDR residues that evolved rapidly. However, the distributions of evolutionary rates of disordered residues were broad and overlapped with those of ordered residues, indicating that some disordered residues were conserved.

Based on these observations, we identified structurally conserved disordered regions with slowly and rapidly evolving residues reflecting constrained and flexible disorder, respectively [19]. For the latter, despite rapid evolution of residues, the change from structurally disordered to ordered

could affect protein function; hence, amino acid substitutions are constrained to residues that confer structural flexibility. This type of IDR typically functions as an entropic spring, flexible linker, or spacer without becoming structured and is frequently located outside the domain region [19,40-42]. In contrast, constrained disorder is associated with protein–protein interaction interfaces that adopt a structured conformation or undergo folding upon binding and are thus constrained in terms of sequence while still requiring flexibility; these modules are usually linear or short linear motifs [19, 43, 44]. There are also domains that confer constrained disorder called IDDs [19]. In this study, this was observed in the MBD—which was predicted to be partly disordered—and in the TRD of MeCP2, which is in accordance with previous reports that structured regions are found only in the MBD while other regions including the TRD are extensively disordered [7, 8, 45]. Most domains with conserved disordered regions are involved in DNA, RNA, and protein binding, which has been demonstrated by both domains of MeCP2 [8, 46].

Insertions and deletions were frequently detected in disordered regions. This is caused by their flexibility, which makes sequence alignment difficult; a tendency of linear motifs to lie between flexible disordered region; and permutation of functional modules with respect to others during evolution that is possible in disordered region, such as SUMO modification sites in *Drosophila* and human p53 that are located before and after the oligomerization domain, respectively [19, 47]. Phosphorylation is important to modulate the balance of proteins between the bound and unbound state and previous study has reported that kinases target disordered proteins as many as twice on average of structured proteins [48,49]. In this Study, most predicted human PTM (phosphorylation) sites in RTT-causing proteins are conserved across chordates and are located in disordered regions; one exception is FOXG1, in which almost half of the phosphorylation sites are located in predicted ordered regions, indicating that functionally constrained disordered regions beyond the domain can be aligned and that RTT-causing proteins can retain their functional modules from phosphorylation despite harboring numerous insertions and deletions.

We also analyzed the distribution of pathogenic RTT-associated missense mutations and found that regardless of the structural order–disorder property, such mutations in RTT-causing proteins tended to occur in the domain region, especially for FOXG1 and CDKL5. Furthermore, although disease-related missense mutations are present in nearly all regions according to RettBase, many were not categorized as pathogenic. For example, R453Q in human MeCP2, which is well documented in UniProt, showed a lower rate of evolution in this study; an *in silico* analysis of wild-type and mutant proteins using IUPred showed a decrease and shift in IUPred scores at positions 335–486 as well as at the C terminus, which initially exhibited a propensity for disorder, although this variant was not categorized as RTT-related [52, 53]. We suggest that pathogenic RTT are predominantly associated with the alteration in domain region. Since this unit is responsible for most of the functions of a protein and harbors residues that evolve more slowly, this can lead to protein aggregation, loss of normal function, and gain of deleterious function if this region fails to adopt its normal conformation [21]. However, this is only relevant for FOXG1 and CDKL5.

Missense mutations in ordered residues are likely to be pathogenic according to the proportion of predicted structural order–disorder in the sequence. Moreover, most identified domains are structured and only 14% of Pfam domains have over 50% of residues that are predicted to be disordered [19]. However, missense mutations in disordered residues have the highest probability of being pathogenic in RTT. This is especially true of mutations in the MBD and TRD of MeCP2, which protein is highly expressed in the brain; these two domains cooperatively mediate transcriptional repression of neuronal genes [8, 52]. The MBD is highly conserved, differing by just four residues between human and *Xenopus* orthologs; the binding properties of this domain can also be modulated by another region outside the domain [8]. Hence, this explains why the presence of missense mutations in other regions such as in rapidly evolving residues that may contribute to flexible disorder could also be deleterious and associated with pathogenic RTT in MeCP2.

We investigated the molecular evolution of MeCP2, CDKL5, and FOXG1 and their interacting proteins by phylogenetic cluster analysis and found that 243 RTT-related molecules formed four clusters—i.e., chordates, metazoans, plants, and eukaryotes. Based on these findings, we propose that

acquisition of each RTT-related gene (RTT-causing proteins and their interaction partners) coincides with the emergence of plants, metazoans, and chordates. Among the three RTT-causing proteins, only FOXG1 was a member of class 2, which comprises genes acquired during metazoan evolution. On the other hand, acquisition of MeCP2 and CDKL5 was correlated with chordate evolution. Thus, acquisition of the *MeCP2* and *CDKL5* genes may have enabled the development of the chordate brain, whereas acquisition of the *FOXG1* gene may have been critical for the development of multicellular systems including the metazoan nervous system.

Among the 237 class 1 or 2 genes, 233 were detected in the cerebral cortex with nine expressed at a high level (Fig. 6). Seven genes were acquired during metazoan evolution, of which four and three encode MeCP2- and FOXG1-interacting molecules, respectively. Since FOXG1 was also acquired during metazoan evolution, acquisition of *FOXG1*, *SATB2*, and *SALL1* may have played an important role in development of the neocortex. FOXG1 is transiently expressed in neuronal progenitor cells and regulates their migration to the cortical plate [51]. During this process, FOXG1 expression is upregulated, which contributes to cortical plate development [54]. Similarly, the FOXG1-interacting chromatin remodeling factor SATB2 was found to be expressed in the cortical plate and regulate neocortical development [55, 56]. Therefore, it is conceivable that transcriptional co-operation between FOXG1 and SATB2 mediates the laminarization of the neocortex. In support of this possibility, patients with *SATB2* mutation exhibit an RTT-like phenotype [57, 58].

MeCP2 was acquired during chordate evolution; a prerequisite for this step was the acquisition of MeCP2-interacting molecules such as ZNF483, SOX2, HIPK2, and HIST2H2A. The MeCP2 kinase HIPK2 was shown to be required for induction of apoptotic cell death in neuronal and other cell types via phosphorylation of the MeCP2 N terminus [59]. Given that CDKL5, another MeCP2 kinase, was also acquired during chordate evolution, it is possible that HIPK2 and CDKL5 cooperate to activate MeCP2 during neocortical development. Since, apoptotic cell death increased in CDKL5 knockout mice brain, CDKL5 probably works suppressive in apoptosis process in contrast to HIPK2 [60]. Therefore, functional division of their kinases through phosphorylation of MeCP2 are important issue. Indeed, the CDKL5-interacting domain was shown to associate with the C terminus of MeCP2 [61]. Hence, CDKL5 may phosphorylates carboxy terminus. Thus, both HIPK2 and CDKL5 may activate MeCP2 by phosphorylating different regions of the protein.

It is important to remember that the features of structural order-disorder and phosphorylation sites in this study have been inferred using linear sequence predictors and the sequences and mutation points were retrieved from databases which data have been collected from studies with various methods. However, the results can still be used and considered as idea for further identification of relationship between IDR and diseases.

4. Materials and Methods

4.1 Sequence retrieval, alignment, and phylogenetic analysis of RTT-causing proteins

Orthologous sequences of human RTT-causing proteins (MeCP2, CDKL5, and FOXG1) in chordates were retrieved from the Kyoto Encyclopedia of Genes and Genomes (KEGG) sequence similarity database with a Smith-Waterman similarity score threshold of 100 and the bidirectional best hits (best-best hits) option [62]. The highest similarity score for each species was used for each RTT-causing protein to minimize redundancy. Datasets were created for each RTT-causing protein and then aligned using MAFFT v.7 with the iterative refinement method (FFT-NS-i), with a maximum of 1000 iterations [63]. Phylogenetic trees were constructed with the maximum likelihood method using RAxML-HPC2 BlackBox with the RAxML automatic bootstrapping option in the CIPRES Science Gateway [64, 65]. The Jones, Taylor, and Thornton (JTT) amino acid substitution model was used for all datasets.

4.2 Structural order–disorder prediction

The structural order–disorder propensity of each protein was predicted using IUPred2A [66] using the option for long disordered regions. This prediction had values ranging from 0 (strong propensity for an ordered structure) to 1 (strong propensity for a disordered structure), with 0.5 as the cut-off between propensity for order and disorder. The results for each site of each protein were mapped onto its sequence alignment and taxon position in the phylogenetic tree using iTOL [67].

4.3 Rate of evolution per site

We calculated the rate of evolution per site of human CDKL5, FOXG1, and MeCP2 relative to their orthologs using Rate4site [68]. The aligned sequences of each protein dataset were calculated using the empirical Bayesian principle with the JTT model and 16 discrete categories of the prior gamma distribution. Gaps were treated as missing data, and outputs were standardized as Z scores. The results of the rate of evolution of each residue were then integrated with the structural order–disorder prediction result and the distribution of the rate of evolution in the structural order and disorder of each protein was evaluated with the Mann-Whitney U test using R software.

4.4 PTM prediction

We predicted phosphorylation sites using NetPhos 3.1 [69] to infer PTM sites conserved between human CDKL5, FOXG1, and MeCP2 sequences and their orthologs. The predictions had values ranging from 0 (strong propensity for obtaining a negative result) to 1 (strong propensity for obtaining a positive result); we used 0.75 as a cut-off to divide the negative and positive results. The prediction results for each sequence were plotted following multiple sequence alignment of each protein dataset. Predicted PTM sites in each dataset were considered as conserved through evolution if they had a positive value according to the 50% majority rule of the amount of sequence in the alignment.

4.5 Point mutations in RTT-causing proteins

Point mutations in CDKL5, FOXG1, and MeCP2 were identified from RettBASE [34]. We selected missense mutations that were associated with pathogenic RTT. To investigate the distribution of point mutations with respect to sequence features, we determined the sites of missense point mutations, calculated the frequency of observed cases with the point mutations, and combined the frequency with three different sequence features including domain or non-domain region, predicted ordered or disordered residue, and rapid or slow evolutionary rate (Supplementary Table S2). The frequency counts of observed cases of RTT-associated mutation in two variables of each sequence feature category were compared with the sequence feature proportion based on the prediction using the chi-square test, Fisher's exact test, and Yates's chi-square test for sequence feature categories with a score greater than or equal to 5 for all variables, a score less than 5, and a score of 0, respectively.

4.6 Phylogenetic profiling and cluster analyses of human MeCP2, CDKL5, and FOXG1 and their interacting proteins

Sequences of human MeCP2, CDKL5, and FOXG1 and their interaction partners identified with BioGRID (release 2019_03) were obtained from the UniProtKB/Swiss-Prot database (release 2019_04) and used as the dataset [35, 36]. We generated phylogenetic profiles of 326 eukaryotes in the KEGG database using the dataset as query [70]. Phylogenetic profiling is a method for detecting the presence or absence of orthologous proteins in a target organism [71]. The presence or absence of proteins homologous to the query in each species was determined using KEGG Ortholog Cluster (release 2019_04) [72]. Profiles were determined based on Manhattan distance and then clustered using Ward's method [73].

4.7 Protein expression in human tissues

Expression levels of human RTT-related proteins in each tissue were extracted from the Human Protein Atlas (release 2019_4) [37] and classified into 37 tissues. Protein expression level was determined using the TPM value, which was corrected for protein expression by gene length. Comparisons of protein expression levels were not shown as a ratio so that proteins with high expression did not skew the results (Equations 1–3). The mean and standard deviation were derived from Equations (1) and (2) and the range was obtained from Equation (3). The range in Equation (3) was taken as the tissue for each of the specifically expressed proteins—i.e., the value was "1" when included in the range of Equation (3) and "0" when it was not included in the expression level of each protein expressed as a percentage. The procedure yielded human protein-specific expression profiles in the context of RTT.

$\mu = \frac{1}{n}\sum_{i=0}^{n} x_i$	(1)
$s = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(x_i - \mu)^2}$	(2)
$\mu + 1.65 \times s \ < x$	(3)

Here, μ, s, n, and x are the mean, standard deviation, number of samples, and one sample, respectively. The value of 1.65 in Equation (3) is the standard confidence factor for extracting data outside the 90% confidence interval.

4. 8 GO analysis

Specific GO categories in the target protein group were obtained using the Panther tool [74]. Categories with an appearance frequency of $P < 0.05$ were defined as protein group-specific. In this study, we obtained GO categories specific for human proteins related to RTT that were classified based on defined functions.

5. Conclusions

Evolutionary analyses of RTT-causing proteins provide insight into the distribution of pathogenic RTT missense mutations, which tend to occur in domain regions and affect ordered residues for CDKL5 and FOXG1. However, missense mutations in disordered residues of MeCP2 may be more deleterious owing to the numerous binding partners, and can also be found outside the domain region since IDDs could function by forming contacts with other sequences in a protein. Our phylogenetic cluster analysis revealed that the RTT-causing proteins MeCP2, CDKL5, and FOXG1 and their interaction partners may have been acquired during metazoan evolution and play essential roles in neocortical development. After the emergence of chordates or vertebrates, interactions between these newly acquired and pre-existing genes may have permitted the evolution of the human neocortex.

Author Contributions conceptualization, M.F., Y.K. and M.I. ; methodology, M.F., G.Y., Y.K. and M.I..; software, M.F. and G.Y.; validation, M.F., G.Y. and K.S.; formal analysis, M.F. and G.Y.; investigation, M.F., G.Y., K.S., S.K., T.K.-K. T.I., Y.K. and M.I.; resources, S.K, T.K.-K. and T.I.; data curation, M.F., Y.K. and M.I. ; writing—original draft preparation, M.F..; writing—review and editing, S.K., T.K.-K. T.I., Y.K. and M.I.; visualization, M.F., G.Y. and K.S.;; supervision, M.I.; project administration, M.I.; funding acquisition, T.I. and M.I.

Funding: This study was supported by MEXT-supported program for the strategic research foundation at private universities (2015-2019 to T.I).

Acknowledgments: We would like to thank Mr. Takahiro Nakamura for support and helpful comments.

Conflicts of Interest: The authors declare no competing interest.

Abbreviations

APOE	Apolipoprotein E
BioGRID	Biological General Repository for Interaction Datasets
CDKL5	Cyclin-dependent kinase-like 5
CIPRES	Cyberinfrastructure for Phylogenetic Research
DOI	digital object identifier
FBD	Forkhead box domain
FOXG1	Forkhead box protein G1
GBD	Groucho-binding domain
GO	Gene Ontology
HIPK2	homeodomain-interacting protein kinase 2
IDD	intrinsically disordered domain
IDR	intrinsically disordered protein
iTOL	Interactive Tree of Life
IUPred	Prediction of Intrinsically Unstructured Proteins
JBD	JARID1B-binding domain
JTT	The Jones, Taylor, and Thornton
KEGG	Kyoto Encyclopedia of Genes and Genomes
MAFFT	Modified Multiple Alignment Fast Fourier Transform
MBD	Methyl-CpG-binding domain
MeCP2	Methyl-CpG-binding protein 2
NES	nuclear export signal
NLS	nuclear localization signal
OMIM	Online Mendelian Inheritance in Man
PTM	post-translational modification
RAxML-HPC2	Randomized Axelerated Maximum Likelihood for High Performance Computing 2
RTT	Rett syndrome
RettBASE	Rett syndrome Variation Database
SALL1	spalt-like transcription factor 1
SATB2	Special AT-rich sequence-binding protein 2
SOX2	SRY-box transcription factor 2
SSDB	Sequence Similarity DataBase
SUMO	Small ubiquitin-related modifier
TPM	transcripts per million
TRD	transcriptional repression domain
Z scores	standardized scores
ZNF483	zinc finger protein 483

References

1. Rett, A. On a unusual brain atrophy syndrome in hyperammonemia in childhood. *Wien Med Wochenschr* **1966**, *116*, 723-726.
2. Hanefeld, F. The clinical pattern of the Rett syndrome. *Brain Dev* **1985**, *7*, 320-325.

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

4. Smeets, E.; Schollen, E.; Moog, U.; Matthijs, G.; Herbergs, J.; Smeets, H.; Curfs, L.; Schrander-Stumpel, C.; Fryns, J.P. Rett syndrome in adolescent and adult females: clinical and molecular genetic findings. *American journal of medical genetics. Part A* **2003**, *122a*, 227-233, doi:10.1002/ajmg.a.20321.

5. Ariani, F.; Hayek, G.; Rondinella, D.; Artuso, R.; Mencarelli, M.A.; Spanhol-Rosseto, A.; Pollazzon, M.; Buoni, S.; Spiga, O.; Ricciardi, S., et al. FOXG1 is responsible for the congenital variant of Rett syndrome. *Am J Hum Genet* **2008**, *83*, 89-93, doi:10.1016/j.ajhg.2008.05.015.

6. Weaving, L.S.; Christodoulou, J.; Williamson, S.L.; Friend, K.L.; McKenzie, O.L.; Archer, H.; Evans, J.; Clarke, A.; Pelka, G.J.; Tam, P.P., et al. Mutations of CDKL5 cause a severe neurodevelopmental disorder with infantile spasms and mental retardation. *Am J Hum Genet* **2004**, *75*, 1079-1093, doi:10.1086/426462.

7. 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. *The Journal of biological chemistry* **2007**, *282*, 15057-15064, doi:10.1074/jbc.M700855200.

8. Ghosh, R.P.; Nikitina, T.; Horowitz-Scherer, R.A.; Gierasch, L.M.; Uversky, V.N.; Hite, K.; Hansen, J.C.; Woodcock, C.L. Unique physical properties and interactions of the domains of methylated DNA binding protein 2. *Biochemistry* **2010**, *49*, 4395-4410, doi:10.1021/bi9019753.

9. Toth-Petroczy, A.; Palmedo, P.; Ingraham, J.; Hopf, T.A.; Berger, B.; Sander, C.; Marks, D.S. Structured States of Disordered Proteins from Genomic Sequences. *Cell* **2016**, *167*, 158-170.e112, doi:10.1016/j.cell.2016.09.010.

10. Canning, P.; Park, K.; Goncalves, J.; Li, C.; Howard, C.J.; Sharpe, T.D.; Holt, L.J.; Pelletier, L.; Bullock, A.N.; Leroux, M.R. CDKL Family Kinases Have Evolved Distinct Structural Features and Ciliary Function. *Cell reports* **2018**, *22*, 885-894, doi:10.1016/j.celrep.2017.12.083.

11. Dunker, A.K.; Lawson, J.D.; Brown, C.J.; Williams, R.M.; Romero, P.; Oh, J.S.; Oldfield, C.J.; Campen, A.M.; Ratliff, C.M.; Hipps, K.W., et al. Intrinsically disordered protein. *Journal of molecular graphics & modelling* **2001**, *19*, 26-59.

12. Dyson, H.J.; Wright, P.E. Equilibrium NMR studies of unfolded and partially folded proteins. *Nature structural biology* **1998**, *5 Suppl*, 499-503, doi:10.1038/739.

13. Uversky, V.N. Protein folding revisited. A polypeptide chain at the folding-misfolding-nonfolding cross-roads: which way to go? *Cellular and molecular life sciences : CMLS* **2003**, *60*, 1852-1871, doi:10.1007/s00018-003-3096-6.

14. Uversky, V.N.; Gillespie, J.R.; Fink, A.L. Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins* **2000**, *41*, 415-427.

15. Romero, P.; Obradovic, Z.; Li, X.; Garner, E.C.; Brown, C.J.; Dunker, A.K. Sequence complexity of disordered protein. *Proteins* **2001**, *42*, 38-48.

16. Dunker, A.K.; Babu, M.M.; Barbar, E.; Blackledge, M.; Bondos, S.E.; Dosztanyi, Z.; Dyson, H.J.; Forman-Kay, J.; Fuxreiter, M.; Gsponer, J., et al. What's in a name? Why these proteins are intrinsically disordered: Why these proteins are intrinsically disordered. *Intrinsically disordered proteins* **2013**, *1*, e24157, doi:10.4161/idp.24157.

17. Tompa, P. Intrinsically unstructured proteins. *Trends in biochemical sciences* **2002**, *27*, 527-533.

18. Tompa, P. The interplay between structure and function in intrinsically unstructured proteins. *FEBS letters* **2005**, *579*, 3346-3354, doi:10.1016/j.febslet.2005.03.072.
19. van der Lee, R.; Buljan, M.; Lang, B.; Weatheritt, R.J.; Daughdrill, G.W.; Dunker, A.K.; Fuxreiter, M.; Gough, J.; Gsponer, J.; Jones, D.T., et al. Classification of intrinsically disordered regions and proteins. *Chemical reviews* **2014**, *114*, 6589-6631, doi:10.1021/cr400525m.
20. Daughdrill, G.W.; Pielak, G.J.; Uversky, V.N.; Cortese, M.S.; Dunker, A.K. Natively disordered proteins. *Protein folding handbook* **2005**, 275-357.
21. Uversky, V.N.; Oldfield, C.J.; Dunker, A.K. Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annual review of biophysics* **2008**, *37*, 215-246, doi:10.1146/annurev.biophys.37.032807.125924.
22. Uversky, V.N. Intrinsically Disordered Proteins and Their "Mysterious" (Meta)Physics. *Frontiers in Physics* **2019**, *7*, doi:10.3389/fphy.2019.00010.
23. Diella, F.; Haslam, N.; Chica, C.; Budd, A.; Michael, S.; Brown, N.P.; Trave, G.; Gibson, T.J. Understanding eukaryotic linear motifs and their role in cell signaling and regulation. *Frontiers in bioscience : a journal and virtual library* **2008**, *13*, 6580-6603.
24. Galea, C.A.; Wang, Y.; Sivakolundu, S.G.; Kriwacki, R.W. Regulation of cell division by intrinsically unstructured proteins: intrinsic flexibility, modularity, and signaling conduits. *Biochemistry* **2008**, *47*, 7598-7609, doi:10.1021/bi8006803.
25. Brown, C.J.; Johnson, A.K.; Dunker, A.K.; Daughdrill, G.W. Evolution and disorder. *Current opinion in structural biology* **2011**, *21*, 441-446, doi:10.1016/j.sbi.2011.02.005.
26. Blekhman, R.; Man, O.; Herrmann, L.; Boyko, A.R.; Indap, A.; Kosiol, C.; Bustamante, C.D.; Teshima, K.M.; Przeworski, M. Natural selection on genes that underlie human disease susceptibility. *Current biology : CB* **2008**, *18*, 883-889, doi:10.1016/j.cub.2008.04.074.
27. Kondrashov, F.A.; Ogurtsov, A.Y.; Kondrashov, A.S. Bioinformatical assay of human gene morbidity. *Nucleic acids research* **2004**, *32*, 1731-1737, doi:10.1093/nar/gkh330.
28. Pritchard, J.K.; Cox, N.J. The allelic architecture of human disease genes: common disease-common variant...or not? *Human molecular genetics* **2002**, *11*, 2417-2423, doi:10.1093/hmg/11.20.2417.
29. Lyst, M.J.; Bird, A. Rett syndrome: a complex disorder with simple roots. *Nat Rev Genet* **2015**, *16*, 261-275, doi:10.1038/nrg3897.
30. Carrette, L.L.G.; Wang, C.Y.; Wei, C.; Press, W.; Ma, W.; Kelleher, R.J., 3rd; Lee, J.T. A mixed modality approach towards Xi reactivation for Rett syndrome and other X-linked disorders. *Proceedings of the National Academy of Sciences of the United States of America* **2018**, *115*, E668-e675, doi:10.1073/pnas.1715124115.
31. Shah, R.R.; Bird, A.P. MeCP2 mutations: progress towards understanding and treating Rett syndrome. *Genome Med* **2017**, *9*, 17, doi:10.1186/s13073-017-0411-7.
32. Mellios, N.; Woodson, J.; Garcia, R.I.; Crawford, B.; Sharma, J.; Sheridan, S.D.; Haggarty, S.J.; Sur, M. beta2-Adrenergic receptor agonist ameliorates phenotypes and corrects microRNA-mediated IGF1 deficits in a mouse model of Rett syndrome. *Proceedings of the National Academy of Sciences of the United States of America* **2014**, *111*, 9947-9952, doi:10.1073/pnas.1309426111.
33. Tropea, D.; Giacometti, E.; Wilson, N.R.; Beard, C.; McCurry, C.; Fu, D.D.; Flannery, R.; Jaenisch, R.; Sur, M. Partial reversal of Rett Syndrome-like symptoms in MeCP2 mutant mice. *Proceedings of the National Academy of Sciences of the United States of America* **2009**, *106*, 2029-2034, doi:10.1073/pnas.0812394106.

34. Krishnaraj, R.; Ho, G.; Christodoulou, J. RettBASE: Rett syndrome database update. *Human mutation* **2017**, *38*, 922-931, doi:10.1002/humu.23263.
35. The Universal Protein Resource (UniProt) in 2010. *Nucleic acids research* **2010**, *38*, D142-148, doi:10.1093/nar/gkp846.
36. Chatr-Aryamontri, A.; Oughtred, R.; Boucher, L.; Rust, J.; Chang, C.; Kolas, N.K.; O'Donnell, L.; Oster, S.; Theesfeld, C.; Sellam, A., et al. The BioGRID interaction database: 2017 update. *Nucleic acids research* **2017**, *45*, D369-d379, doi:10.1093/nar/gkw1102.
37. Uhlen, M.; Fagerberg, L.; Hallstrom, B.M.; Lindskog, C.; Oksvold, P.; Mardinoglu, A.; Sivertsson, A.; Kampf, C.; Sjostedt, E.; Asplund, A., et al. Proteomics. Tissue-based map of the human proteome. *Science* **2015**, *347*, 1260419, doi:10.1126/science.1260419.
38. Chahrour, M.; Zoghbi, H.Y. The story of Rett syndrome: from clinic to neurobiology. *Neuron* **2007**, *56*, 422-437, doi:10.1016/j.neuron.2007.10.001.
39. Francke, U. Mechanisms of disease: neurogenetics of MeCP2 deficiency. *Nature clinical practice. Neurology* **2006**, *2*, 212-221, doi:10.1038/ncpneuro0148.
40. Dyson, H.J.; Wright, P.E. Intrinsically unstructured proteins and their functions. *Nature reviews. Molecular cell biology* **2005**, *6*, 197-208, doi:10.1038/nrm1589.
41. Fahmi, M.; Ito, M. Evolutionary Approach of Intrinsically Disordered CIP/KIP Proteins. *Scientific reports* **2019**, *9*, 1575, doi:10.1038/s41598-018-37917-5.
42. Gsponer, J.; Babu, M.M. The rules of disorder or why disorder rules. *Progress in biophysics and molecular biology* **2009**, *99*, 94-103, doi:10.1016/j.pbiomolbio.2009.03.001.
43. Bellay, J.; Han, S.; Michaut, M.; Kim, T.; Costanzo, M.; Andrews, B.J.; Boone, C.; Bader, G.D.; Myers, C.L.; Kim, P.M. Bringing order to protein disorder through comparative genomics and genetic interactions. *Genome biology* **2011**, *12*, R14, doi:10.1186/gb-2011-12-2-r14.
44. Mi, T.; Merlin, J.C.; Deverasetty, S.; Gryk, M.R.; Bill, T.J.; Brooks, A.W.; Lee, L.Y.; Rathnayake, V.; Ross, C.A.; Sargeant, D.P., et al. Minimotif Miner 3.0: database expansion and significantly improved reduction of false-positive predictions from consensus sequences. *Nucleic acids research* **2012**, *40*, D252-260, doi:10.1093/nar/gkr1189.
45. Wakefield, R.I.; Smith, B.O.; Nan, X.; Free, A.; Soteriou, A.; Uhrin, D.; Bird, A.P.; Barlow, P.N. The solution structure of the domain from MeCP2 that binds to methylated DNA. *Journal of molecular biology* **1999**, *291*, 1055-1065, doi:10.1006/jmbi.1999.3023.
46. Chen, J.W.; Romero, P.; Uversky, V.N.; Dunker, A.K. Conservation of intrinsic disorder in protein domains and families: II. functions of conserved disorder. *Journal of proteome research* **2006**, *5*, 888-898, doi:10.1021/pr060049p.
47. Mauri, F.; McNamee, L.M.; Lunardi, A.; Chiacchiera, F.; Del Sal, G.; Brodsky, M.H.; Collavin, L. Modification of Drosophila p53 by SUMO modulates its transactivation and pro-apoptotic functions. *The Journal of biological chemistry* **2008**, *283*, 20848-20856, doi:10.1074/jbc.M710186200.
48. Grimmler, M.; Wang, Y.; Mund, T.; Cilenšek, Z.; Keidel, E.M.; Waddell, M.B.; Jäkel, H.; Kullmann, M.; Kriwacki, R.W.; Hengst, L. Cdk-inhibitory activity and stability of p27Kip1 are directly regulated by oncogenic tyrosine kinases. Cell **2007**, 128, pp. 269-280, doi: 10.1016/j.cell.2006.11.047.
49. Gsponer, J.; Futschik, M.E.; Teichmann, S.A.; Babu, M.M. Tight regulation of unstructured proteins: from transcript synthesis to protein degradation. Science **2008**, 322, pp.1365-1368, doi: 10.1126/science.1163581.

50. Couvert, P.; Bienvenu, T.; Aquaviva, C.; Poirier, K.; Moraine, C.; Gendrot, C.; Verloes, A.; Andres, C.; Le Fevre, A.C.; Souville, I., et al. MECP2 is highly mutated in X-linked mental retardation. *Human molecular genetics* **2001**, *10*, 941-946, doi:10.1093/hmg/10.9.941.
51. Vacic, V.; Iakoucheva, L.M. Disease mutations in disordered regions--exception to the rule? *Molecular bioSystems* **2012**, *8*, 27-32, doi:10.1039/c1mb05251a.
52. Nan, X.; Campoy, F.J.; Bird, A. MeCP2 is a transcriptional repressor with abundant binding sites in genomic chromatin. *Cell* **1997**, *88*, 471-481.
53. Miyoshi, G.; Fishell, G. Dynamic FoxG1 expression coordinates the integration of multipolar pyramidal neuron precursors into the cortical plate. *Neuron* **2012**, *74*, 1045-1058, doi:10.1016/j.neuron.2012.04.025.
54. Kumamoto, T.; Toma, K.; Gunadi; McKenna, W.L.; Kasukawa, T.; Katzman, S.; Chen, B.; Hanashima, C. Foxg1 coordinates the switch from nonradially to radially migrating glutamatergic subtypes in the neocortex through spatiotemporal repression. *Cell reports* **2013**, *3*, 931-945, doi:10.1016/j.celrep.2013.02.023.
55. Alcamo, E.A.; Chirivella, L.; Dautzenberg, M.; Dobreva, G.; Farinas, I.; Grosschedl, R.; McConnell, S.K. Satb2 regulates callosal projection neuron identity in the developing cerebral cortex. *Neuron* **2008**, *57*, 364-377, doi:10.1016/j.neuron.2007.12.012.
56. Britanova, O.; de Juan Romero, C.; Cheung, A.; Kwan, K.Y.; Schwark, M.; Gyorgy, A.; Vogel, T.; Akopov, S.; Mitkovski, M.; Agoston, D., et al. Satb2 is a postmitotic determinant for upper-layer neuron specification in the neocortex. *Neuron* **2008**, *57*, 378-392, doi:10.1016/j.neuron.2007.12.028.
57. Docker, D.; Schubach, M.; Menzel, M.; Munz, M.; Spaich, C.; Biskup, S.; Bartholdi, D. Further delineation of the SATB2 phenotype. *Eur J Hum Genet* **2014**, *22*, 1034-1039, doi:10.1038/ejhg.2013.280.
58. Lee, J.S.; Yoo, Y.; Lim, B.C.; Kim, K.J.; Choi, M.; Chae, J.H. SATB2-associated syndrome presenting with Rett-like phenotypes. *Clin Genet* **2016**, *89*, 728-732, doi:10.1111/cge.12698.
59. Bracaglia, G.; Conca, B.; Bergo, A.; Rusconi, L.; Zhou, Z.; Greenberg, M.E.; Landsberger, N.; Soddu, S.; Kilstrup-Nielsen, C. Methyl-CpG-binding protein 2 is phosphorylated by homeodomain-interacting protein kinase 2 and contributes to apoptosis. *EMBO Rep* **2009**, *10*, 1327-1333, doi:10.1038/embor.2009.217.
60. Fuchs, C.; Trazzi, S.; Torricella, R.; Viggiano, R.; De Franceschi, M.; Amendola, E.; Gross, C.; Calza, L.; Bartesaghi, R.; Ciani, E. Loss of CDKL5 impairs survival and dendritic growth of newborn neurons by altering AKT/GSK-3beta signaling. *Neurobiology of disease* **2014**, *70*, 53-68, doi:10.1016/j.nbd.2014.06.006.
61. Mari, F.; Azimonti, S.; Bertani, I.; Bolognese, F.; Colombo, E.; Caselli, R.; Scala, E.; Longo, I.; Grosso, S.; Pescucci, C., et al. CDKL5 belongs to the same molecular pathway of MeCP2 and it is responsible for the early-onset seizure variant of Rett syndrome. *Human molecular genetics* **2005**, *14*, 1935-1946, doi:10.1093/hmg/ddi198.
62. Sato, Y.; Nakaya, A.; Shiraishi, K.; Kawashima, S.; Goto, S.; Kanehisa, M. Ssdb: Sequence similarity database in kegg. *Genome Informatics* **2001**, *12*, 230-231.
63. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution* **2013**, *30*, 772-780, doi:10.1093/molbev/mst010.
64. Miller, M.A.; Pfeiffer, W.; Schwartz, T. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In Proceedings of 2010 gateway computing environments workshop (GCE); pp. 1-8.

65. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics (Oxford, England)* **2006**, *22*, 2688-2690, doi:10.1093/bioinformatics/btl446.

66. Meszaros, B.; Erdos, G.; Dosztanyi, Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic acids research* **2018**, *46*, W329-w337, doi:10.1093/nar/gky384.

67. Letunic, I.; Bork, P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics (Oxford, England)* **2007**, *23*, 127-128, doi:10.1093/bioinformatics/btl529.

68. Pupko, T.; Bell, R.E.; Mayrose, I.; Glaser, F.; Ben-Tal, N. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics (Oxford, England)* **2002**, *18 Suppl 1*, S71-77, doi:10.1093/bioinformatics/18.suppl_1.s71.

69. Blom, N.; Gammeltoft, S.; Brunak, S. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *Journal of molecular biology* **1999**, *294*, 1351-1362, doi:10.1006/jmbi.1999.3310.

70. Kanehisa, M.; Furumichi, M.; Tanabe, M.; Sato, Y.; Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research* **2017**, *45*, D353-d361, doi:10.1093/nar/gkw1092.

71. Pellegrini, M.; Marcotte, E.M.; Thompson, M.J.; Eisenberg, D.; Yeates, T.O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America* **1999**, *96*, 4285-4288, doi:10.1073/pnas.96.8.4285.

72. Nakaya, A.; Katayama, T.; Itoh, M.; Hiranuka, K.; Kawashima, S.; Moriya, Y.; Okuda, S.; Tanaka, M.; Tokimatsu, T.; Yamanishi, Y., et al. KEGG OC: a large-scale automatic construction of taxonomy-based ortholog clusters. *Nucleic acids research* **2013**, *41*, D353-357, doi:10.1093/nar/gks1239.

73. Ward, J.H. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association* **1963**, *58*, 236-244, doi:10.1080/01621459.1963.10500845.

74. Mi, H.; Muruganujan, A.; Ebert, D.; Huang, X.; Thomas, P.D. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic acids research* **2019**, *47*, D419-d426, doi:10.1093/nar/gky1038.